

DAIS Qualifying Examination Reading List

I. Information Retrieval

- **Basic concepts**
 - Vector-space retrieval model, TF-IDF weighting, relevance/pseudo feedback, (non-interpolated) average precision, query-likelihood retrieval model, language model smoothing, PageRank, inverted index
- **Background**
 - Modern Information Retrieval: A Brief Overview, Singhal, IEEE Data Engineering Bulletin 24(4), pages 35-43, 2001 ([ps](#))
 - Link Analysis in Web Information Retrieval, Henzinger, IEEE Data Engineering Bulletin 23 (3), pages 3-8, 2000 ([pdf](#))
 - Probabilistic relevance models based on document and query generation, Lafferty and Zhai, Language Modeling and Information Retrieval, Kluwer International Series on Information Retrieval, Vol. 13, 2003 ([ps](#))
 - A study of smoothing methods for language models applied to information retrieval, Zhai and Lafferty, ACM Transactions on Information Systems, Vol. 2, No. 2, pp. 179-214, April 2004 ([pdf](#))
- **More advanced topics**
 - Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis, Hofmann, SIGIR 2003 ([pdf](#))
 - A Formal Study of Information Retrieval Heuristics, Fang et al., SIGIR 2004 ([pdf](#))
 - TeXQuery: A Full-Text Search Extension to XQuery, Amer-Yahia et al., WWW 2004 ([pdf](#))
 - Corpus structure, language models, and ad hoc information retrieval, Kurland and Lee, SIGIR 2004 ([pdf](#))

II. Data Mining and Data Warehousing

- **Basic concepts**
 - Data warehousing: star schema, data cube (be able to list half a dozen typical data cube computation methods), multi-dimensional analysis (OLAP)
 - Data mining: frequent pattern mining (be able to list half a dozen typical methods), sequential pattern mining (be able to list half a dozen typical methods), correlation analysis, classification (be able to list half a dozen typical methods), clustering (be able to list half a dozen typical methods)
- **Background**
 - Data Mining: Concepts and Techniques, 2nd edition, Han and Kamber, Morgan Kaufmann Publishers, 2005. Chapters 3 & 4 (for data warehousing); Chapters 2, 5-7 (for data mining). (Prepublication chapters can be found on the web in the CS498Han Class Notes.)
- **More advanced topics**
 - Data warehousing

- Discovery-driven exploration of OLAP data cubes, Sarawagi, Agrawal, and Megiddo, EDBT 1988 ([pdf](#))
- Multi-Dimensional Regression Analysis of Time-Series Data Streams, Chen et al., VLDB 2002 ([pdf](#))
- Data mining
 - Approximate Frequency Counts over Data Streams, Manku and Motwani, VLDB 2002 ([pdf](#))
 - Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach, Pei et al., IEEE Transactions on Knowledge and Data Engineering, 16(10), 2004 ([pdf](#))
 - A Framework for Clustering Evolving Data Streams, Aggarwal et al., VLDB 2003 ([pdf](#))

III. Databases

- **Basic concepts**
 - Hardware: disk sector, track, block, seek, latency, how to lay out a database page
 - Data modeling: ER, OO, and Object-Relational approaches
 - Concurrency control and recovery: ACID, serializability, two-phase locking, two-phase commit, logging and recovery, the impact of data replication
 - Theory: normalization, dependencies
 - Queries: access methods (hashing, B-trees, multidimensional access methods), how to optimize a query, SQL
 - Benchmarks: TPC-C and TPC-H
- **Background**

You can use any database textbook you like to study the most basic of the concepts listed above; for example, CS411 teaches these concepts. (Note that you will be expected to be able to demonstrate your understanding of the concepts by applying them (as opposed to simply being able to define them).) In the remaining entries, “RDS” refers to Stonebraker’s *Readings in Database Systems*, currently in its 4th edition.

- Generalized Search Trees for Database Systems, Hellerstein et al., VLDB 1995 and RDS ([pdf](#)). We include this paper as the reference for multidimensional access methods; access methods based on B-trees and hashing should be covered in any database textbook.
- New TPC Benchmarks for Decision Support and Web Commerce, Poess and Floyd, SIGMOD Record 29(4), December 2000 ([pdf](#))
- Inclusion of New Types in Relational Data Base Systems, Stonebraker, ICDE 1986 and RDS ([pdf](#)). We include this paper as your reference for understanding the impact of extensibility (as, for example, intended by the object-relational model) on a DBMS.
- **More advanced topics**

Please note that databases are a very broad field. The papers listed here will be changed frequently, to reflect this breadth.

- *Query processing*
 - AutoAdmin 'What-if' Index Analysis Utility, Chaudhuri and Narasayya, SIGMOD 1998 and RDS ([pdf](#))
 - Optimal Aggregation Algorithms for Middleware, Fagin et al., PODS 2001 ([pdf](#))
- *Parallelism*
 - Parallel Database Systems: The Future of High-Performance Database Processing; DeWitt and Gray, CACM 35(6), 1992; and RDS ([pdf](#))
- *Security trends*
 - Hippocratic Databases, Agrawal et al., VLDB 2002 ([pdf](#))
- *Information integration*
 - Applying Model Management to Classical Meta Data Problems, Bernstein, CIDR 2003 and RDS ([pdf](#))