

# CMSC 724 Reading List

Sudarshan S. Chawathe

Spring 2004

Table 1 outlines the planned schedule up to the mid-term. Post-midterm material will be outlined a few weeks into the semester, and will depend on student feedback. The schedule is **approximate** and **likely to change**. *Do not assume any date or event as definite unless you check with me first.* A listing of the planned units follows.

1. **Course overview; introduction to query processing.** Overview of course topics in general. How to implement a very simple SQL query engine and how to improve it.
2. **Course details; introduction to database theory.** What is a query? Why can't we just use first-order logic? Why can't we use a subset of Perl as a query language? What defines a good query language? How can we apply these ideas to the Web?
  - (Review) Chapters 1–3 of [2].
  - Sections 4.0–4.2 of [2].
3. **Introduction to semistructured data.** What is semistructured data? What are the advantages and disadvantages of the relational, object, unstructured, and semistructured models?
  - Chapter 4 of [1]
  - The XPath language specification [15]
4. **Introduction to nontraditional data processing environments.** How do we cope with unbounded or very large streams of data? What query semantics are appropriate? What are some implementation strategies? Which standard methods work and which need major change? How do we support database functionality in a peer-to-peer network? How do we manage data from sensor networks?
  - Streaming XPath processing [29].
  - Peer-to-peer indexes [6].
  - Sensor networks [26, 23].
5. **Standard query processing 1.**
  - Chapters 15 and 16 of [18]
6. **Standard query processing 2.**
  - Graefe's survey: [21]
  - The classic System R paper [31]
7. **Conjunctive Queries.**

#	Date	Planned
1	Jan 27	Unit 1
2	Jan 29	Unit 2
3	Feb 3	Unit 3
4	Feb 5	Unit 4
5	Feb 10	Unit 5
6	Feb 12	Unit 6
7	Feb 17	Unit 7
8	Feb 19	Unit 8
9	Feb 24	Unit 9
10	Feb 26	Unit 10
11	Mar 2	Unit 11
12	Mar 5	catch-up, discussion
–	Mar 6	midterm assigned
13	Mar 9	midterm due 11:00am
14	Mar 11	midterm discussion
13	Mar 16	
14	Mar 18	
15	Mar 23	Unit 9
16	Mar 25	Unit 10
17	Mar 30	–
18	Apr 1	Unit 15
19	Apr 6	Unit 11
20	Apr 8	Unit 12
21	Apr 13	Unit 12
22	Apr 15	Unit 13
23	Apr 20	Unit 16
24	Apr 22	Unit 18
25	Apr 27	catch-up
26	Apr 29	project reports due
27	May 4	demos & discussion
28	May 6	demos & discussion
–	May 8	final assigned
29	May 11	demos & discussion
–	May 13	final due 10:00am

Table 1: Approximate schedule

- Rest of Chapter 4 of [2].
- 8. **First-Order Queries.**
  - Chapters 5 of [2].
- 9. **XML Query Languages 1.**
  - Chapter 5 of [1].
  - The XQuery language specification [8].
- 10. **XML Query Languages 2.**
  - Chapter 6 of [1].
  - Lore and Lorel language [27, 3].
  - UnQL [9].
- 11. **Peer-to-peer networks.**
  - Replication strategies [16].
  - Modeling file-sharing workloads [22].
- 12. **Distributed Query Processing**
  - Survey by Kossmann [25].
- 13. **XML in Relational Databases**
  - A book chapter [14].
- 14. **Historical and Hypothetical Data**
  - Heraclitus [20, 19, 17].
  - A book chapter [13].
- 15. **Theoretical Perspective on Query Optimization**
  - Chapter 6 of [2].
  - Acyclic schemas [5].
- 16. **Beyond First-Order Queries, Part 1**
  - Chapters 16 and 17 of [2].
- 17. **Beyond First-Order Queries, Part 2**
  - Chapter 18 of [2].
- 18. **Debate on Universal Relations.** Half the class will argue in favor of the Universal Relation and the other half against.
  - Chapter 17 of [34].
  - System/U [24].
- 19. **Ontologies for the Web.**
  - OWL [32].
- 20. **Miscellany**
  - Bloom filters
    - I highly recommend that you read the paper that started it all [7].
  - XQuery [11, 12].
  - MGM v. Grokster [30].
  - PPay [35].
  - SplitStream [10].

