

CMSC 724*Reading List

Sudarshan S. Chawathe[†]

Spring 2001

The following list indicates the reading due before the indicated class meeting. **This document will change.**

You should be able to find most of these papers very easily on the Web. In some cases, I have put links to local copies of papers. The ACM Digital Library¹ is a very good source for papers, especially recent ones. (The University has a site subscription so access is free from the umd.edu domain.) The Computer Science Library² also has a good collection of conference proceedings and journals. In most cases, you can download papers from the Web sites of their authors. (You should be able to locate such resources using your favorite search engine.) If you have trouble locating any papers, let me know.

You are required to read the material indicated below *before* the class meeting at which it is due so that you can actively participate in the discussion. You should read the papers critically, noting, for example, the advantages and limitations of the proposed methods. You should be prepared to both ask and answer questions intelligently. The class participation portion of your grade depends on such interactions. More importantly, if you do not do the readings before class, you will not benefit from the classroom discussions (which will assume you have read the material carefully).

Tue., 30 Jan. Topic: introduction; overview of main topics; review of some CMSC 424 material (Datalog, SQL3, algebra, etc.).

Thu., 01 Feb. Topic: Lore and Lorel (semistructured database systems). Material due:

Two papers on Lore and Lorel: [MAG⁺97, AQM⁺96].

UnQL: [BDHS96].

Homework 1

Optional reading: [MW99, AMR⁺98, STH⁺99].

Tue., 06 Feb. Topic: Basics of Information Retrieval. Material due:

*<http://www.cs.umd.edu/class/spring2001/cmssc724/>

[†]<http://www.cs.umd.edu/users/chaw/>

¹<http://www.acm.org/dl/>

²<http://www.cs.umd.edu/Library>

Access Methods for Text [Fal85]: Local copy

The SIFT Information Dissemination System [YGM00]: Local copy

Thu., 08 Feb. Topic: Application of Datalog to Information Integration. Material due:

Information Integration Using Logical Views [Ull97]: Local copy

Theory of Answering Queries using Views [Hal00]: Local copy

Tue., 13 Feb. Topic: Standard query processing techniques. Material due:

Query Evaluation Techniques for Large Databases [Gra93]: Local copy. Much of this paper should be very familiar if you have taken CMSC 624. If you haven't, you'll take longer to read the paper. This paper is very long, so please start reading it well before it is due! You may also want to read a review³ (no endorsement implied) of this paper in the ACM Digital Library⁴.

Thu., 15 Feb. Topic: Standard query processing and database systems issues continued.

Tue., 20 Feb. Topic: Clustering, with a focus on hierarchical clustering. Material due:

Textbook: Chapter 8 (Cluster Analysis) of [HK01].

Papers: BIRCH [ZRL96] (Local copy) and CURE [GRS98] (Local copy).

Optional papers: ROCK [GRS99] (Local copy); Chameleon [KHK99] (Local copy); longer version of the BIRCH paper [ZRL97] (Local copy).

Thu., 22 Feb. Topic: Conjunctive Queries. Material due:

Chapter 4 of [AHV95]

Tue., 27 Feb. Topic: Describing Semistructured Data. Material due:

Textbook: Chapter 7 of [ABS99].

Papers: Representative Objects [NUWC97] (Local copy); Graph Schemas [BDFS96] (Local copy); DataGuides [GW97] (Local copy).

Optional papers: Typing using description logic [CGL98] (Local copy). In addition, there is a large and dynamic collection of schema proposals for XML. (Look for terms like XML-Data, RDF, and XML-Schema at the W3C Web site⁵.)

Thu., 01 Mar. Topic: XML Query Languages. Material due:

Textbook: Chapters 4-6 of [ABS99].

Papers revisited: We will briefly discuss these papers again in the light of the textbook's explanation.

³<http://www.acm.org/pubs/citations/journals/surveys/1993-25-2/p73-graefe/>

⁴<http://www.acm.org/dl/>

⁵<http://www.w3.org/>

Two papers on Lore and Lorel: [MAG⁺97, AQM⁺96].
UnQL: [BDHS96].

Tue., 06 Mar. Topic: First-order queries. Material due:

Textbook: Chapter 5 of [AHV95].

Thu., 08 Mar. Topic: Guest lecture.

Tue., 13 Mar. Topic: Theoretical perspective on optimization of first-order queries. Material due:

Textbook: Chapter 6 of [AHV95].

Spring break: 17th through 25th March; no class meetings.

Tue., 27 Mar. Topic: Evaluation of Datalog. Material due:

Textbook: Chapters 12 and 13 of [AHV95].

