Deciphering Vague Information Searches via Search Result Clustering

Katrina Truebenbach: truebenbach.k@husky.neu.edu

Problem Definition:

Cluster search results from broad queries into topics to enable faster information discovery

- Want clusters with the following properties:
  1. **Internally Coherent**: documents within the cluster represent the same topic
  2. **Externally Distinct**: clusters represent different topics
  3. **Labeled**: clusters have human-readable labels that define their topic
- Usability properties:
  - Relatively *small number* of clusters per search: don't overwhelm users
  - Relatively *evenly sized* clusters: don't want one cluster with all but three results
- Large-scale problem: Need a heuristic for determining the number of clusters across search terms

Existing Methods

Three groups of methods: Trade-off between well-formed clusters and human-readable labels[1]

1. Data-Centric: focus on finding good clusters. Labeling as an after-thought.
2. Description-Aware: cluster and label based on one feature
3. Description-Centric: allow labels to dictate clusters

Proposed Method: Comparison between Data-Centric and Description-Centric

- Form TF-IDF Vector Space Model for documents returned by a search; Euclidean Distance Matrix

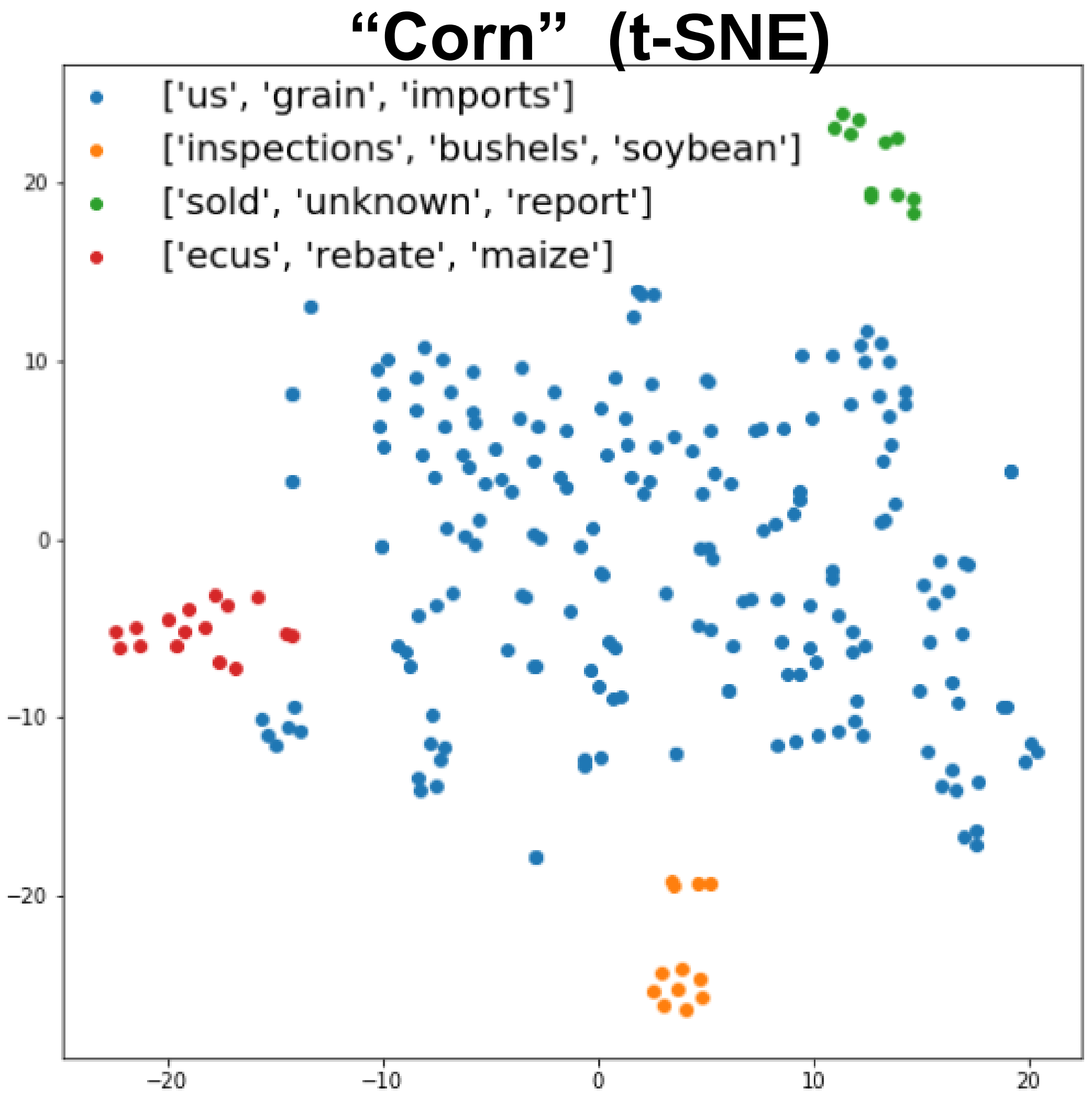
Agglomerative Hierarchical Clustering: *Exactly one cluster assignment per document*

