REPORT DOCUMENTATION PAGE

FORM APPROVED
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	4/4/97	FINAL TECHNICAL RPT, 01 May 94 to 31 Dec 96

4. TITLE AND SUBTITLE OF REPORT
Design and Analysis of Lossless and Lossy Data Compression Methods and Applications to Communication and Caching-Final Report

5. FUNDING NUMBERS
F49620-94-1-0217

6. AUTHOR(S)
Jeffrey S. Vitter

AFOSR-TR-97-0364

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Computer Science
Duke University
Box 90129, LSRC
Durham, NC 27708

8. PERFORMING ORGANIZATION REPORT NUMBER:

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
110 Duncan Avenue, Suite B115
Bolling AFB
Washington, DC 20332-0001

NM

10. SPONSORING/MONITORING AGENCY REPORT NUMBER:

11. SUPPLEMENTARY NOTES:
The views, opinions and/or finds contained in this report are those of the author(s) and should not be construed as an official Air Force Office of Scientific Research position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The work completed in the project dealt with the following areas of data compression and its applications:

* the design and analysis of sophisticated methods for prediction based on data compression techniques, with applications to prefetching, caching, and locality management.

* fast, practical, and code-efficient implementations of arithmetic coding and other coding methods, for use in text and image compression.

* new methods for choosing motion vectors yielding substantially better rate-distortion tradeoffs for video compression in videoconferencing applications.

Duke University recently filed a patent application for the work on prediction.

19971006 045

14. SUBJECT TERMS	15. NUMBER OF PAGES: 9
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT: UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE: UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT: UNCLASSIFIED	20. LIMITATION OF ABSTRACT: UL

Final Project Report

AFOSR Grant No. F49620–94–1–0217
1994–1996

Jeffrey S. Vitter
Lehrman Professor and Chair
Department of Computer Science
Duke University
Durham, NC 27708–0129

1 Research Overview

The work completed in the project dealt with the following areas of data compression and its applications:

- the design and analysis of sophisticated methods for prediction based on data compression techniques, with applications to prefetching, caching, and locality management.
- fast, practical, and code-efficient implementations of arithmetic coding and other coding methods, for use in text and image compression.
- new methods for choosing motion vectors yielding substantially better rate-distortion tradeoffs for video compression in videoconferencing applications.

Duke University recently filed a patent application [14] for the work on prediction.

2 Research Accomplishments

2.1 Optimal Prediction via Data Compression

Prediction for caching and prefetching is very useful for speeding up access time to data on secondary storage. In [15], Prof. Vitter and then-graduate student P. Krishnan develop an optimal universal prefetcher in terms of fault ratio, with particular applications to large-scale databases and hypertext systems. The algorithms are novel in that they are based on data compression techniques that are both theoretically optimal and good in practice. They show for powerful models such as Markov sources and mth-order Markov sources that the page fault rates are optimal in the limit for almost all sequences of page accesses.

In [12] the much stronger form of worst-case analysis is considered, and randomized algorithm is derived that is proven analytically to *converge almost surely to the optimal fault rate in the worst case for every sequence of page requests* with respect to the important class of finite state prefetchers. In particular, no assumption is made about how the sequence of page requests is generated. This analysis model can be looked upon as a generalization of the competitive framework, in that it compares an online algorithm in a worst-case manner over all sequences against a powerful yet non-clairvoyant opponent. It simultaneously achieves the computational goal of implementing our prefetcher in optimal constant expected time per prefetched page, using the optimal dynamic

discrete random variate generator of Matias, Vitter, and Ni (*Proc. 5th Annual SIAM/ACM Symp. Discrete Algorithms*, January 1994).

The more practical aspects of using data compression techniques for prefetching was recently patented by Prof. Vitter and coauthors [14] and is described in a later section. Adapting three well-known data compressors provides three simple, deterministic, and universal prefetchers. The prefetchers are simulated on sequences of page accesses derived from the OO1 and OO7 benchmarks and from CAD applications, and demonstrate significant reductions in fault-rate. Examined issues include cache replacement, the size of the data structure used by the prefetcher, and problems arising from bursts of "fast" page requests (that leave virtually no time between adjacent requests for prefetching and book keeping). The conclusion is that prediction for prefetching based on data compression techniques holds great promise.

