Analysis of Basic Data Reordering Techniques

Tan Apaydin – The Ohio State University
Ali Saman Tosun – University of Texas at San Antonio
Hakan Ferhatosmanoglu – The Ohio State University

Overview

- **Goal:** Study the effectiveness of tuple ordering methods on the bitmap compression performance.
- Bitmap Index
  - 0/1 matrix representing the data
  - Logical bit operations (AND, OR)
  - A concise format, easy to compress
- Data warehouses, scientific databases
- Efficient point and range query execution

Bitmap Example

- Storage
- Query Execution
- Full Data Set
- Transformation between equality and range is 1-1.

Tuple	Equality Encoding					Range Encoding				
	Attribute 1		Attribute 2			Attribute 1		Attribute 2		
	a	b	1	2	3	a	b	1	2	3
t_1 = (b, 3)	0	1	0	0	1	0	1	0	0	1
t_2 = (a, 2)	1	0	0	1	0	1	1	0	1	1
t_3 = (a, 3)	1	0	0	0	1	1	0	0	0	1
t_4 = (b, 2)	0	1	0	1	0	0	1	0	1	1
t_5 = (b, 1)	0	1	1	0	0	0	1	1	1	1
t_6 = (a, 1)	1	0	1	0	0	1	1	1	1	1

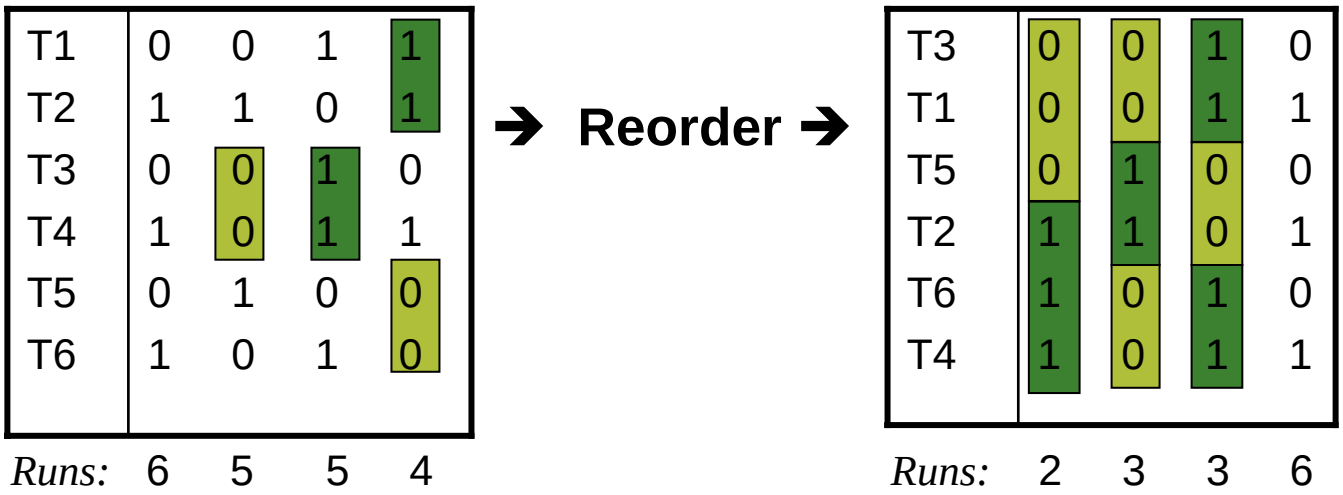
Compressing Bitmap Indices

- Size of bitmap index is still large
- General-purpose compression schemes
- Run-length encoding along the columns
  - Runs of 0s → (0, run-count)
  - Byte-aligned Bitmap Code (BBC) [Oracle '94]
  - Word-Aligned Hybrid Code (WAH) [LBNL '02]
  - No explicit decompression for query processing

Tuple Ordering Improves Compression

Note: Run-Length Compression is applied Column-wise!

Goal: Minimize the hamming distance of adjacent tuples.

