Algorithmic Techniques for Massive Data Sets

Moses S. Charikar

ECPI final report: 2002 – 2005
(Award number DE-FG02-02ER25540)

1 Summary of research

There was significant research progress during this ECPI project. The following papers related to the project were written in either preliminary or final form during the three years of the project. I first list the papers and their current status, deferring details to later sections. (Note that this list excludes several papers on approximation algorithms for optimization problems, a significant part of my recent research, since they are not directly relevant to the ECPI project. Most of these papers can be downloaded from my publications page at http://www.cs.princeton.edu/~moses/papers/

1. *On the Impossibility of Dimension Reduction in ℓ_1, with Bo Brinkman.* (Bo Brinkman was one of my graduate students supported by the DOE ECPI award.) This paper solves a well known open problem in the field of dimension reduction and metric embeddings in the ℓ_1 norm. This work was awarded the **best paper award** at the IEEE Conference on Foundations of Computer Science, 2003, one of the two leading conferences in theoretical computer science. The paper was invited to and appeared in the Journal of the ACM.

2. *Dimension Reduction in the ℓ_1 norm, with Amit Sahai.* This paper was presented at the IEEE Conference on Foundations of Computer Science, 2002. This is precursor to the result mentioned above.

3. *Better Streaming Algorithms for Clustering Problems, with Liadan O'Callaghan and Rina Panigrahy.* This paper was accepted to the forthcoming ACM Symposium on Theory of Computing (STOC), 2003.

4. *Finding Frequent items in Data Streams, with Kevin Chen and Martin Farach-Colton.* This paper was presented at the International Colloquium on Automata, Languages and Programming (ICALP), 2002. The paper was invited by the program committee to appear in a special issue of Theoretical Computer Science devoted to the best papers from ICALP 2003. The version of the paper has since been prepared and has been published.

5. *Clustering with qualitative information, with Venkatesan Guruswami and Tony Wirth.* (Tony Wirth was one of my graduate students partially supported by the DOE ECPI award). This paper designed algorithms for a certain clustering problem called *Correlation Clustering.* It appeared in the IEEE Conference on Foundations of Computer Science, 2003 and was invited to a special issue of JCSS on papers in learning theory.

6. *Maximizing Quadratic Programs: Extending Grothendieck's Inequality, with Tony Wirth.* This paper is a followup to the previous paper and gave algorithms for a previously hard to solve variant of *Correlation Clustering.* This paper appeared in the IEEE Conference on Foundations of Computer Science, 2004.

7. *Aggregating Inconsistent Information: Ranking and Clustering, with Nir Ailon and Alantha Newman.* This paper considered problems involving aggregating possibly inconsistent information from several sources, specifically the two problems of combining rankings and clusterings. It appeared in the ACM Symposium on Theory of Computing, 2005.

8. *Fitting Tree Metrics: Hierarchical Clustering and Phylogeny, with Nir Ailon.* This extended some of the techniques in the previous paper to the setting of hierarchical clustering and applied the methods to fitting a tree structure to a given data set so as to minimize fitting error. It appeared in the IEEE Conference on Foundations of Computing, 2005.

9. *Image Similarity Search with Compact Data Structures, with Chrstine Lv and Kai Li.* This paper applied theoretical ideas for designing compact representations to the problem of searching a collection of images for images similar to a given query image. It appeared in the ACM Conference on Information and Knowledge Management, 2004.

In the following section, I describe the results mentioned above in greater detail. In the next section, I describe undergraduate research projects advised by me in the past year and outreach activities related to this ECPI project.

2 Research Highlights

2.1 Dimension reduction

A common problem that is encountered in dealing with high dimensional data is the fact that most algorithms for such data scale very poorly with the dimension, typically exhibiting an exponential dependence. Dimension reduction is often used to combat this "curse of dimensionality." PI Charikar and graduate student Bo Brinkman proved that dimension reduction in ℓ_1 norm is impossible, solving a long standing open problem in the area. In sharp contrast to the dimension reduction in ℓ_2 where n point metrics can be mapped to $O((\log n)/\epsilon^2)$ dimensions with distortion at most $1+\epsilon$, Brinkman and Charikar exhibited an n-point ℓ_1 metric where $n^{\Omega(1/\delta^2)}$ dimensions are needed in order to guarantee distortion at most δ. The proof introduced novel new ideas to this field, using linear programming duality in proving dimension lower bounds. The work was very well received by computer scientists and mathematicians. This paper was awarded the best paper award at IEEE FOCS 2003.

