

CMSC 724*Reading List

Sudarshan S. Chawathe[†]

Spring 2002

The following list indicates the reading due before the indicated class meeting. **This document will change, so please check it often.**

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).

1 Schedule

This schedule is only a rough outline and the actual schedule will depend on how quickly we cover material, feedback from the class, and other factors. In particular, the exams will be scheduled after a few class meetings.

01 Feb 2002:

08 Feb 2002: Topic: Standard query processing techniques.

*<http://www.cs.umd.edu/class/spring2002/cmssc724/>

†<http://www.cs.umd.edu/users/chaw/>

¹<http://www.acm.org/dl/>

²<http://www.cs.umd.edu/Library>

Query Evaluation Techniques for Large Databases [Gra93]: Local copy. This paper presents a very good overview of standard query processing techniques but is very long, so please start reading it well before it is due.

15 Feb 2002: Conjunctive and First Order Queries.

Chapters 4 and 5 of [AHV95]

22 Feb 2002: XML Query Languages.

Chapters 4-6 of [ABS99]

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

UnQL: [BDHS96].

01 Mar 2002: Clustering, Classification, and Prediction.

Chapters 7 and 8 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).

08 Mar 2002: Structure Extraction.

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 site³.)

15 Mar 2002: Datalog: Evaluation and Applications.

Textbook Chapters 12 and 13 of [AHV95]

Information Integration Using Logical Views [Ull97]: Local copy

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

22 Mar 2002: Recursion and Negation; Expressiveness; Complexity.

Chapters 14, 15, and 16 of [AHV95]

29 Mar 2002: Spring break; no class meeting.

05 Apr 2002: To be decided...

12 Apr 2002:

³<http://www.w3.org/>

19 Apr 2002:

26 Apr 2002:

03 May 2002:

10 May 2002:

10 May 2002:

18 May 2002: Official (university) final exam date; actual final exam schedule TBA.

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 Library⁴: Requires a subscription, but UMD has a site-wide subscription that gives access from all local machines.
- The DBLP Bibliography Server⁵ has extremely good coverage of the Database and Logic Programming fields.
- ACM SIGMOD⁶.
- VLDB Foundation⁷.

⁴<http://www.acm.org/dl/>

⁵<http://www.purl.org/net/dblp>

⁶<http://www.acm.org/sigmod/>

⁷<http://www.vldb.org>

- SIGMOD Record⁸
- IEEE Data Engineering Bulletin⁹
- Maryland Database Group¹⁰ 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.
- [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.
- [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.

⁸<http://www.acm.org/sigmod/record/>

⁹<http://www.research.microsoft.com/research/db/debull>

¹⁰<http://www.cs.umd.edu/areas/db/>

- [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.
- [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.
- [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 Ozsü 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.
- [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.
- [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.