- PCA dimensionality reduction: *Retain 80% of variance*
- Linkage Matrix: *Ward metric*
- Cut dendrogram at k clusters: for k in a restricted range, find an elbow in the rate of change of distortion
- Label: 3 highest scored words in top-level cluster centroid (mean vector of cluster)
- Sort: silhouette score * size of cluster

Modified Lingo Algorithm:[2] *Some documents in multiple clusters and some documents in no clusters*

- SVD with k dimensions: for each possible k, find elbow in the rate of change of retained variance
- Label: 3 highest valued words in each column (concept/cluster) of reduced V.
- Assign documents to clusters: *Document-label strength > 0.1*
- Combine clusters with overlapping labels
- Sort: highest label score * size of cluster

Hierarchical Clustering: "Corn" (t-SNE)



Data Description & Experimental Setup

Data: **Reuters financial newswire articles**

10,788 articles with 90 topic labels.

Topic labels used as proxy search terms; clustering within.

- Remove symbols, punctuation, numbers, stop words. Stem words.

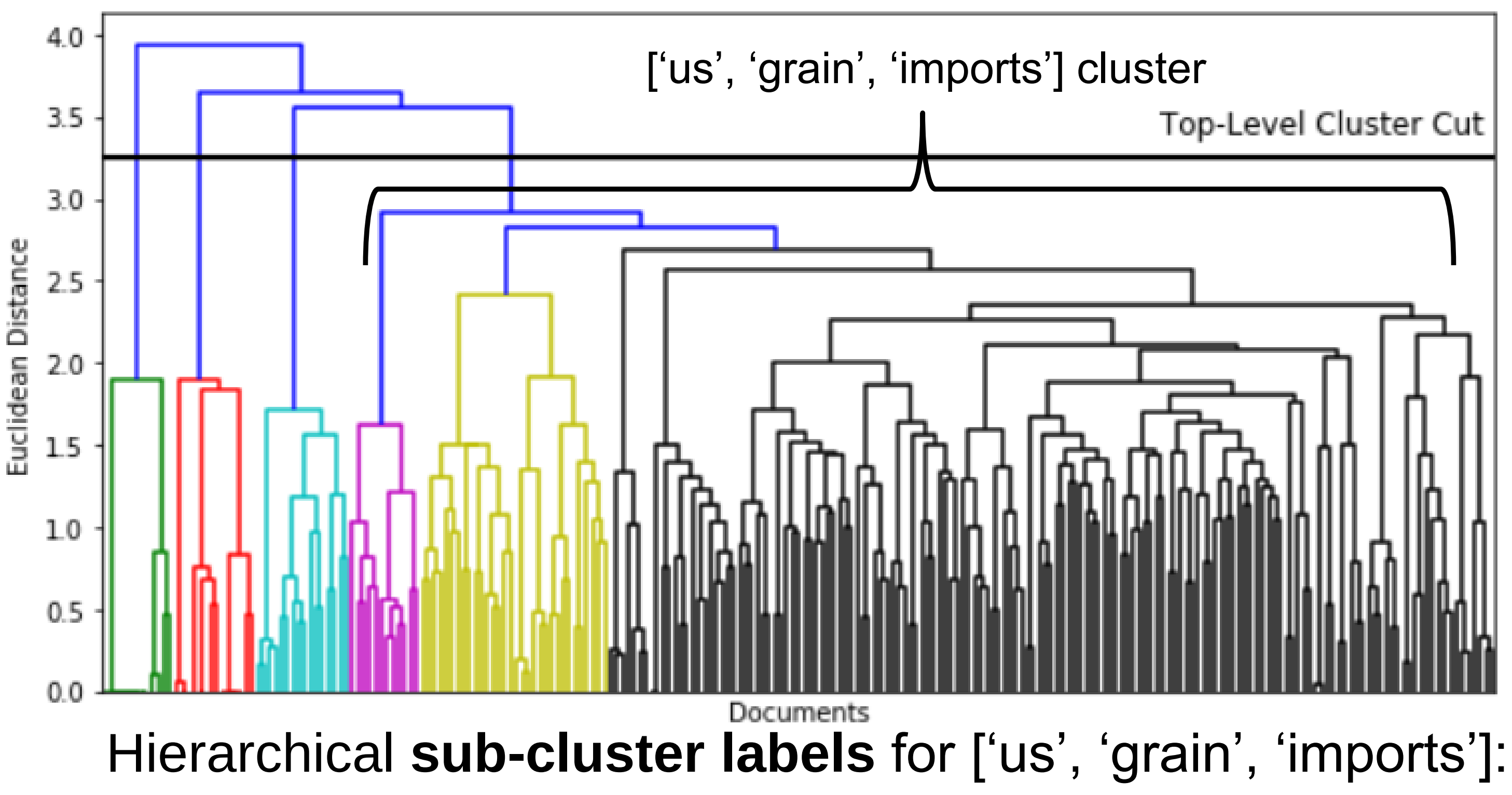
Results

	Hierarchical	Lingo
Silhouette Coefficient	0.42	0.14
Avg. Distortion	0.50	0.85
Avg. # Clusters / Search Term	4.97	3.89
Avg. # Documents / Cluster	21.88	30.07
Avg. % Docs w/o a Cluster	0	24.56

Lingo's silhouette and distortion negatively affected by documents in multiple clusters

Lingo "corn" labels (5 clusters): ['us', 'export', 'soviet'] ; ['soviet', 'acres', 'agreement'] ; ['inspections', 'price', 'bushels'] ; ['contract', 'stocks', 'futures'] ; ['trades', 'ago', 'gulf']

