LEXIDB: A SCALABLE CORPUS DATABASE MANAGEMENT SYSTEM

MATTHEW COOLE, PAUL RAYSON, JOHN MARIANI

BACKGROUND

Corpora have grown from millions to billions of words in recent years.

Brown Corpus (1961) ~1 million words

BNC (1994) ~100 million words

Historical Hansard (2005) ~1.68 billion words

EEBO-TCP ~4 billion words

Simple tool and concordancers e.g. AntConc, Wmatrix etc. cannot handle this scale.

LEXIDB

Lightweight distributed corpus DBMS.

Supports 4 query types;

- Concordance Lines
- Collocations
- Clusters (N-Grams)
- Word lists

Supports corpora in the range of ~billions of words.

Token level annotations supported.

LEXIDB ARCHITECTURE

All nodes in network utilised as peers.

Data split into regions to allow easy migration and data balancing.

Uses a column-family store designed for zipfian data.

Full text index of both text and annotation to support regular expressions.

DEMO

BENCHMARK SETUP

2 corpora

- Historical Hansard (1.68 billion words)
- EEBO-TCP Phase 1 (0.91 billion words)

AWS test system (8 vCPUs, 30GB RAM, 2 x 80GB SSD) – 1,2 & 4 node configurations

BENCHMARKS RESULTS

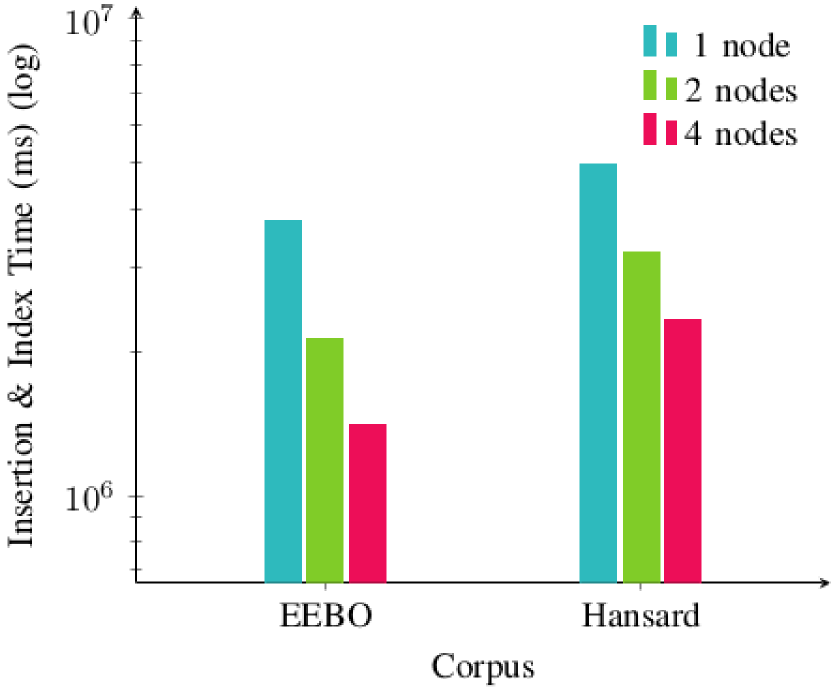

Fig. 3. Insertion and Indexing

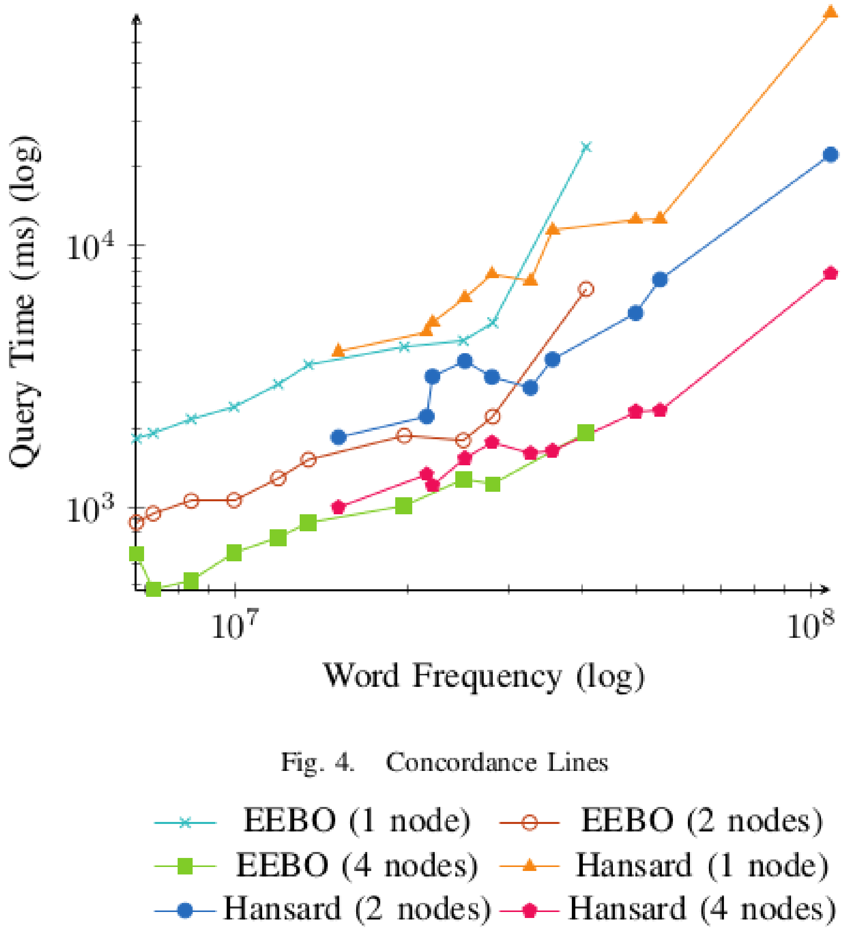

Fig. 4. Concordance Lines

BENCHMARKS RESULTS (2)

