

Massive Data Sets

...Reflections On a Workshop

Jon R. Kettenring
Mathematical Sciences Research Center
Telcordia Technologies
445 South Street
Morristown, NJ 07960

E-mail: jon@research.telcordia.com

In 1995, the National Research Council's Committee on Applied and Theoretical Statistics held an interdisciplinary workshop on massive data sets with the encouragement and support of the National Security Agency. The workshop participants considered a range of practical applications and used them to generate an overall perspective on the statistical and computational challenges as well as a number of specific research opportunities. In this paper, I will review the transactions and comment on what has happened subsequently. Extra attention will be given to the area of information retrieval and the methodology of latent semantic indexing.

Key Words: cluster analysis, data quality, latent semantic indexing, singular value decomposition, visualization of data

1. Introduction

One purpose of this paper is to reflect on the events leading up to a workshop on massive data sets that took place in July of 1995 under the auspices of the Committee on Applied and Theoretical Statistics at the National Research Council. A second purpose is to recall some of the content of the workshop, and its follow up, not only for its historical interest but also because of its continuing relevance. Since much of the discussion at the workshop centered on the method of latent semantic indexing for information retrieval, this topic is covered in somewhat more detail. Finally, a few thoughts are offered on what lies ahead in terms of progress and priorities.

2. The 1994-95 Scene

The Committee on Applied and Theoretical Statistics, affectionately known as CATS, has been operating as part of the Board on Mathematical Sciences at The National Research Council for over 20 years. Its official mission is to “provide a locus of activity and concern for the statistical sciences.” Unofficially, it has served as a watchdog for critical issues involving statistics and spotlighted them as they have arisen. In 1994, through its normal process of priority setting and with the strong urging of the National Security Agency (via Jim Maar, in particular, who was then Chief of Statistical Research Techniques), CATS decided that the time was right to create a special focus on the technical challenges presented by increasingly large data sets.

This focus took the form of a workshop that was carefully designed to bring together practitioners from a wide range of fields as well as representatives from the statistics community. Potential participants were invited to represent their views and experiences via electronic position statements. These were shared and commented on in the spirit of electronic brainstorming. They proved useful for structuring the program and provided considerable momentum for the workshop itself.

Ultimately, 10 formal presentations—all dealing with real-world applications and associated massive data challenges—were selected as the anchor points for the workshop. They served as the stimuli for all discussion about underlying research issues and needs.

3. Follow Up and Reflections

The proceedings of the workshop were published (National Research Council 1996) and in 1999 five of the applications papers (plus other related ones) appeared in the *Journal of Computational and Graphical Statistics* in a special section devoted to the workshop. The featured applications were about earth observation satellite data (Kahn and Braverman 1999), magnetic resonance imaging data (Eddy, Fitzgerald, Genovese, Lazar, Mockus, and Welling 1999), healthcare data (Goodall 1999), telephone network data (McIntosh 1999), and atmospheric measurements data (Levy, Pu, and Sampson 1999).

Buja and Keller-McNulty (1999) introduced the special section by noting:

In retrospect, it appears that this workshop took place ahead of its time: The term “massive datasets” was not yet common currency, and the number of published articles on the topic in the statistics literature was probably less than a handful. It is therefore not surprising how fresh many of the contributions to the workshop appear four years later.

Indeed, they seem just as relevant six years later! We are still facing many of the same challenges even though one can point to progress in several directions.

Much of the discussion at the workshop dealt with the definition of a massive data set. At that time it seemed appropriate to stick with a murky definition even though it caused frustration for some. Here is one version:

A massive data set is one for which the size, heterogeneity, and general complexity cause serious pain for the analyst(s).

Several comments are in order. The “pain” cannot be relieved by simple random sampling. This is usually due to the extensive heterogeneity in the data and the risk that important characteristics will be missed or sorely under represented. One aspect of the pain is that the time involved to do what needs to be done—modeling, data analysis, etc.—is potentially enormous. In many instances the number of dimensions to the data may itself be huge. There have been several attempts to make the definition more precise. Huber (1999) commented that “the aggravation starts around 10^8 bytes.” See also Wegman (1995). Nevertheless, there is much to be said for a fuzzier definition. In this vein, Sacks (National Research Council 1996) commented: “I don’t know what it is, but I know it when I see it.”

In Kettenring (1997), reflecting my own thinking about the underlying themes of the workshop, I laid down six approaches that seemed to be valuable to pursue. The first is adaptive sampling by which I meant the use of “intelligent” sampling to detect and preserve important structure. This is a vague concept but the spirit is to learn as you go or sample based on what has already been observed. Such an approach will likely lack the mathematical rigor of traditional sampling, but the gain may still be worth the price. See Wegman (1995) for a similar suggestion.

The second is reliance on approximations. As in the first approach the intent is a pragmatic one: forget the comfort of optimal methods in order to achieve progress. One way this may play out is through asymptotic arguments that give rise to approximate but quick and “good enough” answers.

