Can Econometrics Answer When High Dimensional Text Clusters have Short Signatures?

Vinay Deolalikar, Deependra Singh
Groupon Data Mining
Palo Alto, CA 94040

***Abstract*—The past decade has witnessed an unprecedented generation of unstructured information. Given this mountain of unstructured information, a paradigm that is of increasing importance is to extract a small amount of this information for use by applications. As a separate, recent development, quantities that were earlier modeled using the classical Pareto 80-20 rule were found to display a distinct phenomenon the so-called long-tail. This phenomenon, seen, for example, in the effect of the internet on sales curves, has influenced recent thinking in econometrics.**

In this interdisciplinary paper, we demonstrate links between the above two seemingly disparate issues. We show that the problem of extracting a small amount of information in data mining contexts is amenable to analysis using recent ideas from econometrics. In particular, we view the centroids produced during document clustering, which capture important statistical information about the corpus, through the econometric lens of Pareto vs. long-tail. We ask when such clusters have short signatures, and construct a framework using ideas from econometrics that gives us intuitively satisfying answers to this question. We also unearth a new structural phenomenon that we call "saturation" that should give impetus to further study.