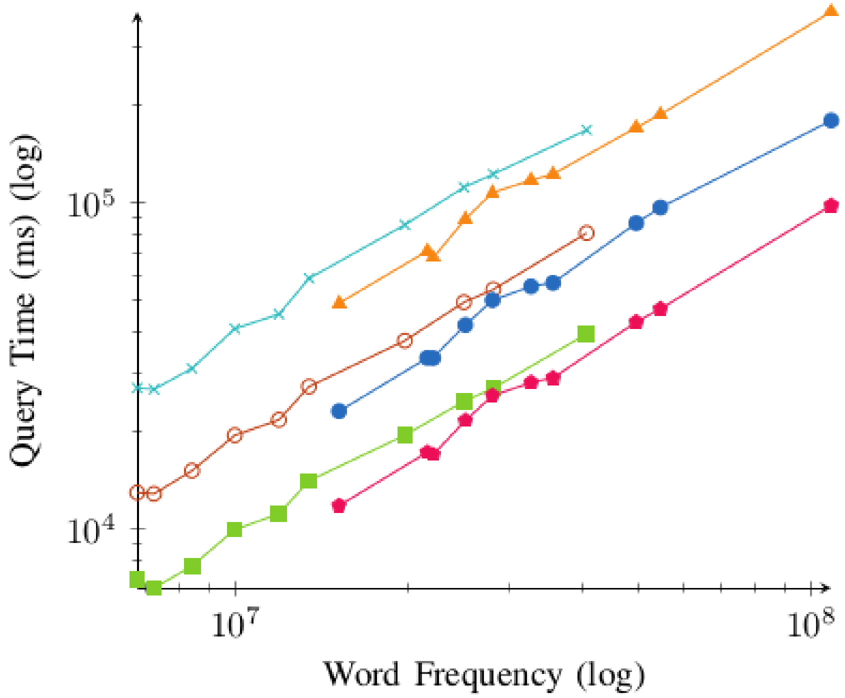

Fig. 5. Collocations

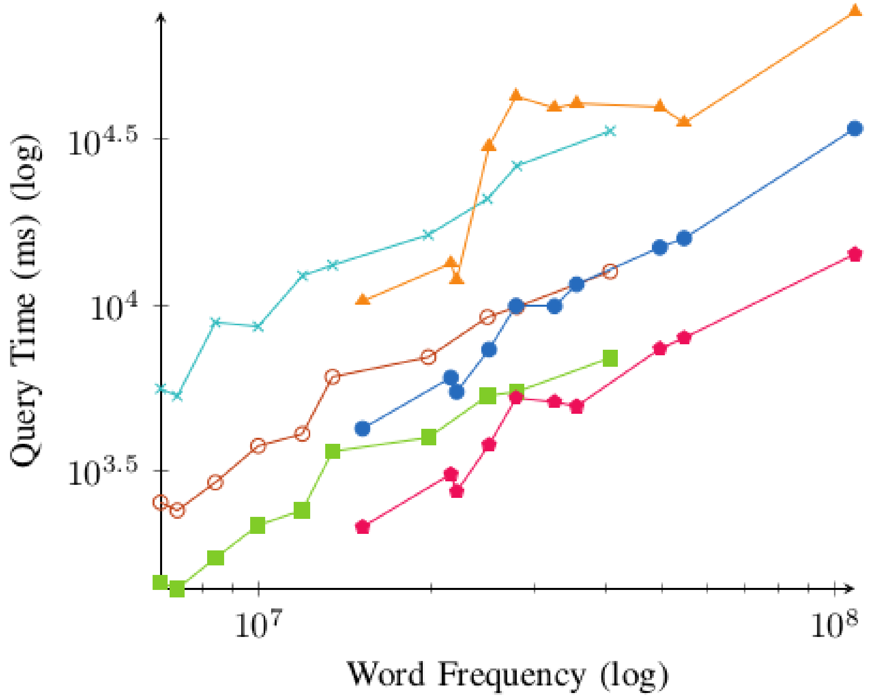

Fig. 6. Clusters (n-grams)

BENCHMARK RESULTS (3)

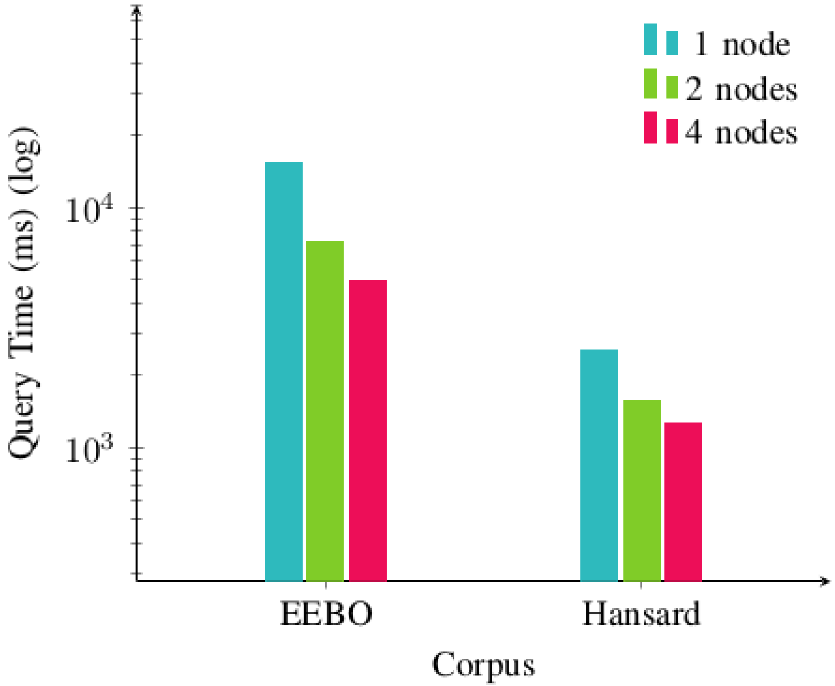

Fig. 7. Frequency List

FUTURE WORK

Chord (DHT) for more robust scalability.

Support for further column families to support metadata.

Query language extension to move towards a simplified CQL.

ACKNOWLEDGEMENTS

Research funded by EPSRC doctoral training grant.

Corpus Data Project (CDP) – Laurence Anthony, Stephen Wattam, John Vidler.

Code available at; https://github.com/matthewcoole/lexidb