To be completed...

**Resources** It's probably unnecessary to state that a lot of interesting material can be found by searching the Internet using your favorite tools. However, please do not rely only on the Internet for all your literature searches. In particular, do not use search engines as oracles for proving non-existence of related work. For example, if nothing relevant results from a search on "peer to peer concurrency control", it does not mean there is no relevant work. One needs

to try many different search terms, at the very least. More important, at least some of the other resources should be consulted. Further, do not rely on automated search alone. It is important to browse papers as well. A good start is to flip through the pages of SIGMOD and VLDB proceedings from the last three years. (You can certainly flip pages electronically but make sure you use visual scanning and not keyword search as the tool.)

- The ACM Digital Library: <http://www.acm.org/dl/> Requires a subscription, but UMD has a site-wide subscription that gives access from all local machines.
- The DBLP Bibliography Server: <http://www.purl.org/net/dblp> has good coverage of the Database and Logic Programming fields.
- ACM SIGMOD: <http://www.acm.org/sigmod/>. This site includes pointers to many database-related resources in addition to information about the SIGMOD conferences.
- VLDB Foundation: <http://www.vldb.org/>. The two main items here are the VLDB conferences and the VLDB journal.
- ACM journals: ACM TODS <http://www.acm.org/tods/> is the main database-related journal. Others, such as TOIS <http://www.acm.org/tois/> and TOIT <http://www.acm.org/toit/> are also relevant.
- IEEE TKDE: <http://www.computer.org/tkde/index.htm>.
- SIGMOD Record: <http://www.acm.org/sigmod/record/>. There are often articles describing past and current research trends, interviews, and other less formal articles here.
- IEEE Data Engineering Bulletin: <http://www.research.microsoft.com/research/db/debull>.
- Maryland Database Group: <http://www.cs.umd.edu/areas/db/> has pointers to DBChat other local database-related matters.
- Modern Information Retrieval [4]. Use this book for an overview of Information Retrieval. The huge list of references is a big plus.
- Readings in Database Systems [33]. 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 [28]. Look here for distributed query optimization, distributed transaction processing, etc.
- The databases section of Citeseer: <http://citeseer.nj.nec.com/Databases/>. Citeseer

provides a quick way to browse related papers. It is not a substitute for real browsing, but is helpful in the early stages of literature search.

