# CMSC601 Research Area

Patrick Trinkle
Dept. of Computer Science and Electrical Engineering,
University of Maryland Baltimore County,
Baltimore, MD, 21250
tri1@umbc.edu

February 28th, 2010

## 1 Research Area

I have strong interests in interacting with VLDB that have the data clustered, or clustered indexes to provide interactive querying. The interaction is not strictly through queries which take advantage of clustering, but run-time manipulation of results and restructuring of the temporary data to fit more closely to the intent of the query itself.

## 2 Background Documents

Berry et al's paper in 1994 [1] appears to be the second paper published with the notion of using linear algebra to better enable query processing and document indexing. Previous methods involved databases the documents with keywords and searching against these. Berry et al identify the disparity in language for describing the same document contents. This is a seminal paper in its role redefining information retrieval. Much of document clustering and query processing formed its base on utilizing linear algebra and term vector spaces for comparing documents. This paper was preceeded by Deerwester et al [2] in 1990. I'm under the impression the Berry paper is an extension of the original seminal paper released in 1990.

CURE [3] clusters data managed in a database. The paper is highly cited and also references R*-trees, which Dr Kalpakis taught my databases course were efficient for querying databases under certain circumstances. Because the CURE paper references R*-trees, I will be examaing the paper defining these. To further my understanding of R-trees I'll read through Beckmann et al's "The R*-tree: an efficient and robust access method for points and rectangles" [4].

## References

[1] Dumais S. T. Berry, M. W. and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Rev.*, pages 573–595, 1994. *This paper appears to be the second published work on using linear algebra for query processing. They describe the method as Latent Semantic Analysis. Google Scholar reports it has over 1100 citations.*

[2] Dumais S. T. Landauer T. K. Furnas G. W. Deerwester, S. C. and R. A. Harshman. Indexing by latent semantic analysis. *Journal of American Society for Information Science*, 41:391–407, 1990. *This*

*appears to be the first paper that uses latent semantic analysis for indexing documents to support better querying. Google Scholar lists the citations as nearly 5000.*

[3] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: an efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, 1998. *Interesting implementation of clustering with large databases. A possible good starting point for examining data clustering of managed data.*

[4] Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, and Bernhard Seeger. The r*-tree: an efficient and robust access method for points and rectangles. In *Proceedings of the 1990 ACM SIGMOD international conference on Management of data*, 1990. *Appears to be an important improvement over the state of the art in searching through database data. I remember Dr Kalpakis mentioning R-trees in graduate databases as interesting and it is cited over 650 times.*