Primary Interests

* Information Assurance
* Information Retrieval & Document Clustering

Databases to Search:

ACM Digital Library

* ACM Transactions on Database Systems (TODS) **(Scanned back to Volume 33)**
* ACM Transactions on Information Systems (TOIS) **(Scanned back to Volume 25)**
* ACM Transactions on Knowledge Discovery from Data (TKDD) **(Scanned back to Volume 1)**
* ACM Transactions on Speech and Language Processing (TSLP)
* SIGIR: Annual ACM Conference on Research and Development in Information Retrieval
* SIGMOD: International Conference on Management of Data
* SSS: Workshop On Storage Security And Survivability
* **TDES: Trends and Direction in Expert Systems (for work project?\_**
* **IRP2PN: Information Retrieval In Peer-To-Peer Networks (interesting...?)**
* DMKD: Data Mining And Knowledge Discovery

TOIS

**Learning author-topic models from text corpora - drawing correlation between the authors and their topics, etc (amusingly enough they examine the Enron corpus). (cited 0) (dl/bk for further review)**

**published: January 2010, 2004**

**possible important citations:**

1. Berry, M. W., Dumais, S. T., and O'Brien, G. W. 1994. Using linear algebra for intelligent information retrieval. SIAM Rev. 573—595 **(cited by 1146 – googlescholar dl)**
2. Box, G. E. P. and Tiao, G. C. 1973. Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, MA.
3. Brooks, S. 1998. Markov chain Monte Carlo method and its application. Statistician 47, 69--100.
4. Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. 1990. Indexing by latent semantic analysis. J. Amer. Soc. Inform. Sci. 41, 6, 391--407. **(cited by 4994 – googlescholar dl)**
5. (Diederich, J., Kindermann, J., Leopold, E., and Paass, G. 2003. Authorship attribution with support vector machines. Appl. Intell. 19, 1, 109--123.) Not necessarily awesome, but something Nicholas suggested I do.
6. Mosteller, F. and Wallace, D. 1964. Inference and Disputed Authorship: The Federalist Papers. Addison-Wesley, Reading, MA.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. Inform. Retriev. 1, 1-2, 69--90. **(cited by 1351 – googlescholar dl)**

ACM Transactions on Knowledge Discovery from Data (TKDD)

**Feature-preserved sampling over streaming data (cited 0) (dl/bk for further review)**

**published:** 2009

**possible important citations:**

1. Sudipto Guha , Rajeev Rastogi , Kyuseok Shim, CURE: an efficient clustering algorithm for large databases, Proceedings of the 1998 ACM SIGMOD international conference on Management of data, p.73-84, June 01-04, 1998, Seattle, Washington, United States **(See below!)**

SIGMOD '98 Proceedings of the 1998 ACM SIGMOD international conference on Management of data

**CURE: an efficient clustering algorithm for large databases (cited 234) (dl/bk for further review)**

**published: 1998**

**possible important citations:**

1. Norbert Beckmann , Hans-Peter Kriegel , Ralf Schneider , Bernhard Seeger, The R\*-tree: an efficient and robust access method for points and rectangles, Proceedings of the 1990 ACM SIGMOD international conference on Management of data, p.322-331, May 23-26, 1990, Atlantic City, New Jersey, United States **(cited 676) (dl/bk for further review)**
2. Raymond T. Ng , Jiawei Han, Efficient and Effective Clustering Methods for Spatial Data Mining, Proceedings of the 20th International Conference on Very Large Data Bases, p.144-155, September 12-15, 1994 **(cited 267) (bk for further review) (See below!)**

VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases

Efficient and Effective Clustering Methods for Spatial Data Mining **(cited 267) (bk for further review)**

**published:** 1994

**note:** Full text not immediately available

**possible important citations:**

1. Thomas Brinkhoff , Hans-Peter Kriegel , Bernhard Seeger, Efficient processing of spatial joins using R-trees, Proceedings of the 1993 ACM SIGMOD international conference on Management of data, p.237-246, May 25-28, 1993, Washington, D.C., United States **(cited 119) (dl/bk for further review)**

Ideas:

Unsupervised high dimensional data clustering in parallel with gossiping systems?

Unsupervised data clustering and classification and visualization thereof.

Interacting with VLDB clusters, Interactive responsive queries, like the project Dr Kalpakis made us build, etc.

**Assignment:** Write a paragraph, or two at the most, describing the area in which you want to do your research project. Be as specific as you can. List at least three papers that you'll look at. Wikipedia doesn't count. Of those papers, indicate which may be seminal, and why.  
  
I'm requiring that you build your bibliography using BibTex, and the paragraph in LaTeX. So you'll get a chance to practice that!