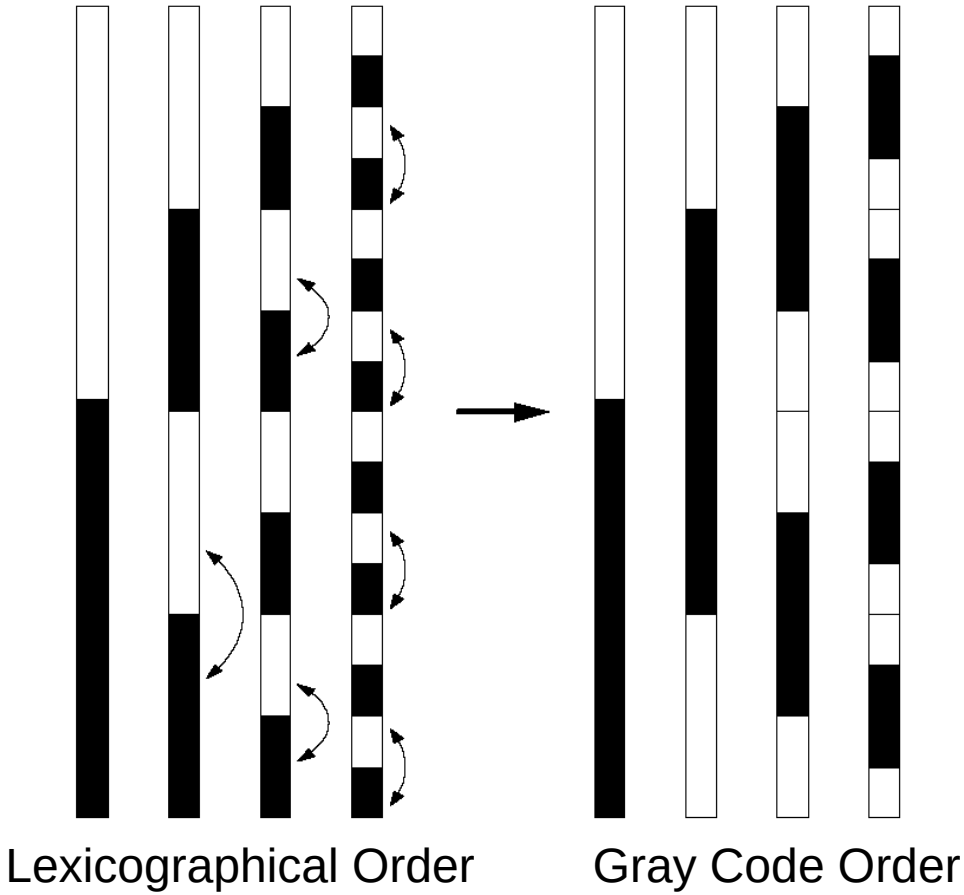


- 20 runs before ordering
- 14 runs after ordering

Tuple Ordering Problem

- NP-Complete
- Most TSP heuristics are ineffective
- **Gray-code**: A space filling-curve for hamming space
  - In-place, linear
  - Improves lossless compression over already compressed bitmaps

Gray Code Ordering



Goal: Less number of runs (better compression).

Analysis Goals (Compression Context)

- Compare Equality and Range Encodings
- Compare Lexicographic and GCO
- Calculate total number of runs in above scenarios

Equality Encoding

- **Theorem 1:** For full data, the number of runs in Lexicographic order of A attributes , where A>1, using equality encoding is:

$$F(C_1) + \sum_{i=2}^{A} \left(F(C_i) \prod_{j=1}^{i-1} C_j - \left[(C_i - 2) \left((\prod_{j=1}^{i-1} C_j) - 1 \right) \right] \right)$$

Define F(x) as F(x) = 3x.

- Proof in the paper.

- **Theorem 2:** Number of runs in GCO using equality encoding is equal to the number of runs in Lexicographic order.

Range Encoding for Lexicographic Order

- **Theorem 3:** For full data, the number of runs in Lexicographic order of A attributes, where A>1, using range encoding is:

$$E(C_1) + \sum_{i=2}^{A} \left(E(C_i) \prod_{j=1}^{i-1} C_j - \left[(\prod_{j=1}^{i-1} C_j) - 1 \right] \right)$$

Define E(x) as E(x) = 2x - 1.

- **Corollary 1 (Equality Lexico vs. Range Lexico):** For Lexicographic order of full data, range encoding produces fewer runs than equality encoding.

Range Encoding for GCO

- **Theorem 4:** For full data, the number of runs in GCO of A attributes, where A>1, using range encoding is:

$$E(C_1) + \sum_{i=2}^{A} \left(E(C_i) \prod_{j=1}^{i-1} C_j - C_i \left[\left(\prod_{j=1}^{i-1} C_j \right) - 1 \right] \right)$$

E(x) = 2x - 1.