2.2 Arithmetic Coding

Arithmetic coding, in conjunction with a suitable probabilistic model, can provide nearly optimal data compression. In [8] we show how arithmetic coding works and describe an efficient implementation that uses table lookup as a fast alternative to arithmetic operations. The reduced-precision arithmetic has a provably negligible effect on the amount of compression achieved. We can speed up the implementation further by use of parallel processing. We discuss the role of probability models and how they provide probability information to the arithmetic coder. We conclude with perspectives on the comparative advantages and disadvantages of arithmetic coding.

2.3 Lossless Image Compression

In [7] we present a method for progressive lossless compression of still grayscale images that combines the speed of our earlier FELICS method with the progressivity of our earlier MLP method. We use MLP's pyramid-based pixel sequence, and image and error modeling and coding based on that of FELICS. In addition, we introduce a new prefix code with some advantages over the previously used Golomb and Rice codes. Our new progressive method gives compression ratios and speeds similar to those of non-progressive FELICS and those of JPEG lossless mode, also a non-progressive method.

The image model in Progressive FELICS is based on a simple function of four nearby pixels. We select two of the four nearest known pixels, using the two with the middle (non-extreme) values. Then we code the pixel's intensity relative to the selected pixels, using single bits, adjusted binary codes, and simple prefix codes like Golomb codes, Rice codes, or the new family of prefix codes introduced here. We estimate the coding parameter adaptively for each context, the context being the absolute value of the difference of the predicting pixels; we adjust the adaptation statistics at the beginning of each level in the progressive pixel sequence.

2.4 Text Compression

We give a detailed algorithm for fast text compression in [9]. Our algorithm, related to the PPM methods, which are the state-of-the-art methods for maximum text compression, simplifies the modeling phase by eliminating the *escape* mechanism, and speeds up coding by using a combination of *quasi-arithmetic coding* and Rice coding. We provide details of the use of quasi-arithmetic code tables, and analyze their compression performance. Our *Fast PPM* method is shown experimentally to be almost twice as fast as the PPMC method, while giving comparable compression.

The novel idea explored in [13] is re-representing the alphabet so that a representation of a character reflects its properties as a predictor of future text. This enables us to use an estimator from a restricted class to map contexts to predictions of upcoming characters. We describe an algorithm that uses this idea in conjunction with neural networks. The performance of this implementation is

compared to other compression methods, such as UNIX compress, gzip, PPMC, and an alternative neural network approach.

2.5 Video Compression and Rate Control

In [5] we compare methods for choosing motion vectors for motion-compensated video compression. Our primary focus is on videophone and videoconferencing applications, where very low bit rates are necessary, where the motion is usually limited, and where the frames must be coded in the order they are generated. We provide evidence, using established benchmark videos of this type, that choosing motion vectors to minimize codelength subject to (implicit) constraints on quality yields substantially better rate-distortion tradeoffs than minimizing notions of prediction error. We illustrate this point using an algorithm within the $p \times 64$ standard. We show that using quadtrees to code the motion vectors in conjunction with explicit codelength minimization yields further improvement. We describe a dynamic-programming algorithm for choosing a quadtree to minimize the codelength. Current research is aimed at heuristics for speeding up the processing time and use of similar ideas for gaining improvements in static image compression. More recent work [6] gives a fast, practical implementation of the optimizations mentioned earlier, and in some cases it even achieves better rate-distortion performance than the computationally-intensive approach.