## References

- [1] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann, first edition, Oct. 1999.
- [2] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
- [3] 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, Nov. 1996.
- [4] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, first edition, May 1999.
- [5] C. Beeri, R. Fagin, D. Maier, and M. Yannakakis. On the desirability of acyclic database schemes. *Journal of the ACM (JACM)*, 30(3):479–513, 1983.
- [6] B. Bhattacharjee, S. Chawathe, V. Gopalkrishnan, P. Keleher, and B. Silaghi. Efficient peer-to-peer searches using result-caching. In *Proceedings of the International Workshop on Peer-to-Peer Systems (IPTPS)*, Berkeley, California, Feb. 2003. To appear.
- [7] B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, July 1970.
- [8] S. Boag, D. Chamberlin, M. F. Fernandez, D. Florescu, J. Robie, and J. Simeon. XQuery 1.0: An XML query language. W3C Working Draft 12, Nov. 2003.
- [9] 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.
- [10] M. Castro, P. Druschel, A.-M. Kermarrec, A. Nandi, A. Rowstron, and A. Singh. SplitStream: high-bandwidth multicast in cooperative environments. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, pages 298–313. ACM Press, 2003.
- [11] D. Chamberlin. XQuery: An XML query language. *IBM Systems Journal*, 41(4):597–615, 2002.
- [12] D. Chamberlin. XQuery: A query language for XML. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, page 682. ACM Press, 2003. Tutorial notes at [http://www.almaden.ibm.com/cs/people/chamberlin/sigmod03\\_xquery.pdf](http://www.almaden.ibm.com/cs/people/chamberlin/sigmod03_xquery.pdf).
- [13] S. S. Chawathe. *Managing Historical XML Data*, volume 57 of *Advances in Computers*, chapter 3, pages 109–169. Elsevier Science, 2003.
- [14] S. S. Chawathe. *Semistructured Data in Relational Databases*, chapter 3. Practical Handbook of Internet Computing. CRC Press, 2004. To appear.
- [15] J. Clark and S. DeRose. XML path language (XPath) version 1.0. W3C Recommendation <http://www.w3.org/>, Nov. 1999.
- [16] E. Cohen and S. Shenker. Replication strategies in unstructured peer-to-peer networks. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM)*, pages 177–190, Pittsburgh, Pennsylvania, Aug. 2002.
- [17] M. Doherty, R. Hull, and M. Rupawalla. Structures for manipulating proposed updates in object-oriented databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Montréal, Québec, 1996.
- [18] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice-Hall, 2002.
- [19] S. Ghandeharizadeh, R. Hull, and D. Jacobs. Implementation of delayed updates in Heraclitus. In *Advances in Database Technology—EDBT '92, Lecture Notes in Computer Science 580*, pages 261–276. Springer-Verlag, Berlin, Mar. 1992.
- [20] S. Ghandeharizadeh, R. Hull, and D. Jacobs. Heraclitus: Elevating deltas to be first-class citizens in a database programming language. *ACM Transactions on Database Systems*, 21(3):370–426, Sept. 1996.
- [21] G. Graefe. Query evaluation techniques for large databases. *ACM Computing Surveys*, 25(2):73–169, 1993.
- [22] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, pages 314–329, Pittsburgh, Pennsylvania, Aug. 2003.
- [23] H. Gupta, S. R. Das, and Q. Gu. Connected sensor cover: Self-organization of sensor networks for efficient query execution. In *Proceedings of the 4th ACM International Symposium on Mobile ad hoc Networking and Computing*, pages 189–200, Annapolis, Maryland, June 2003.
- [24] H. F. Korth, G. M. Kuper, J. Feigenbaum, A. van Gelder, and J. D. Ullman. SYSTEM/U: a database system based on the universal relation assumption. *ACM Transactions on Computer Systems (TOCS)*, 9(3):331–347, 1984.
- [25] D. Kossmann. The state of the art in distributed query processing. *ACM Comput. Surv.*, 32(4):422–469, 2000.

- [26] S. R. Madden, M. J. Franklin, J. M. Hellerstein, , and W. Hong. The design of an acquisitional query processor for sensor networks. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, June 2003.
- [27] 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, Sept. 1997.
- [28] M. T. Ozsu and P. Valduriez. *Principles of Distributed Database Systems*. Prentice-Hall, Upper Saddle River, New Jersey, second edition, 1999.
- [29] F. Peng and S. S. Chawathe. XPath queries on streaming data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, San Diego, California, June 2003. To appear. Available at <http://www.cs.umd.edu/projects/xsq/>.
- [30] P. Samuelson. What’s at stake in MGM v. Grokster? *Communications of the ACM*, 47(2):15–20, 2004.
- [31] P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price. Access path selection in a relational database management system. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 23–34, 1979.
- [32] M. K. Smith, C. Welty, and D. L. McGuinness. OWL Web ontology language guide. W3C Recommendation, Feb. 2004.
- [33] M. Stonebraker and J. Hellerstein, editors. *Readings in Database Systems*. Morgan Kaufmaann, San Francisco, California, third edition, 1998.
- [34] J. Ullman. *Principles of Database and Knowledge-Base Systems*, volume 2. Computer Science Press, 1989.
- [35] B. Yang and H. Garcia-Molina. PPay: micropayments for peer-to-peer systems. In *Proceedings of the 10th ACM conference on Computer and communication security*, pages 300–310. ACM Press, 2003.