2.2 Streaming Algorithms for Clustering Problems

In recent years, the streaming model has received a lot of attention as a fundamental model for developing algorithms for massive data sets. Here, an algorithm must process its input by making one pass (or a small number of passes) over it, using a limited amount of memory. This is a common model used for settings where the size of the input far exceeds the size of the main memory available and the only feasible access to the data is by making one or more passes over it. In a recent paper with Liadan O'Callaghan and Rina Panigrahy, we examined clustering problems in the streaming model. Clustering problems have been studied extensively across several disciplines. They are typically formulated as optimization problems, where the input is a set of points with a distance function defined on them and the goal is to find a clustering solution (a partion into clusters) that optimizes a certain objective function. A common objective function for clustering is the k-median objective: find k centers (medians) in a set of n points so as to minimize the sum of distances of points to their closest centers. A previous paper of Guha *et al* gave a streaming algorithm for the k-median problem. Their algorithm uses space n^{ϵ} and produces an $O(2^{O(1/\epsilon)})$-approximation. Our research was motivated by the following question, left open by the work of Guha *et al*: *Is is possible to devise a streaming algorithm for the k-median problem that uses only $O(k\, poly \log(n))$ space and produces a constant (or even logarithmic) approximation factor ?* Our main result solves this open problem. We give a streaming algorithm for the k-median problem that uses $O(k\, poly \log n)$ space and produces a constant factor approximation; the algorithm is randomized and works with high probability.

2.3 Correlation clustering

Clustering is a common algorithmic technique to deal with large data sets, serving to organize and interpret unstructured data. PI Charikar and graduate student Tony Wirth collaborated with Venkat Guruswami on correlation clustering - a clustering problem motivated by learning applications. Consider a setting where we want to partition a set of items into disjoint clusters, given access to pairwise similarity/dissimilarity information from a classifier. The problem can be formalizes as follows: Given a graph with edges labeled with $+$ or $-$ (possibly with weights), the goal is to partition the vertices into clusters which is as consistent as possible with the labeling of pairs. Intuitively, we would like $+$ pairs to be inside clusters, and $-$ pairs to be across clusters. We could consider a variety of optimization problems here, such as minimizing disagreements, maximizing agreements, and so on. Charikar, Guruswami and Wirth gave a simple LP based 4-approximation for minimizing disagreements for complete graphs, an $O(\log n)$ approximation for general graphs, an SDP based constant factor approximation for the maximization problem and a variety of hardness of approximation results. This paper appeared in FOCS 2003 and was invited to a JCSS special issue for papers on learning theory.

In recent work, Charikar and graduate student Wirth obtained the first non-trivial results on the problem of maximizing agreements-disagreements for correlation clustering, obtaining an $\Omega(1/\log n)$ approximation. Prior to this, no non-trivial approximation was known for this problem. The new result is based on an SDP based approximation algorithm for a general class of quadratic maximization problems. The analysis can be viewed as an extension of Grothendieck's inequality, a fundamental inequality in functional analysis. This paper was accepted and will appear in the forthcoming IEEE FOCS 2004 conference.

2.4 Aggregating Inconsistent Information

In joint work with graduate student Nir Ailon and visiting postdoctoral student Alantha Newman, we studied the problem of aggregating information from different inconsistent sources, specifically in the setting of ranking and clustering. For example, given rankings of web pages for the same query by different search engines, how should one combine the results into a single ordered list ? Given the output of different clustering methods on the same data set, i.e. different partitions of the same data, how should one combine these to produce a single clustering of the data ? Such problems were proposed and considered by researchers in many different research areas before. Our work gave extremely simple algorithms to solve these problems with very good performance guarantees. For example, the algorithm for aggregating rankings is a variant of the quicksort sorting algorithm. In later work, we applied these techniques to hierarchical clusterings and made substantial progress on a longstanding open problem on fitting a tree metric to a given set of pairwise distances.