2.6 Machine Learning

In [2] we introduce a new technique which enables a learner without access to hidden information to learn nearly as well as a learner with access to hidden information. We apply our technique to solve an open problem of Maass and Turán. We describe analogous results for two generalizations of this model to function learning, and apply those results to bound the difficulty of learning in the harder of these models in terms of the difficulty of learning in the easier model. We bound the difficulty of learning unions of k concepts from a class F in terms of the difficulty of learning F. We bound the difficulty of learning in a noisy environment for deterministic algorithms in terms of the difficulty of learning in a noise-free environment. We apply a variant of our technique to develop an algorithm transformation that allows probabilistic learning algorithms to nearly optimally cope with noise. A second variant enables us to improve a general lower bound of Turán for the PAC-learning model (with queries). Finally, we show that logarithmically many membership queries never help to obtain computationally efficient learning algorithms.

In [4], we consider the problem of learning real-valued functions from random examples when the function values are corrupted with noise. With mild conditions on independent observation noise, we provide characterizations of the learnability of a real-valued function class in terms of a generalization of the Vapnik-Chervonenkis dimension, the fat-shattering function, introduced by Kearns and Schapire. We show that, given some restrictions on the noise, a function class is learnable in our model if and only if its fat-shattering function is finite. With different (also quite mild) restrictions, satisfied for example by gaussian noise, we show that a function class is learnable from polynomially many examples if and only if its fat-shattering function grows polynomially. We prove analogous results in an agnostic setting, where there is no assumption of an underlying function class.

In [3] a new general-purpose algorithm is developed for learning classes of $[0, 1]$-valued functions, and a general upper bound on the expected absolute error of this algorithm in terms of a scale-sensitive Vapnik dimension is derived. We apply this result to obtain new upper bounds on packing numbers in terms of this scale-sensitive notion of dimension. We obtain new bounds on packing numbers in terms of Kearns and Schapire's fat-shattering function. We show how to apply both packing bounds to obtain improved bounds on the sample complexity of agnostic learning.

In [11] we develop a provably optimal and computationally efficient algorithm for the rent-to-

buy problem and evaluate its practical merit for the disk spindown scenario via simulation studies. Our algorithm uses $O(\sqrt{t})$ time and space, and its expected cost for the tth resource use converges to optimal as $O(\sqrt{\log t/t})$, for any bounded probability distribution on the resource use times. Alternatively, using $O(1)$ time and space, the algorithm almost converges to optimal. We describe the results of simulating our algorithm for the disk spindown problem using disk access traces obtained from an HP workstation environment. We introduce the natural notion of effective cost, which merges the effects of energy conservation and response time performance into a single metric. We show that our algorithm is best in terms of effective cost for almost all relevant parameters by 6–25% over the optimal online algorithm in the competitive model. In addition, our algorithm reduces excess energy by 17-60%, and when compared against the 5 second threshold reduces excess energy by 6-42%.

In [10] we present a natural online matching problem motivated by problems in mobile computing. A total of n customers connect and disconnect sequentially, and each customer has an associated set of stations to which it may connect. Each station has a capacity limit. We allow the network to preemptively switch a customer between allowed stations to make room for a new arrival. We wish to minimize the total number of switches required to provide service to every customer. When each customer can be connected to at most two stations, some intuitive algorithms have lower bounds of $\Omega(n)$ and $\Omega(n/\log n)$. When the station capacities are 1, there is an upper bound of $O(\sqrt{n})$. When customers do not disconnect and the station capacity is 1, we achieve a competitive ratio of $O(\log n)$. There is a lower bound of $\Omega(\sqrt{n})$ when the station capacities are 2. We present optimal algorithms when the station capacity is arbitrary in special cases.

In the load balancing problem, which we study in [1], there is a set of servers, and jobs arrive sequentially. Each job can be run on some subset of the servers, and must be assigned to one of them in an online fashion. Traditionally, the assignment of jobs to servers is measured by the norm; in other words, an assignment of jobs to servers is quantified by the maximum load assigned to any server. In this measure the performance of the greedy load balancing algorithm may be a logarithmic factor higher than the offline optimal. In many applications, the norm is not a suitable way to measure how well the jobs are balanced. If each job sees a delay that is proportional to the number of jobs on its server, then the average delay among all jobs is proportional to the sum of the squares of the numbers of jobs assigned to the servers. Minimizing the average delay is equivalent to minimizing the Euclidean (or L_2) norm. For any fixed p, $1 \leq p < \infty$, we show that the greedy algorithm performs within a constant factor of the offline optimal with respect to the L_p norm. The constant grows linearly with p, which is best possible, but does not depend on the number of servers and jobs.

