Measuring Swampiness: Quantifying Chaos in Large Heterogeneous Data Repositories

Extended Abstract

Luann Jung
Massachusetts Institute of Technology
luju@mit.edu

Brendan Whitaker
Ohio State University
whitaker.213@osu.edu

Kyle Chard (advisor)
University of Chicago
chard@uchicago.edu

Aaron J. Elmore (advisor)
University of Chicago
aelmore@cs.uchicago.edu

ABSTRACT

As scientific data repositories and filesystems grow in size and complexity, they become increasingly disorganized. The coupling of massive quantities of data with poor organization makes it challenging for scientists to locate and utilize relevant data, thus slowing the process of analyzing data of interest. To address these issues, we explore an automated clustering approach for quantifying the organization of data repositories. Our parallel pipeline processes heterogeneous filetypes (e.g., text and tabular data), automatically clusters files based on content and metadata similarities, and computes a novel "cleanliness" score from the resulting clustering. We demonstrate the generation and accuracy of our cleanliness measure using both synthetic and real datasets, and conclude that it is more consistent than other potential cleanliness measures.

ACM Reference Format:
Luann Jung, Brendan Whitaker, Kyle Chard (advisor), and Aaron J. Elmore (advisor). 2018. Measuring Swampiness: Quantifying Chaos in Large Heterogeneous Data Repositories: Extended Abstract. In *Proceedings of ACM Student Research Competition (SC'18).* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

Traditional modes of organizing data repositories and filesystems are increasingly ineffective due to the size, heterogeneity, and complexity of data. Researchers are now turning to alternative organizational models such as data lakes—repositories for large quantities of raw data that are integrated in a pay-as-you-go fashion [5, 6]. However, users are often unwilling to spend time describing and organizing data, causing repositories to become opaque "data swamps"[4] with poor metadata and confusing directory structures.

To combat this problem, we propose a set of tools that automate the process of identifying content-based relationships between files. We present a parallel pipeline that crawls repositories, collects key information regarding data composition and distribution, and automatically clusters files based on extracted content and metadata. Our unsupervised clustering models aim to detect latent similarities in file subject, provenance, or purpose [2] and then clusters accordingly. We use these clusters to define a novel "cleanliness" measure to quantify the organization of the data repository. This measure consists of a newly proposed frequency drop score which takes into account the directory composition and density of clusters generated by the pipeline. We explore the efficacy of our approach using synthetic data as well as a real-world climate science dataset [8].

SC'18, November 2018, Dallas, Texas USA

ACM ISBN N/A.

2 METHODOLOGY

We implement a clustering-based pipeline to identify similar data irrespective of how it is organized. The pipeline is composed of four major steps: crawling, preprocessing, clustering, and calculating cleanliness.

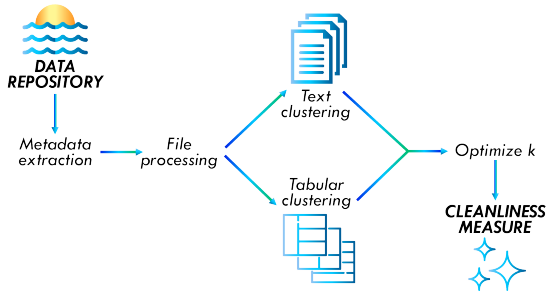

Figure 1: Clustering pipeline.

We focus on two data types: unstructured text and structured tabular data. First, we convert files into common formats (`.txt`/`.csv`). Then, we preprocess file contents according to their data type. Text data are tokenized, stemmed, and vectorized into a TF-IDF matrix, while schemas are extracted from tabular data and used to compute a pairwise Jaccard distance matrix.

For text files, we implement classic k-means clustering and the faster MiniBatch k-means clustering. For tabular files, we use agglomerative hierarchical clustering since it does not rely on centroids or other features of Euclidean space. After clustering both filetypes, we generate output clusters, composition statistics, and a dataset cleanliness score. The pipeline is then repeated over a user-specified range of k values to optimize the k which best represents the data.

To measure cleanliness, we first define the frequency drop score for a clustering of some dataset A by examining the distribution

of directories constituting each cluster C_i. Given the number of files from each directory in a cluster, we identify the location of the largest "frequency drop"—representing the point where the tail of the distribution begins. Let $\{ D_1, ..., D_m \}$ be the set of all directories containing files from cluster $C_i \subseteq A$. We define the head H_i as the set of all directories before the drop, and the tail T_i as the set of all remaining directories of C_i. Under the assumptions that similar data are physically close in well-organized datasets and that the clustering $C = \{ C_1, ..., C_k \}$ is sufficiently cohesive, the function $\mathcal{S}(C)$ yields a value in $[0, 1]$ representing the cleanliness of the dataset. We define a logarithm-like function which is well-defined for a base of 1:

$$\sigma(a, b) = \begin{cases} \log_a b & \text{if } a > 1 \\ 0 & \text{if } a = 1 \end{cases}. \quad (1)$$

The frequency drop score for each cluster is given by

$$\text{drop}(C_i) = \begin{cases} \frac{1-\sigma(m-1,|H_i|)}{|C_i|} \sum_{D_j \in H_i} |D_j| & \text{if } m > 1 \\ 1 & \text{otherwise} \end{cases}, \quad (2)$$

and the score for the entire clustering is given by

$$\mathcal{S}(C) = \frac{|C_i|}{|A|} \sum_{i=1}^{k} \text{drop}(C_i). \quad (3)$$

3 EVALUATION

We evaluate our approach using synthetic data as well as the Carbon Dioxide Information and Analysis Center's (CDIAC) data repository.

