Outline: Indexing Methods for Web Archives

March 4, 2013

1 Introduction - (5 pages)

Motivate FT search on Archives and need for specific workloads. Argue that these workloads are important as "primitive" operations which enable various text and time analysis and extraction tasks. Present use cases for each scenarios which will be presented as motivations to more-or-less concrete research questions later.

1.1 Research Questions - (1 - 1.5 pages)

Each of these correspond to a core chapter in the thesis.

- Efficient Indexing and Maintenance for Temporal Queries
- Enabling approximate results for temporal queries
- Supporting efficient phrase queries

1.2 Contributions and Publications - (1 page)

List contributions as solutions to the research questions and the correspoinding publications.

1.3 Outline of Thesis (0.5 page max)

2 Foundations and Technical Background - (20-25 pages)

2.1 Web Archiving (1 - 1.5 page)

Keeping it general and prsent an overview rather than details.

2.2 Information Retrieval (10 pages)

Fundamentals of IR, scoring model, ranking, evaluation(???), link analysis, doc. ordered vs score ordered index lists.

2.2.1 Indexing Text - (7 pages)

- Introduce inverted indexes, dictionaries, data structures
- Explain different payloads for postings, query semantics
- Construction of inverted indexes etc.

2.2.2 Query Processing Techinques - (1-2 pages)

DAAT, TAAT, WAND, NRA, TA, CA etc.

2.2.3 Phrase Queries - (2 pages)

Auxillary indexes. Bi-gram indexes etc.

2.2.4 Handling index updates - (2 pages)

Index update schemes - geometric, logarithmic, query-log based etc.

2.2.5 Compression and Caching (2 pages)

Compression techniques based on gaps. explain different Do we need caching at all ? or do I introduce it in the context of phrase queries later.

2.3 Data Management

2.3.1 Temporal Databases - (1 page)

Plan to keep it short and concise.

2.4 KMV Synopsis (1 page)

2.5 Indexing Archives - (7-10 pages)

2.5.1 Time-travel Queries and indexing

Content about TTIX and initial notation which will be re-used throughout the dissertation.

2.5.2 Compression in indexes for archives

Coalescing and mostly work from Jinru He. Other related works on compressing document collections.

3 Query Optimization for approximate queries (23 pages)

3.1 Introduction and Problem statement (3 pages)

3.2 Related Work (0.5 - 1 page)

3.3 Model(0.5 page)

3.4 Index Organization(1 page)

Synopsis Index and the vertically partitioned index.

3.5 Partition Selection (1 page)

Partition Selection as a query optimization step for approximate queries and formal problem statement.

3.6 Single-Term Partition Selection (5 pages)

3.6.1 Size-based Partition

3.6.2 Equi-cost Partition Selection

3.7 Multi-Term Partition Selection (4 pages)

3.7.1 Size-based Partition (3 pages)

3.7.2 Equi-cost Partition Selection

3.8 Pratical Issues and Optimizations (1 pages)

3.8.1 Cost Model and Heuristic algorithm

3.9 Experimental Evaluation (6 pages)

3.10 Summary

4 Efficient Indexing and Maintenance for Temporal Queries - (30-35 pages)

Presents index sharding as an alternate index organization way for temporal queries. Discusses alternate sharding methods. Addresses index maintenance and presents experiments to justify claims.

4.1 Introduction and Problem statement (2 pages)

4.1.1 Approach

4.1.2 Organization

4.2 Related Work (1 page)

4.3 Model (0.5 page)

4.4 Index Organization (1 page)

4.5 Index Sharding (1 pages)

4.6 Idealized Index Sharding (3 pages)

4.6.1 Algorithm and Proof of Optimality

4.7 Cost-Aware Shard Merging (3 pages)

4.7.1 Cost Model and Heuristic algorithm

4.8 Index Maintenance (2 pages)

Introduce index maintenance and extra notation as required.

4.9 Incremental Sharding (5 pages)

4.9.1 Algorithm and Proof of Approximation guarantee

4.10 System Architecture (1 pages)

4.11 Experimental Evaluation(10 pages)

4.12 Summary

5 Phrase Indexing and Querying - (23 pages)

5.1 Introduction (1 page)

5.2 Model and Indexing Organization (3 pages)

5.3 Related Work (1 page)

5.4 Query Optimization (4 pages)

5.4.1 Optimal Solution

5.4.2 Approximation Guarantee

5.5 Phrase Selection (4 pages)

5.5.1 Query-Optmizer-Based Selection

5.5.2 Coverage-Based Selection

5.6 Experimental Evaluation (7 pages)

Add experiments for robustnes and final plots for wall-clock-times.

5.7 Summary

6 Conclusion - (2-3 pages)

7 Appendix

Queries. Examples from the output of the query optimizer from Phrase querying.