The third approach is to develop guided visualization methods that would be effective at “pan and zoom.” Because many massive data problems have a natural hierarchical structure to them, visualization systems that are tuned to handle such data should prove very useful. Another idea, suggested by Tierney (National Research Council 1996) and others, is to find “some way of using design ideas to help us...decide what parts are worth looking at.” While it is encouraging that much progress has been made in developing various imaging tools (Lawler 2001), general-purpose data visualization systems for dealing with massive data remain a significant challenge for the statistics research community.

Number four is distributed work. This has both a technological and sociological aspect to it. On the latter, Carr (National Research Council 1996) commented: “if [statisticians] want to be part of big science...we

need to learn to work on teams...and talk with people from other disciplines.” This message has now been repeated often enough that it is probably widely accepted, and an increasing number of statistics departments appear to be preparing students for cross-disciplinary work.

Fifth: divide and conquer. At the workshop there was general understanding and agreement that direct global modeling does not work. Instead, modeling needs to proceed in stages. First, one works to find homogeneous regions in the data. Global considerations enter later in what might be labeled a consolidation phase. St. Amant and Cohen (National Research Council 1996) talked about the regions in terms of data geography: “we need to know either where to look for patterns or what kind of patterns to look for.” Fayyad’s comment (National Research Council 1996) that “partitions of the data, clusters, are the key to dealing with large data sets” captures much of the dialogue about how one needs to proceed in the absence of other information that can help in the partitioning or regionalization.

The last approach is to exploit the context. Indeed, this one is so important that it should probably have been first on the list. The nature of the problem will strongly influence, if not dictate, how one should proceed. This point is closely related to Mallows’ 1997 R. A. Fisher Lecture on the “zeroth problem” (Mallows 1998).

Immediately following the workshop, Huber wrote up his reflections on the event, which he entitled “The Morning After.” Subsequently, he edited his thoughts and added additional perspective in Huber (1999). The paper is filled with important and provocative comments. A few examples are included here.

He describes visualization as an “oxymoron—the art is to reduce size before one visualizes. The contradiction (and challenge) is that we may need to visualize first in order to find out how to reduce size.”

His advice on complexity of algorithms is to keep them below $O(n^{3/2})$. [Wegman (1995) provides extensive analysis of the limits of computational and visualization feasibility.]

On strategy and analysis, Huber’s view is that work typically begins with “task and subject-matter specific, complex preprocessing, or by extracting systematic subsets on the basis of a priori considerations, or a combination of the two.” Subsets can often be defined “by windows in space and time” or by finding “remarkable features” and then “extracting all data in the immediate neighborhood of such features.” As for summaries, they “enter only later.” And then with a dig at data mining, he adds that identifying “noteworthy but otherwise unspecified features by machine...is a hopeless search for the Holy Grail.” Huber has much to say also about the infrastructure one needs for massive data analysis including programming environments, database management systems, and massively parallel machines.

4. Latent Semantic Indexing

Dumais’ workshop presentation (National Research Council 1996) was based on her pioneering paper, Deerwester, Dumais, Landauer, Furnas, and Harshman (1990). The context is information retrieval from online textual databases. Their approach falls into the category of vector space modeling and is commonly referred to as latent semantic indexing (LSI). The idea is to form a term-by-document data matrix for a collection of documents of interest. Each cell of the matrix is a count—possibly weighted to reflect the relative importance of the words and documents—of the number of times a particular word appears in a particular document. The resulting matrix tends to be huge even though almost all the cells may be empty. To reduce dimensionality, and—roughly speaking—to smooth out the data and drop unneeded portions, the singular value decomposition (SVD) is used in the spirit of principal components analysis (PCA) to obtain a lower rank approximation to the data matrix. (Even with the reduction, the number of dimensions or singular values retained is typically still very large.) Documents are represented in the implied reduced space and retrieval is done by projecting a query into the same space and identifying documents nearest (in some sense) to it. The tendency is for documents that are similar in content to be located near enough to each other so that if one is retrieved they all will be.

Given the complexity of the information retrieval challenge, and the relative crudeness of the calculations, it has always impressed me how well the LSI methodology performs. Moreover, the variety of applications

has also been impressive. Many requests have come to Telcordia for copies of LSI software. Intended uses include cross-language document retrieval, essay scoring, electronic newspaper personalization, linguistics education, and analysis of language patterns of patients with neurological problems.

When I first learned about LSI in the late 1980's, I had the reaction that information retrieval problem cried out for a cluster analysis approach because clustering is in essence a much more localized calculation than SVD or LSI. Of course, a moment's reflection on the magnitude of the clustering task, with so many words and documents to consider, was enough to scare anyone off at that time. Nevertheless, there has been a series of attempts to tackle information retrieval via various clustering methods with several of them occurring in the last five years.

To limit the discussion, I will only mention the recent work of Dhillon and Modha (2001) and Dhillon, Fan, and Guan (2001) as one example of this line of research. They develop a version of k-means clustering that is tailored to the data matrix at hand and efficient in memory usage. The overall complexity of their algorithm depends on the product of the number of non-zero cells in the data matrix, the number of clusters assumed, and the number of iterations.

