July 24, 2026

Efficient Sparse Retrieval with Lightweight Superblock Pruning

Parker Carlson, Wentai Xie, Rohil Shah, Tao Yang

University of California, Santa Barbara

Background & Motivation

- Recent retrieval algorithms BMP and SP [SIGIR ’24, ’25] use a very large number of clusters to accelerate retrieval

Background & Motivation

- Recent retrieval algorithms BMP and SP [SIGIR ’24, ’25] use a very large number of clusters to accelerate retrieval
- However, they both have large index sizes and struggle with cluster overpruning during approximate retrieval

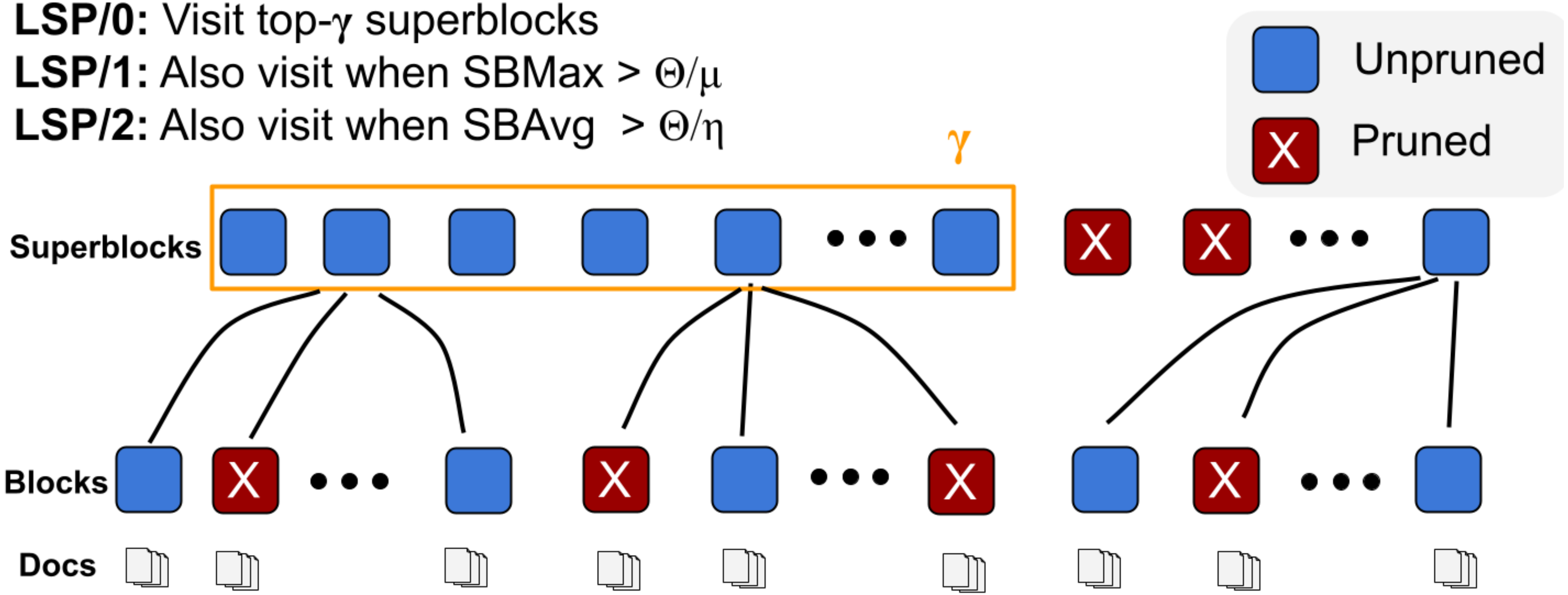
Background & Motivation

- Recent retrieval algorithms BMP and SP [SIGIR ’24, ’25] use a very large number of clusters to accelerate retrieval
- However, they both have large index sizes and struggle with cluster overpruning during approximate retrieval
- In this work, we investigate the design choices of BMP and SP and propose improvements to address their weaknesses

LSP Architecture

$$SBBoundSum(C_i); = \sum_{t \in Q} q_t \cdot \max_{d \in C_i} w_{t,d};$$

LSP/0: Visit top-γ superblocks
LSP/1: Also visit when SBMax > Θ/μ
LSP/2: Also visit when SBAvg > Θ/η



UC SANTA BARBARA

Goals of Lightweight Superblock Pruning (LSP)

1. Address the superblock overpruning problem

Goals of Lightweight Superblock Pruning (LSP)

1. Address the superblock overpruning problem
2. What superblock heuristic is best?

Goals of Lightweight Superblock Pruning (LSP)