- **Corollary 2 (Equality GCO vs. Range GCO):** For GCO of full data, range encoding produces fewer runs than equality encoding.

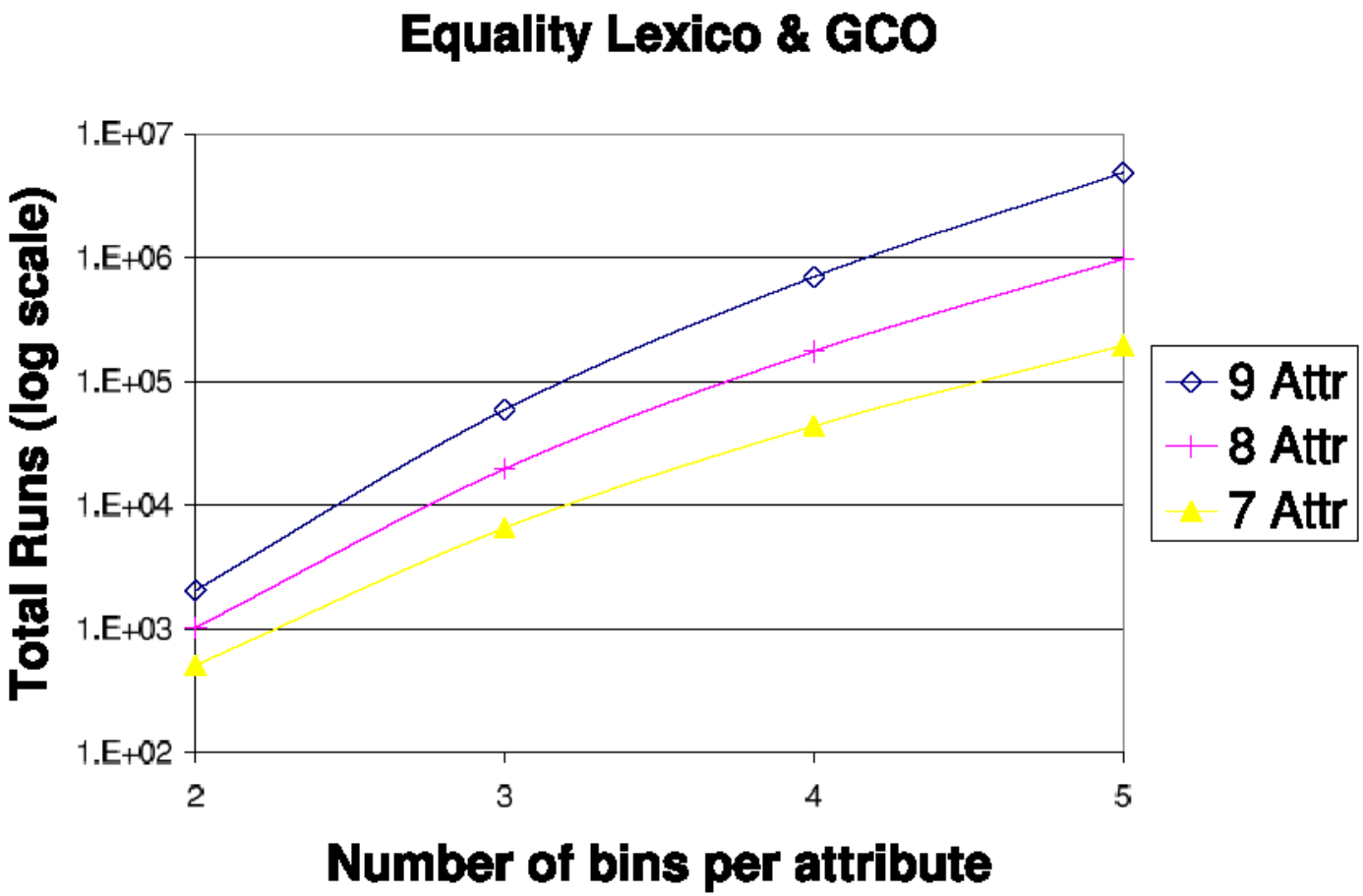
Lexicographic Order vs. GCO in Range Encoding

- **Corollary 3 (Range Lexico vs. Range GCO):** For range encoding with full data, GCO produces fewer number of runs than Lexicographic order.
- In fact, we were able to prove that GCO is optimum for range encoding.
- Ongoing work: An optimum and fast reordering algorithm suited for equality encoding.