Thu., 29 Mar. Midterm (take-home) assigned. **Completed midterms due in AVW 4179 by 11:00am on Monday, 02 April 2001.** No class.

Tue., 03 Apr. Evaluation of Datalog (catchup); Chapter 13 of [AHV95].

Thu., 05 Apr. Mining Association Rules. Material due:

Textbook: Chapter 6 of [HK01].

Tue., 10 Apr. Classification and Prediction. Material due:

Textbook: Chapter 7 of [HK01].

Thu., 12 Apr. Attend colloquium 4:00pm–5:00pm in AVW 3258. Material due:

Papers: Two papers on index selection in OLAP [GHR⁺97] Local copy; [GM99] Local copy.

Tue., 17 Apr. Topic: Recursion and Negation (together). Material due:

Textbook: Chapter 14 of [AHV95].

Thu., 19 Apr. Topic: Recursion and Negation (together), continued; Expressiveness and Complexity. Material due:

Textbook: Chapters 15 and 16 of [AHV95].

Tue., 24 Apr. Topic: Beyond First-Order Queries. Material due:

Textbook: Chapters 17 and 18 of [AHV95].

Thu., 26 Apr. Attend colloquium 4:00pm–5:00pm in AVW 3258. Material due:

Papers: [ALU01] Local copy; [LBU01] Local copy.

Tue., 01 May. Catch-up and wrap-up of main topics.

Thu., 03 May. Material due:

Preliminary term project report

Tue., 08 May. Material due:

Term project presentation and discussion

Thu., 10 May. Material due:

Term project presentation and discussion

Tue., 15 May. Material due:

Final term project report

Fri., 18 May. Take-home final exam assigned.

Fri., 25 May. Completed final exam due in AVW 4179 by 11:00am on Friday, 25 May 2001.

1 Books

Textbooks

Data on the Web: From Relations to Semistructured Data and XML [ABS99]. We'll cover most of the material in this book, and use the topics as a launch pad into more detailed investigations in other areas.

Foundations of Databases [AHV95]. We will cover selected topics from this book, using it as the main reference for Database Theory.

Data Mining: Concepts and Techniques [HK01]. This book presents an overview of the current state of Data Mining research. We will cover a few chapters of this book, using papers to supplement the material.

Reference Books

Modern Information Retrieval [BYRN99]. Use this book for an overview of Information Retrieval. The huge list of references is a big plus.

A First Course in Database Systems [UW97]. This is the textbook I currently use for CMSC 424, and covers most of the user-level database issues. It includes easily digestible chapters on OODBs (ODL/OQL) and Datalog, which are topics often not covered in introductory database classes.

Database System Implementation [GMUW00]. This book is a good one if you need to brush up on basic database implementation topics covered in CMSC 624 (e.g., query optimization, concurrency control, recovery).

Readings in Database Systems [SH98]. This collection of papers is typically covered in CMSC 624 and similar courses. It includes many famous papers, such as “the System R paper,” “the ARIES paper,” and Gray et al.’s locking paper.

Principles of Distributed Database Systems [OV99]. Look here for distributed query optimization, distributed transaction processing, etc.

2 Resources

- The ACM Digital Library⁶: Requires a subscription, but UMD has a site-wide subscription that gives access from all local machines.
- The DBLP Bibliography Server⁷ has extremely good coverage of the Database and Logic Programming fields.
- ACM SIGMOD⁸.
- VLDB Foundation⁹.
- SIGMOD Record¹⁰
- IEEE Data Engineering Bulletin¹¹
- Maryland Database Group¹² with pointers to other relevant DB resources.

References

- [ABS99] Serge Abiteboul, Peter Buneman, and Dan Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann, first edition, October 1999.
- [AHV95] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