1. Address the superblock overpruning problem
2. What superblock heuristic is best?
3. Compress the index without affecting latency

Goals of Lightweight Superblock Pruning (LSP)

1. Address the superblock overpruning problem
2. What superblock heuristic is best?
3. Compress the index without affecting latency
4. Simplify parameter choices & enable zero-shot approximate retrieval

Goals of Lightweight Superblock Pruning (LSP)

1. Address the superblock overpruning problem
2. What superblock heuristic is best?
3. Compress the index without affecting latency
4. Simplify parameter choices & enable zero-shot approximate retrieval
5. Scale to larger datasets

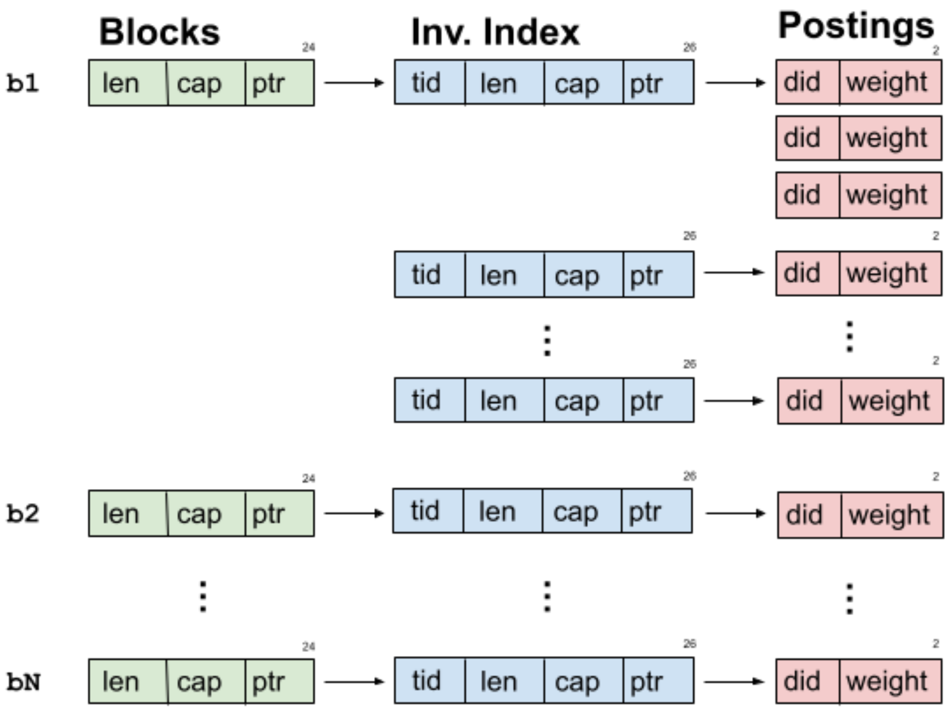
Goals of Lightweight Superblock Pruning (LSP)

1. Address the superblock overpruning problem
2. What superblock heuristic is best?
3. Compress the index without affecting latency
4. Simplify parameter choices & enable zero-shot approximate retrieval
5. Scale to larger datasets