2.5 Indexing large collections of high dimensional data using compact signatures

In ongoing collaborative work with faculty member Kai Li and graduate students Christine Lv, William Josephson, and Zhe Wang, we are investigating techniques to efficiently index and search large collections of high dimensional data leveraging recent developed theoretical techniques for finite metric spaces and their applications to compact signatures for complex high-dimensional data sets. Our initial results were for images which we have since extended to a general purpose toolkit for a variety of different data types. The toolkit is still under development and has been applied to images, sounds, 3D shapes and genomic data. The eventual goal is to release this software for use by other researchers.

2.6 Embedding edit distance in ℓ_1

Devising embeddings of special metrics into ℓ_1 is valuable as it facilitates the use of “off the shelf” algorithmic techniques devised for ℓ_1, such as efficient algorithms for nearest neighbor search and so on. Motivated by these connections, a number of special metrics have been investigated to understand the minimimum distortion required to embed them into ℓ_1. One such special metrics which have so far not been understood very well is the string edit distance metric. For a pair of strings, edit distance measures the minimum number of character insert and delete operations required to transform one string into another. This is a natural computationally defined metric for documents and positive algorithmic results for this could lead to advances in indexing large collections of documents. PI Charikar and collaborator Robi Krauthgamer from IBM Almaden made progress on this by devising a simple embedding for collections of non-repetitive strings, i.e. strings where no small substring occurs repeatedly. This embedding has a distortion of $O(\log n)$ where n is the maximum length of a string in the collection. In contrast to other known embeddings for strings, this new embedding is very simple and could eventually be useful for solving the general edit distance problem.

3 Outreach

Charikar organized a workshop on finite metric spaces and their applications at Princeton in August 2003. This brought together leading researchers in theoretical computer science, computer science applications as well as mathematics to discuss current research in this newly emerging and quickly developing research area. In December 2003, Charikar gave an invited talk on ”Streaming Algorithms” at a workshop in Bombay, India as part of FSTTCS (Foundations of Software Technology and Theoretical Computer Science). Streaming algorithms are very efficient algorithms which operate on very large data sets using only a small amount of storage space. In July 2004, he also gave an invited talk on algorithmic aspects of finite metric spaces at COLT 2004 (the annual conference on computational learning theory). Similar invited talks were given at the Annual Fall Workshop in Computational Geometry (Nov 2004), the Yale CS Colloquium (April 2005) and the Statistics and Optimization of Clustering Workshop in Windsor, England (July

2005). PI Charikar also co-taught a PhD summer school on *Finite Metric Spaces and their Algorithmic Applications* at IT University, Copenhagen in the summer of 2004.

4 Graduate and undergraduate training

In the past year, the PI advised six graduate students and worked with six undergraduate students on semester long or year long research projects, many on problems related to this project. One grad and one undergrad were female.

4.1 Undergraduate student projects

In the past year, PI Charikar advised five undergraduate independent work projects and one senior thesis. Most of these projects were closely tied to ongoing research work and were designed to give undergraduates a glimpse of current research. Emily Huang worked on compact representation schemes for images, experimenting with techniques to devise compact image signatures so that approximate image similarity could be determined from them. Lev Reyzin worked on clustering linguistic data in an on-line fashion, specifically focussing on clustering news articles in one pass so as to discover emerging news stories. Bryce Liu worked on improved computer input methods for Chinese through segmentation of Pinyin. The approach used was to train an algorithm to learn the segmentation rules on the basis of a large corpus of Pinyin. Mike Dinitz worked on generating gap examples for a certain semidefinite programming relaxation in connection with Charikar and Wirth's recent work on correlation clustering. Lorenzo Orecchia worked on analyzing semidefinite programming relaxations for maximum acyclic subgraph, attempting to beat the longstanding 2-approximation for the problem. Finally, Rahul Bharagava did a senior thesis on investigating the minimum distortion of embedding planar graphs into ℓ_1.

4.2 Undergraduate and graduate education

PI Charikar has continued to teach an undergraduate course on Discrete Mathematics for undergraduates. The PI is in the process of continually updating and revamping the course to reflect current research trends and modern applications of discrete mathematics. Charikar has also continued development of a graduate course on Advanced Algorithms.