- 18% of documents in zero clusters
- 70% of remaining documents in at least 2 clusters



Hierarchical **sub-cluster labels** for ['us', 'grain', 'imports']:
['acres', 'acreage', 'program'],
['soviet', 'us', 'sale'],
['us', 'imports', 'grain']

Discussion of Results & Takeaways

Difficult problem:

- Difficult to create distinct clusters because documents already related within search: Strong thresholds needed
- Overall k heuristic creates inconsistency in quality. (ex. dendrogram implies "corn" has 6 natural clusters?)

Comparison of algorithms:

- Hierarchical: **larger number of small clusters**. Clusters well-formed. Labels sometimes overlapping, confusing.
- Lingo: **smaller number of large clusters**. Clusters cohesive, not distinct. Labels distinct and logical.

Hierarchal preferred: gives user agency to recognize overlapping clusters, explore sub-clusters.
Lingo too often leaves out documents and combines sub-topics in favor of distinct labels.

Future Work

- Use phrases as labels instead of single words (Suffix-Tree Stemming)
- Use semantic embedding of documents and labels to define similarity, eliminate overlap (word2vec, doc2vec)
- Consider how to decrease the computational complexity of hierarchical clustering at scale

1. Carpineto, Caludio et al. "Survey of Web Clustering Engines". *ACM Computing Surveys* 41(3). July 2009.
2. Osinski, Stanislaw et al. "Lingo: Search Result Clustering Algorithm Based on Singular Value Decomposition." *Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'03 Conference held in Zakopane, Poland, June 2-5.* January 2004