Compression Results

Document Representations

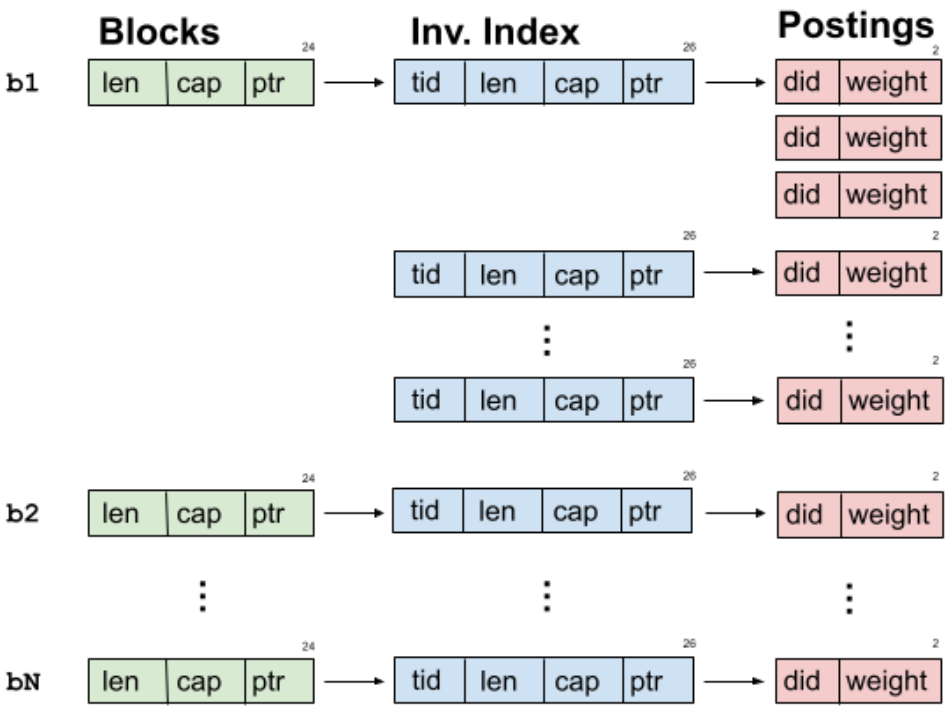
(a) BMP's Block Forward Index



UC SANTA BARBARA

Document Representations

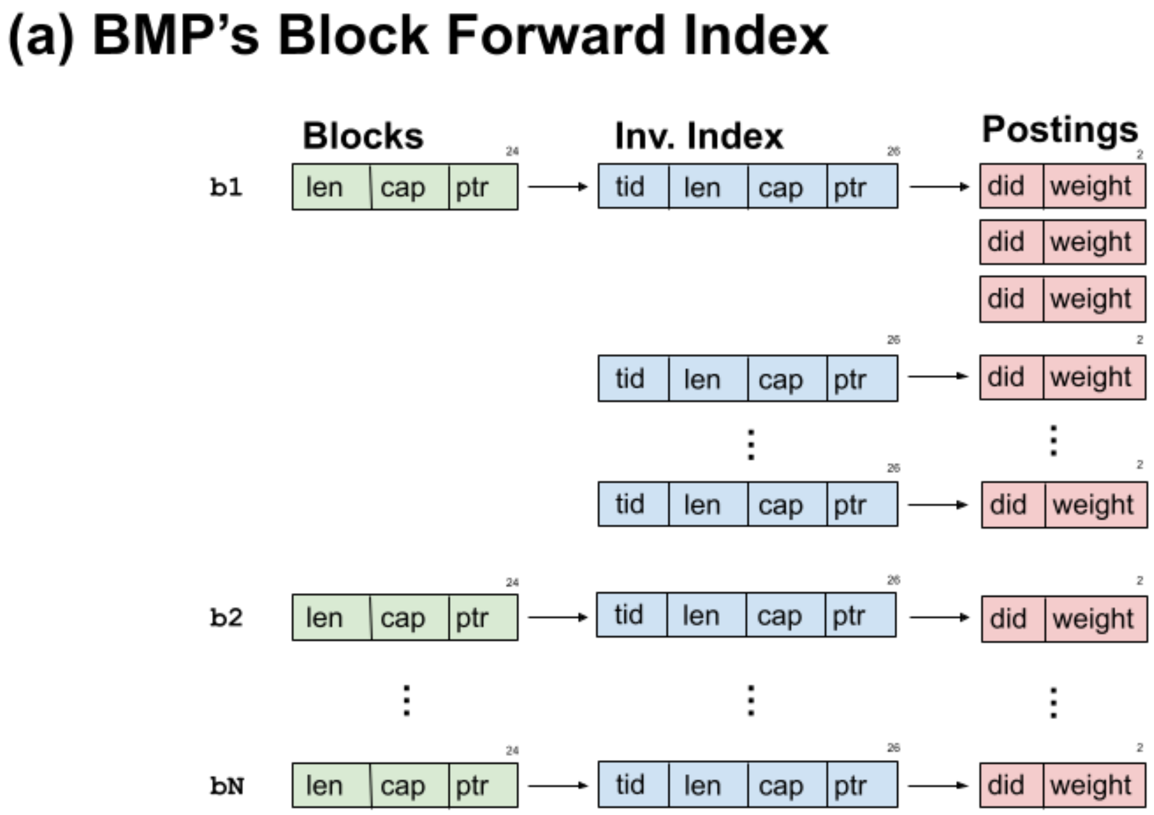
(a) BMP's Block Forward Index



On a 64-bit system, each vector requires 24 bytes!
(len: u64, cap: u64, ptr: u64)