As a baseline, we generated three synthetic datasets based on N-ary trees. Each synthetic dataset includes one parent directory (root node) with N children, each of which has N children, extended to any chosen height h. Each leaf node contains twenty `.txt` files and twenty `.csv` files, with each file containing the same word repeated 100 times. Each word is unique to its leaf node, such that the number of expected clusters is equal to the number of leaf nodes. These datasets, when run through our pipeline, yield:

- perfect clusters where each cluster contains only and all of the files with the same word.
- a cleanliness score of 1.0.

With this as a baseline, we then shuffled the datasets such that files were randomly assigned to leaf directories. Table 1 shows that the cleanliness scores decrease as the dataset is shuffled.

	% Scrambled					
Dataset	0%	20%	40%	60%	80%	100%
2-ary, 5-height	1.000	0.806	0.619	0.420	0.227	0.093
3-ary, 3-height	0.963	0.765	0.595	0.429	0.188	0.079
6-ary, 2-height	1.000	0.792	0.593	0.451	0.225	0.106
40-ary, 1-height	0.950	0.780	0.579	0.341	0.217	0.109

Table 1: Cleanliness scores for shuffled synthetic datasets.

We compared our cleanliness score with two other measures: cluster cohesion and a modified Silhouette score [7], both computed with naïve filesystem tree distance. Figure 2 shows these measures calculated on progressively more shuffled synthetic datasets and real scientific data (from the pub8 subset of CDIAC). We conclude that the silhouette scores are inconsistent and noisy when compared to our cleanliness measure. The naïve tree distance score is comparable, but still fails to discriminate between repositories with vastly different organizational structures in some adversarial examples.

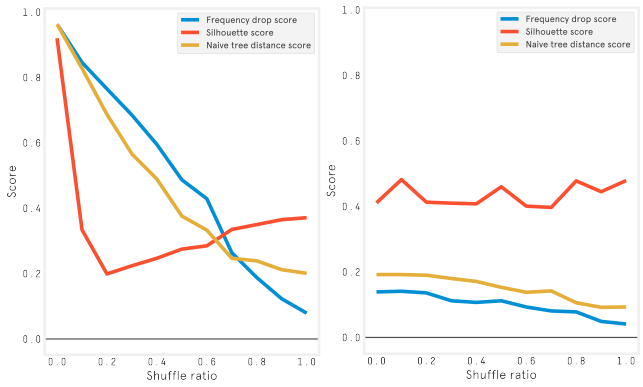

Figure 2: Comparison of cleanliness measures - 3-ary tree synthetic dataset of tabular files with height 2 (left), and tabular files from pub8 (right).

4 SUMMARY

We introduce a parallel pipeline for automated content-based clustering of files from large heterogeneous data repositories. These clusters are then used to derive a novel measure of the organizational cleanliness of a repository. The measure we developed exhibits better consistency than existing measures when tested on a variety of datasets. The code for our pipeline is available here: https://github.com/lollyluann/cluster-datalake

REFERENCES

[1] Paul Beckman, Tyler J Skluzacek, Kyle Chard, and Ian Foster. 2017. Skluma: A statistical learning pipeline for taming unkempt data repositories. In *29th International Conference on Scientific and Statistical Database Management*. 41.

[2] Will Brackenbury, Rui Liu, Mainack Mondal, Aaron J. Elmore, Blase Ur, Kyle Chard, and Michael J. Franklin. 2018. Draining the Data Swamp: A Similarity-based Approach. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA'18)*. ACM, New York, NY, USA, Article 13, 7 pages. https://doi.org/10.1145/3209900.3209911

[3] M. Chessell, F. Scheepers, N. Nguyen, R. van Kessel, and R. van der Starre. 2014. *Governing and Managing Big Data for Analytics and Decision Makers*. http://www.redbooks.ibm.com/redpapers/pdfs/redp5120.pdf

[4] Rihan Hai, Sandra Geisler, and Christoph Quix. 2016. Constance: An intelligent data lake system. In *Proceedings of the 2016 International Conference on Management of Data*. ACM, 2097–2100.

[5] Shawn R Jeffery, Michael J Franklin, and Alon Y Halevy. 2008. Pay-as-you-go user feedback for dataspace systems. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 847–860.

[6] Jayant Madhavan, Shawn R Jeffery, Shirley Cohen, Xin Dong, David Ko, Cong Yu, and Alon Halevy. 2007. Web-scale data integration: You can only afford to pay as you go. CIDR.

[7] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53 – 65. https://doi.org/10.1016/0377-0427(87)90125-7

[8] U.S. Dept. of Energy. 2017. Carbon Dioxide Information Analysis Center. ftp://cdiac.ornl.gov. Visited Feb. 28, 2017.