Quantifying Chaos in Large Heterogeneous Data Repositories.

Luann Jung[1], Brendan Whitaker[2]
Kyle Chard[3], Aaron Elmore[3] (advisors)

[1]MIT, [2]Ohio State University, [3]University of Chicago

Overview of this talk:

- Theory and background: what is the problem?
- Implementation: what is our proposed solution?
- Relevance: How does  fit into the picture?

How do we measure “structure” in the filesystem?

Clustering

Inputs

- Dataset
- Range of k

Assumptions

- Define density as number of files/number of directories.
- Organized dataset ⇒ files in a directory tend to be similar, i.e. similar things are close together
- Clustering groups similar things together
- Disorganized dataset ⇒ clusters may reach a higher max density
- We say k is optimal where the clusters reach max density
- The cleanliness function measures density of the clusters

Conclusion

- The max density of the clusters is an approximate measure of dataset organization
- Since our cleanliness function measures cluster density, we maximize this function over our range of k

- Hypothesis: our proposed method accurately measures organizational structure of a dataset in-filesystem.
- Challenge: there is no established way to evaluate a measure/score.
- Significance: understanding when our data becomes intractable or unusable due to poor management can save time and money in research and industry.

Implementation

Clustering Pipeline

DATA REPOSITORY

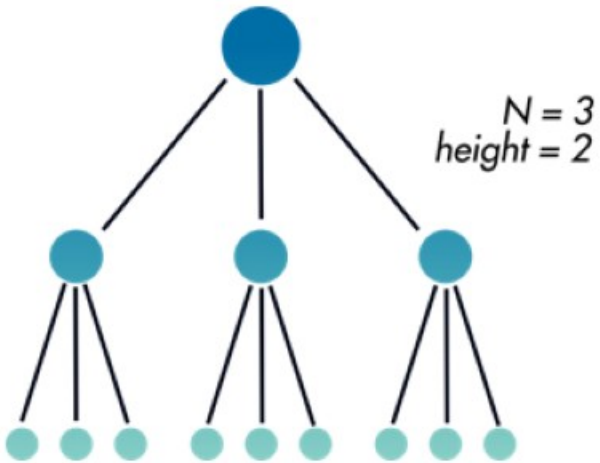
Metadata extraction → *File preprocessing* → *Text clustering* / *Tabular clustering* → *Optimize k* → **CLEANLINESS MEASURE**

Evaluation on Real Data

CDIAC

- Department of Energy Carbon Dioxide Information and Analysis Center's climate data repository (CDIAC)
- ~400 GB
- We did not have a domain expert to verify our clusters
- However, in the future we can work with domain experts to verify and refine our clustering pipeline.

Evaluation

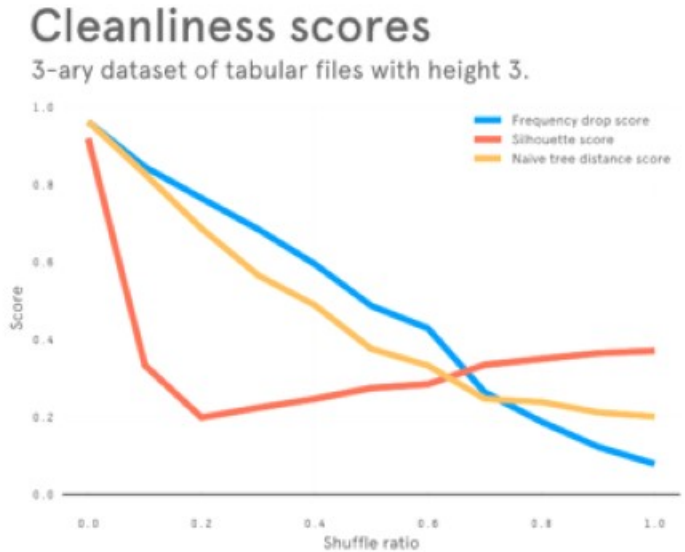


Synthetic Datasets

- As a baseline comparison, we generated 4 "perfect" datasets based on N-ary trees
- We shuffled these datasets by increasing amounts to observe how our cleanliness score performs

Comparison of Scores

- We compared our score with two others:
  - Naïve tree distance (cohesion)
  - Silhouette score (cohesion and separation)
- Frequency drop score more accurately portrays the level of organization present



Resource requirements

- K-means requires vectors in R^n.
- Vectorization of many documents is very memory-hungry.
- Skylake nodes (192GB memory).
- Had serious trouble with network connectivity, and upload speeds.
- Working with filesystem dumps, prohibitively slow.
- Moved to AWS for the second half of our experiments.

Special Thanks to...