UC SANTA BARBARA

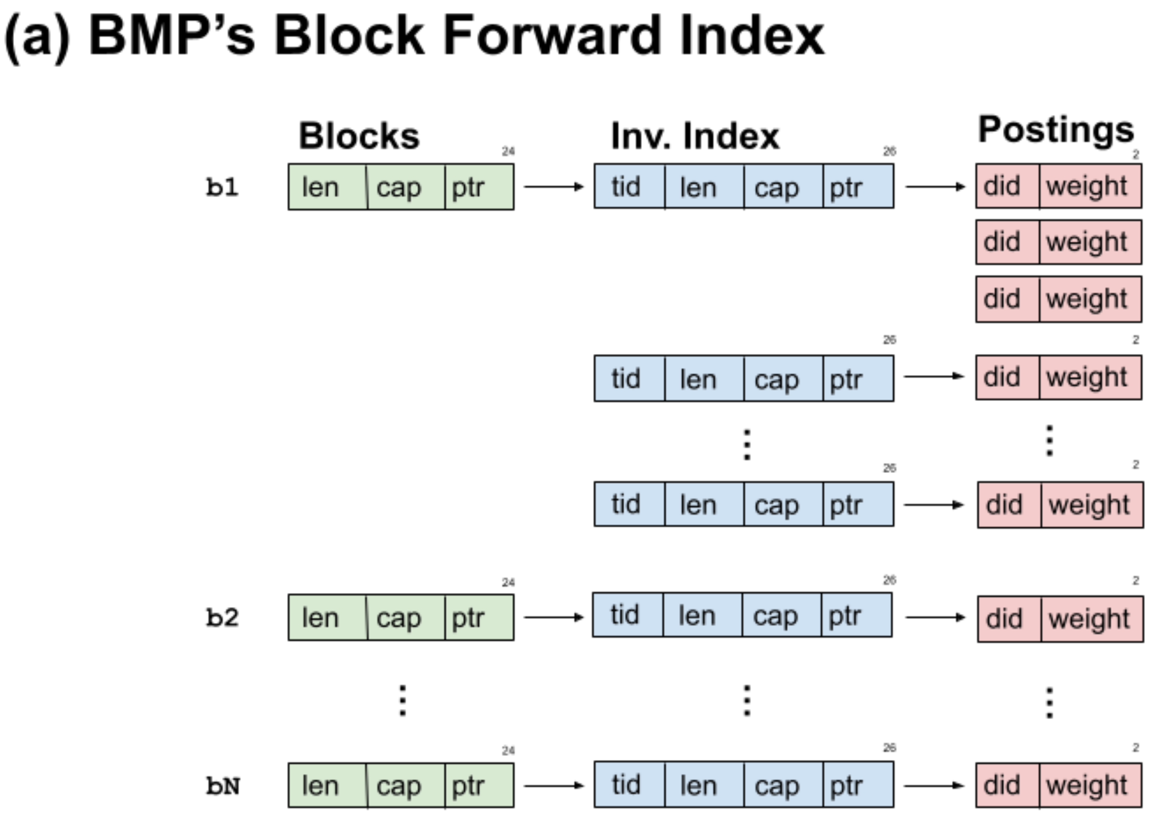
Document Representations

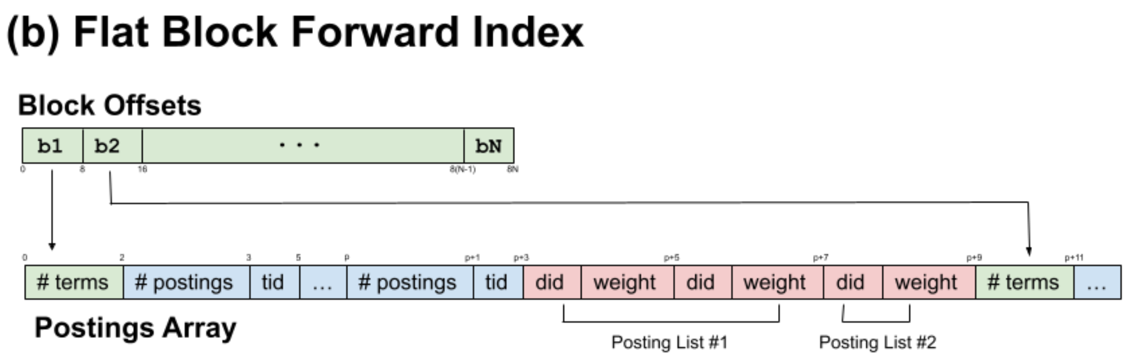


On a 64-bit system, each vector requires 24 bytes!
(len: u64, cap: u64, ptr: u64)

On MS MARCO with 1.1M clusters, vectors require 15 GB, but postings are only 3 GB!

Document Representations





UC SANTA BARBARA

Document Representations — 4 Options

1. BMP-Inv (original)

Document Representations — 4 Options

1. BMP-Inv (original)
2. Compact-Inv (original with less vector overhead)

Document Representations — 4 Options

1. BMP-Inv (original)
2. Compact-Inv (original with less vector overhead)

Assume index is fixed and remove redundant length and capacity.