3 Patents

The more practical aspects of using data compression techniques for prefetching was recently patented by Prof. Vitter and coauthors [14]. The authors adapting three well-known data compressors to get three simple, deterministic, and universal prefetchers. The prefetchers are simulated on sequences of page accesses derived from the OO1 and OO7 benchmarks and from CAD applications, and demonstrate significant reductions in fault-rate. Examined issues include cache replacement, the size of the data structure used by the prefetcher, and problems arising from bursts of "fast" page requests (that leave virtually no time between adjacent requests for prefetching and book keeping). The conclusion is that prediction for prefetching based on data compression techniques holds great promise.

4 References

1. A. Awerbuch, A. Azar, E. F. Grove, P. Krishnan, M.-Y. Kao, and J. S. Vitter. "Load Balancing in the L_p Norm," *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science (FOCS '95)*, Milwaukee, WI, October 1995.

2. P. Auer and P. M. Long. "Simulating access to hidden information while learning," *Proceedings of the 26th Annual ACM Symposium on the Theory of Computation*, 1994.

3. P. L. Bartlett and P. M. Long. "More theorems about scale-sensitive dimensions and learning," *Proceedings of the 1995 Conference on Computational Learning Theory*, July 1995.

4. P. L. Bartlett, P. M. Long and R. C. Williamson. "Fat Shattering and the Learnability of Real-valued Functions," *Proceedings of the 1994 Workshop on Computational Learning Theory.*

5. D. T. Hoang, P. M. Long, and J. S. Vitter, "Explicit Bit Minimization for Motion-Compensated Video Coding," *Proceedings of the 1994 IEEE Data Compression Conference (DCC '94)*, Snowbird, UT, March 1994.

6. D. T. Hoang, P. M. Long, and J. S. Vitter. "Efficient Cost Measures for Motion Compensation at Low Bit Rates," submitted. A shorter and earlier version appears in *Proceedings of the 1996 IEEE Data Compression Conference (DCC '96)*, Snowbird, UT, April 1996.

7. P. G. Howard and J. S. Vitter. "Fast Progressive Lossless Image Compression," *Proceedings of the 1994 IST/SPIE International Symposium on Electronic Imaging Science and Technology*, San Jose, CA, February 1994.

8. P. G. Howard and J. S. Vitter. "Arithmetic Coding for Data Compression," *Proceedings of the IEEE*, 82(6), June 1994.

9. P. G. Howard and J. S. Vitter. "Design and Analysis of Fast Text Compression Based on Quasi-Arithmetic Coding," *Journal of Information Processing and Management*, **30** (6), 1994, 777–790.

10. E. F. Grove, M. Kao, P. Krishnan, and J. S. Vitter. "Online Perfect Matching and Mobile Computing", *Proceedings of the Fourth Workshop on Algorithms and Data Structures (WADS '95)*, Kingston, Ontario, August 1995.

11. P. Krishnan, P. M. Long, and J. S. Vitter. "Learning to Make Rent-to-Buy Decisions with Systems Applications," *Machine Learning: Proceedings of the Twelfth International Conference*, Armand Prieditis and Stuart Russell, eds., Morgan Kaufmann Publishers, San Francisco, CA, 1995. 322–330.

12. P. Krishnan and J. S. Vitter. "Optimal Prediction for Prefetching in the Worst Case," *Proceedings of the 5th Annual SIAM/ACM Symposium on Discrete Algorithms (SODA '94)*, Alexandria, VA, January 1994, 392–401.