⁶<http://www.acm.org/dl/>

⁷<http://www.purl.org/net/dblp>

⁸<http://www.acm.org/sigmod/>

⁹<http://www.vldb.org>

¹⁰<http://www.acm.org/sigmod/record/>

¹¹<http://www.research.microsoft.com/research/db/debull>

¹²<http://www.cs.umd.edu/areas/db/>

- [ALU01] Foto Afrati, Chen Li, , and Jeff Ullman. Generating efficient plans for queries using views. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Santa Barbara, California, May 2001. To appear.
- [AMR⁺98] S. Abiteboul, J. McHugh, M. Rys, V. Vassalos, and J. Wiener. Incremental maintenance for materialized views over semistructured data. In *Proceedings of the Twenty-second International Conference on Very Large Data Bases*, New York, New York, 1998.
- [AQM⁺96] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. Wiener. The Lorel query language for semistructured data. *Journal of Digital Libraries*, 1(1):68–88, November 1996.
- [BDFS96] P. Buneman, S. Davidson, M. Fernandez, and D. Suciu. Adding structure to unstructured data. Technical Report MS-CIS-96-21, University of Pennsylvania, Computer and Information Science Department, 1996.
- [BDHS96] P. Buneman, S. Davidson, G. Hillebrand, and D. Suciu. A query language and optimization techniques for unstructured data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 505–516, Montréal, Québec, June 1996.
- [BYRN99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, first edition, May 1999.
- [CGL98] D. Calvanese, G. Giacomo, and M. Lenzerini. What can knowledge representation do for semi-structured data? In *Proceedings of the National Conference on Artificial Intelligence*, 1998.
- [Fal85] Christos Faloutsos. Access methods for text. *ACM Computing Surveys*, 17(1):49–74, 1985.
- [GHR⁺97] H. Gupta, V. Harinarayan, A. Rajaraman, , and J. D. Ullman. Index selection for OLAP. In *Proceedings of the International Conference on Data Engineering*, Birmingham, U.K., April 1997.
- [GM99] H. Gupta and I.S. Mumick. Selection of views to materialize under a maintenance-time constraint. In *Proceedings of the International Conference on Database Theory*, Jerusalem, Israel, January 1999.
- [GMUW00] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database System Implementation*. Prentice-Hall, Upper Saddle River, New Jersey, 2000.
- [Gra93] Goetz Graefe. Query evaluation techniques for large databases. *ACM Computing Surveys*, 25(2):73–169, 1993.
- [GRS98] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 73–84, Seattle, Washington, June 1998.

- [GRS99] S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. In *Proceedings of the International Conference on Data Engineering*, pages 512–521, Sydney, Australia, March 1999.
- [GW97] R. Goldman and J. Widom. DataGuides: Enabling query formulation and optimization in semistructured databases. In *Proceedings of the Twenty-third International Conference on Very Large Data Bases*, Athens, Greece, 1997.
- [Hal00] Alon Y. Halevy. Theory of answering queries using views. *SIGMOD Record*, 29(4), December 2000.
- [HK01] Jiawei Han and Micheline Kamber. *Data Mining: concepts and techniques*. Morgan Kaufmann, San Francisco, California, 2001.
- [KHK99] G. Karypis, E.-H. Han, and V. Kumar. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *IEEE Computer*, 32(8):68–75, 1999.
- [LBU01] Chen Li, Mayank Bawa, , and Jeff Ullman. Minimizing view sets without losing query-answering power. In *Proceedings of the International Conference on Database Theory*, London, U.K., January 2001.
- [MAG⁺97] J. McHugh, S. Abiteboul, R. Goldman, D. Quass, and J. Widom. Lore: A database management system for semistructured data. *SIGMOD Record*, 26(3):54–66, September 1997.
- [MW99] Jason McHugh and Jennifer Widom. Query optimization for XML. In *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB)*, pages 315–326, Edinburgh, Scotland, U.K., September 1999.
- [NUWC97] S. Nestorov, J. Ullman, J. Wiener, and S. Chawathe. Representative objects: Concise representations of semistructured, hierarchical data. In *Proceedings of the International Conference on Data Engineering*, pages 79–90, 1997.
- [OV99] M. Tamer Ozsu and Patrick Valduriez. *Principles of Distributed Database Systems*. Prentice-Hall, Upper Saddle River, New Jersey, second edition, 1999.
- [SH98] M. Stonebraker and J. Hellerstein, editors. *Readings in Database Systems*. Morgan Kaufmann, San Francisco, California, third edition, 1998.
- [STH⁺99] Jayavel Shanmugasundaram, Kristin Tufte, Gang He, Chun Zhang, and David DeWitt and Jeffrey Naughton. Relational databases for querying XML documents: Limitations and opportunities. In *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB)*, pages 302–314, Edinburgh, Scotland, U.K., September 1999.
- [Ull97] Jeffrey D. Ullman. Information integration using logical views. In *Proceedings of the International Conference on Database Theory*, 1997.

- [UW97] J. D. Ullman and J. Widom. *A first course in database systems*. Prentice-Hall, Upper Saddle River, New Jersey, 1997.
- [YGM00] Tak Yan and Hector Garcia-Molina. The SIFT information dissemination system. Technical report, Stanford University Database Group, 2000. Available at <http://www-db.stanford.edu/pub/papers/main.ps> .
- [ZRL96] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 103–114, Montreal, Canada, June 1996.
- [ZRL97] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–181, 1997.