Assume maximum block size of 256 docs/block (len: u8, ptr: u64)

Limit to 2^16 terms to rightsize termids

Document Representations — 4 Options

1. BMP-Inv (original)
2. Compact-Inv (original with less vector overhead)
3. Flat-Inv (proposed data structure)

Document Representations — 4 Options

1. BMP-Inv (original)
2. Compact-Inv (original with less vector overhead)
3. Flat-Inv (proposed data structure)
4. Forward (Seismic implementation)

Index Size - Document index

4.4M clusters 69K clusters

Block Size	2	4	8	16	32	64	128	256
Document index								
BMP-Inv	28	23	19	15	12	9.6	7.6	6.1
Compact-Inv	22	18	14	12	9.5	7.7	6.3	5.1
Flat-Inv	4.5	4.0	3.6	3.2	**2.9**	**2.6**	**2.4**	**2.4**
Fwd.	**3.1**	**3.1**	**3.1**	**3.1**	3.1	3.1	3.1	3.1

Latency Ablation - Forward vs Inverted Index

	Block Size (b)					
Index Type	8	16	32	64	96	128
99% recall budget						
Flat-Inv	4.29	9.00	17.0	31.3	44.0	**55.6**
Fwd.	**2.43**	**5.30**	**11.4**	**25.8**	**43.1**	60.5
Safe retrieval						
Flat-Inv	11.6	17.0	31.3	56.4	76.5	**91.6**
Fwd.	**9.65**	**11.5**	**22.1**	**48.0**	**76.3**	103

(Super)Block Maximum Weight Compression

1. Dense (BMP, SP)

(Super)Block Maximum Weight Compression

1. Dense (BMP, SP)
2. Sparse (BMP)

(Super)Block Maximum Weight Compression

1. Dense (BMP, SP)
2. Sparse (BMP)
3. SIMDBP w/ random access

(Super)Block Maximum Weight Compression

1. Dense (BMP, SP)
2. Sparse (BMP)
3. SIMDBP w/ random access
4. SIMDBP + quantization (4 bits)

Index Size — Block Maximums

4.4M clusters — 69K clusters

Block Size	2	4	8	16	32	64	128	256
Superblock/block maximum weights								
BMP-Dense	125	63	32	16	8.0	4.0	2.0	1.0
BMP-Sparse	15	9.8	7.6	5.6	4.2	3.0	2.2	1.6
SIMDBP-256*	21	14	9.4	6.0	3.7	2.3	1.3	0.75
+ 4-bit Quant.	**7.4**	**5.3**	**3.5**	**2.4**	**1.5**	**0.95**	**0.59**	**0.33**

Latency & Storage Ablation

	Latency (ms), b=8			Latency (ms), b=128			Space (GB)	
Recall Budget	97%	99%	Safe	97%	99%	Safe	b=8	128
Uncompressed	**0.51**	0.84	**3.3**	12.0	13.3	14.2	51	9.6

UC SANTA BARBARA

Latency & Storage Ablation

Recall Budget	Latency (ms), b=8 97%	99%	Safe	Latency (ms), b=128 97%	99%	Safe	Space (GB) b=8	128
Uncompressed	**0.51**	0.84	**3.3**	12.0	13.3	14.2	51	9.6
Add superblock/block maximum weight compression								
BMP-Sparse	2.7	18	250	73	106	240	26	9.8
SIMDBP-256*	0.83	1.7	6.1	12.2	13.5	14.6	28	8.9
+4-bit Quant.	0.52	0.82	4.7*	7.8	8.3	8.9*	22	8.2

Latency & Storage Ablation

	Latency (ms), b=8			Latency (ms), b=128			Space (GB)	
Recall Budget	97%	99%	Safe	97%	99%	Safe	b=8	128
Uncompressed	**0.51**	0.84	**3.3**	12.0	13.3	14.2	51	9.6
Add superblock/block maximum weight compression								
BMP-Sparse	2.7	18	250	73	106	240	26	9.8
SIMDBP-256*	0.83	1.7	6.1	12.2	13.5	14.6	28	8.9
+4-bit Quant.	0.52	0.82	4.7*	7.8	8.3	8.9*	22	8.2
Add document index compression								
Compact-Inv	0.59	0.95	5.4*	**7.5**	**7.9**	**8.6***	18	6.9
Flat-Inv	0.56	0.85	4.7*	9.4	9.9	10.7*	7.3	**3.2**
Fwd.	0.52	**0.77**	4.8*	9.9	10.6	11.8*	**6.8**	3.7

This work is supported by the
SIGIR Student Travel Grant

UC SANTA BARBARA