13. P. M. Long, A. I. Natsev, and J. S. Vitter. "Text Compression Via Alphabet Re-Representation," *Proceedings of the 1997 IEEE Data Compression Conference (DCC '97)*, Snowbird, UT, March 1997.

14. J. S. Vitter, K. M. Curewitz, and P. Krishnan. "Online Background Predictors and Prefetchers," United States Patent No. 5,485,609, Duke University, January 16, 1996.

15. J. S. Vitter and P. Krishnan. "Optimal Prefetching via Data Compression," *Journal of the ACM*, **43**(5) September 1996.

5 Various Statistics

5.1 Number of researchers supported

Faculty (including the P.I.): 1
Postdocs: 2 (Prof. Philip M. Long and Dr. Paul G. Howard)
Graduate Students: 3 (Dr. Dzung T. Hoang, Mr. Paul Natsev, Mr. T. M. Murali)

5.2 Professional honors and participation of the P.I.

- Member of Program Committee, 1993 Workshop on Algorithms and Data Structures (WADS '93), Montreal, Canada, August 1993.
- Member of Program Committee, 1994 IEEE Data Compression Conference (DCC '94), Snowbird, UT, March 1994.
- Member of Program Committee, 7th Annual ACM Conference on Computational Learning Theory (COLT '94), New Brunswick, NJ, July 1994.
- Conference Vice-Chair, Member of Program Committee, and Session Chair, 6th IEEE International Conference on Tools with Artificial Intelligence (TAI '94), New Orleans, LA, November 1994.
- Workshop organizer, Workshop on Modeling and Specification of I/O (MSIO '95), as part of the Seventh IEEE Symposium on Parallel and Distributed Processing (SPDP '95), San Antonio, TX, October 1995.
- Member of Program Committee, First ACM International Conference on Mobile Computing and Networking, November 1995.
- Member of Program Committee, Sixth Annual International Symposium on Algorithms and Computation (ISAAC '95), Cairns, Australia, December 1995.
- Member of Program Committee and Session Chair, 1996 ACM-SIAM Symposium on Discrete Algorithms (SODA '96), Atlanta, GA, January 1996.
- Member of Program Committee, Workshop on Randomized Parallel Computing (WRPC '96), as part of the 1996 International Parallel Processing Symposium (IPPS '96), Honolulu, HI, April 1996.
- Session Chair, 1996 ACM Workshop on Parallel Algorithms (WOPA '96), 1996 ACM Federated Computer Research Conference, Philadelphia, PA, May 1996.
- Member of Program Committee and Session Chair, ACM-IEEE Workshop on I/O in Parallel And Distributed Systems (IOPADS '96), 1996 ACM Federated Computer Research Conference, Philadelphia, PA, May 1996.
- Co-Chair of Working Group on Storage I/O Issues in Large-Scale Computing, ACM Workshop on Strategic Directions in Computing Research, Massachusetts Institute of Technology, Cambridge, MA, June 1996.
- Member of Working Group on Computational Geometry, ACM Workshop on Strategic Directions in Computing Research, Massachusetts Institute of Technology, Cambridge, MA, June 1996.

- Member of Program Committee and Session Chair, First CGC Workshop on Computational Geometry, Baltimore, MD, October 1996.
- Member of Program Committee and Session Chair, 1997 IEEE Data Compression Conference (DCC '97), Snowbird, UT, March 1997.
- Member of Program Committee, Workshop on Randomized Parallel Computing (WRPC '97), as part of the 1997 International Parallel Processing Symposium (IPPS '97), Geneva, Switzerland, April 1997.
- Member of Program Committee, 16th Annual ACM Symposium on Principles of Database Systems (PODS '97), Tucson, AZ, May 1997.
- Member of Program Committee, International Conference on Compression and Complexity of Sequences (SEQUENCES '97), Positano, Italy, June 1997.
- Conference Co-Chair, Member of Program Committee, and Session Chair, Second CGC Workshop on Computational Geometry, Durham, NC, October 1997.
- Member of Program Committee, ACM-IEEE Workshop on I/O in Parallel And Distributed Systems (IOPADS '97), San Jose, CA, November 1997.