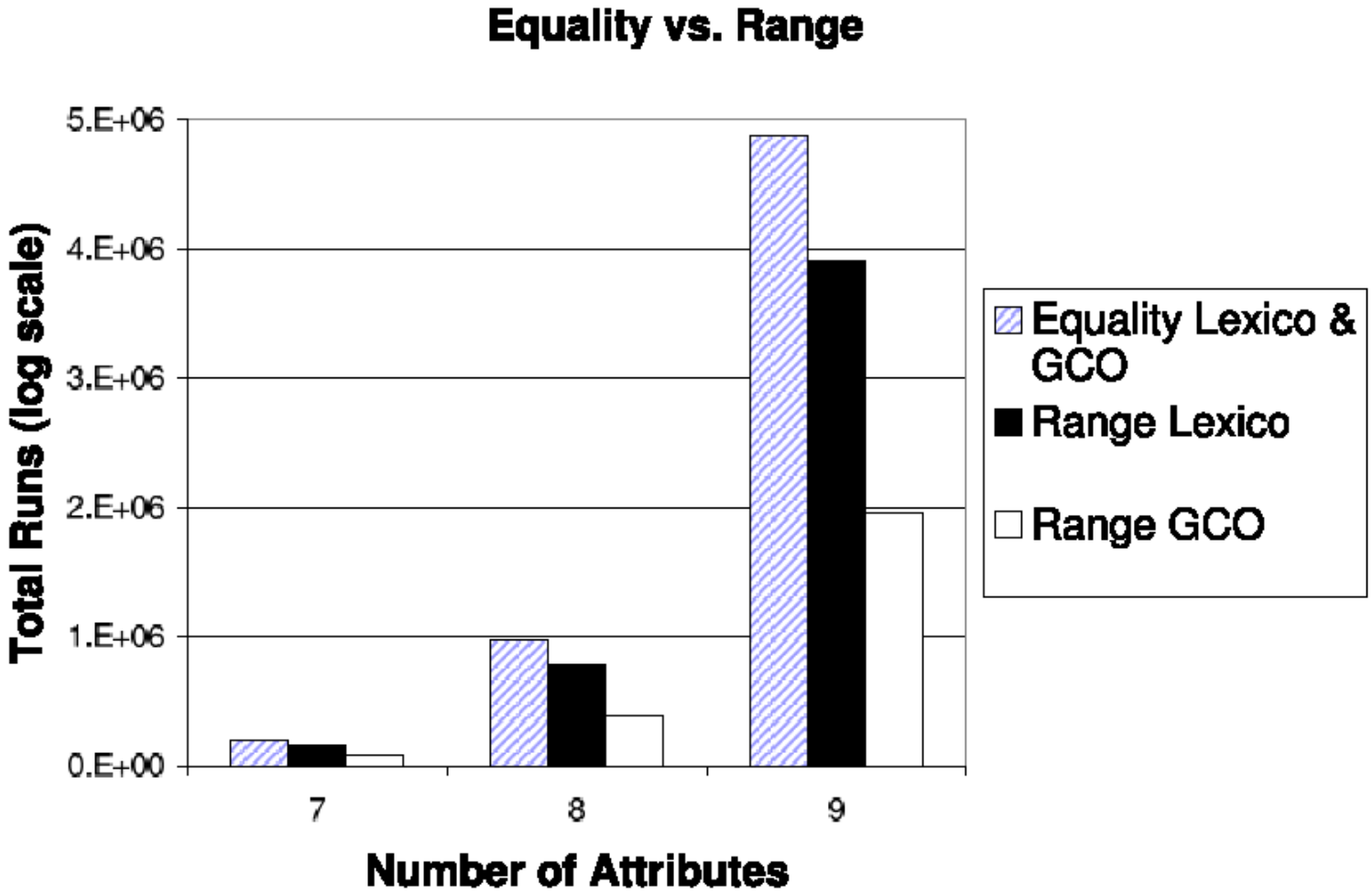
Experiments

- Data sets: used full data sets with varying
  - number of attributes
  - cardinality of attributes

Equality Encoding



Equality Encoding vs. Range Encoding



Conclusion

- Studied effectiveness of tuple reordering methods on compression performances.
- Theoretical foundations and performance analysis of lexicographic order and GCO.
- Two encodings:
  - Equality and Range
- Range encoding provides better compression both for Lexico and GCO.
- GCO is optimum for range encoding.

Questions and Comments

- Thank you!

Email: apaydin@cse.ohio-state.edu