Before clustering, the document vectors of the data matrix are normalized to (Euclidean) length one. After clustering, the centroid of each cluster is computed and also normalized to length one to yield what are called concept vectors. These vectors are the closest to their respective clusters according to cosine similarity, a measure which is emphasized in LSI. Furthermore, Dhillon and Modha note an empirical tendency for the concept vectors to be less dense than the (left) singular vectors from the corresponding SVD. They also tend to be orthogonal and to span the same space as the leading k left singular vectors.

Dhillon and Modha use the concept vectors to do a regression-like decomposition of the data matrix as an alternative to the SVD. With this in hand one can proceed with document retrieval in the same manner as with LSI.

It will be worth following how successful this new approach turns out to be. One still has the challenge of choosing the "right" value of k just as one must determine how many terms to keep in the SVD. There is the additional annoyance of worrying about the starting points for the k-means algorithm and the stability and optimality of the solution. Will it be necessary generally to have as many clusters as retained terms in the SVD? And what will happen when the number of clusters in the data is huge? Extensive experimentation will be needed to gauge the relative performance and practicality of these two methods.

5. Looking Ahead

Since the workshop in 1995, there has been steady progress on and heightened awareness of the challenges presented by massive data sets. The topic has clearly become a popular one in the research community. As one example, take a look at the table of contents of this conference, Interface '01!

While progress is being made on many fronts, breakthroughs in data visualization systems and clustering methodologies for massive data problems would seem to be among the highest priorities. Since many of the large problems have a natural hierarchical structure to them, visualization systems that are tuned to that type of data would be especially welcome. Clustering methods not only need to scale well to deal with size but also to reflect the statistical nature of the data. Algorithms that are clever computationally but fail to deal with issues such as statistical variability within clusters will fall short of what is really needed.

At the same time that we are struggling to increase our capabilities for dealing with data quantity, the companion problems of data quality need to be tackled as well. In particular, while industry is awash with massive data, it is weighed down by the enormous costs associated with various shortcomings and errors in these collections. The need for increased attention to data quality is also reflected in experiences with data on the Web.

Still, simple extrapolation from 1995 on the rate of progress suggests that practitioners will continue to feel the pain of massive data challenges in the years ahead. As the Director of the National Security Agency

put it in a recent discussion of communications data (Wall Street Journal, 5/23/01, p. 1), “there’s simply too much out there, and it’s too hard to understand.”

Acknowledgements

Thanks to Daryl Pregibon of AT&T Labs for his help in designing the Massive Data Sets Workshop, William Szewczyk of the National Security Agency for organizing the NSA Overview session in which this paper was presented at Interface '01, and Cliff Behrens of Telcordia Technologies for his insights into LSI.

REFERENCES

Buja, A. and Keller-McNulty, S. (1999), “Introduction to the Special Section on Massive Datasets,” *Journal of Computational and Graphical Statistics*, 8, 544.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G.W., and Harshman, R. A. (1990), “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science*, 41, 391-407.

Dhillon, I. S., Fan, J., and Guan, Y. (2001), “Efficient Clustering of Very large Document Collections,” in *Data Mining for Scientific and Engineering Applications*, to appear.

Dhillon, I. S. and Modha, D. S. (2001), “Concept Decompositions for Large Sparse Text Data Using Clustering,” *Machine Learning*, 42, 143-175.

Eddy, W. F., Fitzgerald, M., Genovese, C., Lazar, N., Mockus, A., and Welling, J. (1999), “The Challenge of Functional Magnetic Resonance Imaging,” *Journal of Computational and Graphical Statistics*, 8, 545-558.

Goodal, C. (1999), “Data Mining of Massive Data Sets in Healthcare,” *Journal of Computational and Graphical Statistics*, 8, 620-634.

Huber, P. (1999), “Massive Datasets Workshop: Four Years Later,” *Journal of Computational and Graphical Statistics*, 8, 635-652.

Kahn, R. and Braverman, A. (1999), “What Shall We Do With the Data We Are Expecting From Upcoming Earth Observation Satellites?,” *Journal of Computational and Graphical Statistics*, 8, 575-588.

Lawler, A. (2001), “New Imaging Tools Put the Art Back Into Science,” *Science*, 292, 1044-1047.

Levy, G., Pu, C., and Sampson, P. D. (1999), “Statistical, Physical, and Computational Aspects of Massive Data Analysis and Assimilation in Atmospheric Applications,” *Journal of Computational and Graphical Statistics*, 8, 559-574.

Mallows, C. L. (1998), “The Zeroth Problem,” *The American Statistician*, 52, 1-9.

McInstosh, A. (1999), “Analyzing Telephone Network Data,” *Journal of Computational and Graphical Statistics*, 8, 611-619.

National Research Council (1996), *Massive Data Sets*, Washington, DC: National Academy Press.

Wall Street Journal, May 23, 2001.

Wegman, E. (1995), “Huge Datasets and the Frontiers of Computational Feasibility,” *Journal of Computational and Graphical Statistics*, 4, 281-295.