

Fast Searching over Compressed Text using A New Coding Technique: Tagged Sub-optimal Code (TSC)

Abdelghani Bellaachia and Iehab AL Rasan, George Washington University, U.S.A.

A new coding technique, we call Tagged Sub-optimal code (TSC), is proposed. TSC is a variable-length sub-optimal code that supports minimal prefix property. It always determines its codeword boundary without traversing a tree or lookup table. TSC technique may be beneficial in many types of applications: speeding up string matching over compressed text, speeding decoding process, robustness of error detection and recovery during transmission, as well as any general-purpose integer representation code. Experimental results show that TSC is 8.9 times faster than string matching over compressed text using Huffman encoding, and 3 times faster in the decoding process. On the other hand, it is 6% less than the compression ratio of Huffman encoding. Additionally, TSC is 14 times faster than Byte Pair Encoding (BPE) compression process.

Although many studies have been conducted on string matching over compressed text using

well-known compression schemes [1], [2], and [3], these compression schemes are not intended to speed up data processing over compressed data and significant delay would be involved in their process. TSC is based on traversing a quad tree that generates variable-length codewords every level of the tree, and is delimited with the sequence of 01 or 10 bits. Figure 1 shows the quad tree representation of TSC technique. The Number of codewords available = $2^{h+1} - 2$, where h is the height of quad tree > 0 . Maximum codeword length = $\lfloor \log_2 x + 1 \rfloor \times 2$.

Table 1 shows the elapsed time in milliseconds for searching compressed Brown

corpus and Calgary/ Canterbury corpus files, encoding, and decoding process.

References

1. A. Amir and G . Benson. Efficient Two-dimensional Compressed Matching. In Proc. Second IEEE Data Compression Conference, pages 279-288, March 1992.
2. M. Farach and M. Thorup. String-Matching in Lempel-Ziv Compressed Strings. In 27th ACM STOC, pages 703-713,1995.
3. Nivio Ziviani, Edleno de Moura, Gonzalo Navarro and Ricardo Baeza-Yates. Compression: A Key for Next-Generation Text Retrieval Systems. IEEE Computer, v. 33 issue 11, pages 37-44, Nov 2000.

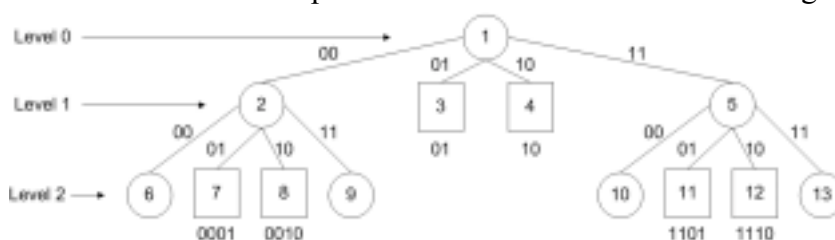


Figure 1 Quad Tree Representations of TSC

TABLE 1 : STRING MATCHING OVER COMPRESSED TEXT, ENCODING AND DECODING ELAPSED TIME IN MILLISECONDS

File	Search over TSC	Search over Huffman Text	Proposed TSC Coding		BPE		Huffman	
			Encoding	Decoding	Encoding	Decoding	Encoding	Decoding
Brown Corpus	350	3245	1573	1322	22793	1311	1543	3525
Book1	50	420	230	80	3154	50	220	440
Book2	40	350	180	60	2083	40	180	360
News	30	250	120	30	1192	30	120	260
Paper1	10	40	20	10	160	10	20	40
Paper2	10	50	30	10	290	10	30	50
Average	82	726	359	252	4945	242	352	779