5.3 Invited talks

- "Locality, Dynamic, and Prediction Issues in DIS," Panel Member, ARO Workshop on Virtual, Distributed Interactive Simulation, Research Triangle Park, NC.
- "Obstacles in the Implementation of Parallel Algorithms," Panel Member, Workshop on Parallel I/O and Databases, Dartmouth Institute for Advanced Graduate Studies (DAGS '93), Hanover, NH.
- "Load Balancing Paradigms for Optimal Use of Parallel Disks and Parallel Memory Hierarchies," Workshop on Parallel I/O and Databases, Dartmouth Institute for Advanced Graduate Studies (DAGS '93), Hanover, NH.
- "Average-Case Analysis of Prediction," Dagstuhl-Seminar on Average-Case Analysis of Algorithms, Schloß Dagstuhl, Wadern, Germany.
- "Paradigms for Optimal Sorting and Computational Geometry in Large-Scale Parallel Memories," Max Planck Institute, Saarbrücken, Germany.
- "Models for Parallel Secondary and Hierarchical Storage," Workshop on Models, Architectures, and Technologies for Parallel Computation, DIMACS, Rutgers University, New Brunswick, NJ.
- "Load Balancing Paradigms for Optimal Use of Parallel Disks and Parallel Memory Hierarchies," Stanford University, Stanford, CA.
- "Optimal Prediction via Data Compression," University of Texas at Dallas, Dallas, TX.
- "Load Balancing Paradigms for Optimal Use of Parallel Disks and Parallel Memory Hierarchies," Keynote address at Workshop on Algorithmic Research in the Midsouthwest (WARM '93), University of North Texas, Denton, TX.

- "Predictive Techniques for Caching and Locality Management," Microsoft Corporation, Redmond, WA.
- "Efficient Processing of Large-Scale Data," Mathematisches Forschungsinstitut Oberwolfach, Germany.
- "Data Compression and Applications," Air Force Office of Scientific Research, Bolling Air Force Base, D. C.
- "Efficient Processing of Large-Scale Data in External Memory," Distinguished Lecturer Series, Johns Hopkins University, Baltimore, MD.
- "How to Predict Well," Tulane University, New Orleans, LA.
- "How to Predict Well," Supercomputing Research Center, Bowie, MD.
- "Future Communication Issues in Dealing with Large-Scale Data," Army Research Office Math/CS Investment Strategy Meeting, Lake Buena Vista, FL.
- "Data Compression Techniques for Networks," Air Force Office of Scientific Research Initiative Planning Meeting, Raleigh, NC.
- "Future Trends and Issues in I/O," Panel moderator, Workshop on Modeling and Specification of I/O (MSIO '95), as part of the Seventh IEEE Symposium on Parallel and Distributed Processing (SPDP '95), San Antonio, TX.
- "I/O Environments for Geometric Computation," Army Research Office MURI Advisory Board, Johns Hopkins University, Baltimore, MD.
- "Theory and Practice of I/O-Efficient Computation," Distinguished Lecturer Series, Northwestern University, Chicago, IL.
- "Predicting Fast and Reliably," keynote speaker at the Midwest Theory Day, Washington University at St. Louis, St. Louis, IL.
- "Predicting Fast and Reliably," University of Venice, Venice, Italy.
- "I/O-Efficient Computation," University of Rome, La Sapienza, Rome, Italy.
- "Algorithms for Processing Line Segments in External Memory, with Applications to Databases and Geographic Information Systems," Dagstuhl-Seminar on Computational Cartography, Schloß Dagstuhl, Wadern, Germany.
- "I/O-Efficient Geometry with TPIE," Army Research Office MURI Advisory Board, University of Pennsylvania. Philadelphia, PA.
- "Sequence Sorting in Secondary Storage," International Conference on Compression and Complexity of Sequences, Positano, Italy.