Oh That Looks Familiar: A Novel Similarity Measure for Spreadsheet Template Discovery

Ananad Krishnakumar*, **Vengadesh Ravikumaran***
Ekimetrics
*Equal contribution

Abstract

Traditional methods for identifying structurally similar spreadsheets fail to capture the spatial layouts and type patterns defining templates. To quantify spreadsheet similarity, we introduce a hybrid distance metric that combines semantic embeddings, data type information, and spatial positioning. In order to calculate spreadsheet similarity, our method converts spreadsheets into cell-level embeddings and then uses aggregation techniques like Chamfer and Hausdorff distances. Experiments across template families demonstrate superior unsupervised clustering performance compared to the graph-based Mondrian baseline, achieving perfect template reconstruction (Adjusted Rand Index of 1.00 versus 0.90) on the FUSTE dataset. Our approach facilitates large-scale automated template discovery, which in turn enables downstream applications such as retrieval-augmented generation over tabular collections, model training, and bulk data cleaning.

1 Introduction

Spreadsheets are ubiquitous in enterprises, yet collections are often messy and difficult to leverage at scale for LLM or ML applications. A key bottleneck is that similar spreadsheets—those following the same template or layout—are scattered across repositories, hindering automated processing and workflow integration.

We define spreadsheets as *similar* if they share consistent header arrangements, data regions, and content distributions. Organizing spreadsheets by similarity enables enterprises to treat template families as unified objects—critical for emerging applications like table-based RAG systems, automated data wrangling pipelines, and foundation model pretraining over structured data.

Existing methods for spreadsheet similarity vary in their approaches: content-based embeddings [4] focus primarily on semantic information while potentially overlooking layout structure, while graph-based approaches like Mondrian [12] capture topological relationships through structural graphs. We propose a hybrid cell-level distance metric that jointly encodes spatial positioning, type patterns, and semantic content. By combining Euclidean layout similarity with type-aware semantic matching and aggregation strategies (Chamfer [5] and Hausdorff [6] distances), our method effectively identifies spreadsheet template families for downstream processing.

The primary contribution of our paper is a **hybrid cell-level distance metric** for grouping spreadsheets into template families. We demonstrate superior unsupervised clustering performance compared to the graph-based Mondrian benchmark [12], achieving perfect template reconstruction.

2 Related Work

Our work intersects two primary research areas: spreadsheet representation methods and similarity measures. Prior spreadsheet understanding ranges from vision-based approaches [3, 2] to sequential

AI for Tabular Data workshop at EurIPS 2025.

models [10, 8, 13] and modern LLM encodings [14, 9, 11]. Modern representation methods have explored various encoding formats (Markdown, HTML, JSON) for large language models [14, 9, 11], providing foundations for our encoding methodology, though they focus primarily on content rather than structural patterns.

For similarity measurement, content-based methods [1] focus on semantics while potentially underweighting spatial structure. Graph-based approaches like Mondrian [12], our primary baseline, capture topology but exhibit a critical limitation: they consider content or structure independently. Our hybrid approach addresses this gap by jointly encoding spatial positioning, data types, and semantic content.

3 Methodology

We measure spreadsheet similarity through hybrid distance metrics combining spatial layout, type information, and semantic content.

3.1 Definitions

Definition 1 (Embedding). For a spreadsheet S with dimensions $m \times n$, we define its embedding as:

$$\Phi(S) = \{(i, j, t, s) : (i, j) \in [m] \times [n],\ t \in \mathcal{T},\ s \in \mathbb{R}^n\} \tag{1}$$

where $\mathcal{T}$ is a collection of data types $\{\text{Integer, Float, Date, String}, \ldots\}$. Each element $(i, j, t, s) \in \mathbb{N}^2 \times \mathcal{T} \times \mathbb{R}^n$ represents a non-empty cell at position (i, j) with data type t, where s is a vector representation of the semantic meaning encoded using *sentence-transformers/all-minilm-l6-v2* [15]. In our implementation, we map data types to integer encoding, details found in A. This representation captures spatial positioning, type structure and semantic meaning.

3.2 Cell-Level Distance (Hybrid Sub-Metric)

For two cells $u = (i_1, j_1, t_1, s_1)$ and $v = (i_2, j_2, t_2, s_2)$, we define:

$$d_c = w_{\text{spatial}} \cdot d_{\text{spatial}} + w_{\text{type}} \cdot d_{\text{type}} + w_{\text{semantic}} \cdot d_{\text{semantic}} \tag{2}$$

where $w_{\text{spatial}}, w_{\text{semantic}}, w_{\text{type}} \in [0, 1]$ and $w_{\text{spatial}} + w_{\text{semantic}} + w_{\text{type}} = 1$. d_c is a weighted average combination of the 3 dimensions.

Spatial component: Normalized Euclidean distance on cell positions

$$d_{\text{spatial}}(u, v) = \frac{\sqrt{(i_1 - i_2)^2 + (j_1 - j_2)^2}}{\sqrt{M_{\max}^2 + N_{\max}^2}} \tag{3}$$

where $M_{\max}$ and $N_{\max}$ are the maximum row and column dimensions across both spreadsheets. This ensures $d_{\text{spatial}} \in [0, 1]$ and makes distances comparable across different spreadsheet sizes.

Type component: Binary indicator for type mismatch

$$d_{\text{type}}(u, v) = \mathbf{1}_{t_1 \neq t_2} = \begin{cases} 0 & \text{if } t_1 = t_2 \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

Semantic component: Based on cosine similarity of cells

$$d_{\text{semantic}}(u, v) = \frac{1}{2} \cdot \left(1 - \frac{s_1 \cdot s_2}{\|s_1\| \|s_2\|}\right) \tag{5}$$

Since all components are normalized to $[0, 1]$, their average d_c is also a metric on $\mathbb{N}^2 \times \mathcal{T} \times \mathbb{R}^n$ with $d_c(u, v) \in [0, 1]$.

3.3 Spreadsheet-Level Distance (Aggregation Strategies)

Given spreadsheets S_1, S_2 with embeddings $\Phi(S_1) = \{x_1, \ldots, x_m\}$ and $\Phi(S_2) = \{y_1, \ldots, y_n\}$, we aggregate cell-level distances d_c to compute spreadsheet-level distances.Since d_c is a metric, all aggregation methods inherit metric properties (symmetry, non-negativity, triangle inequality) on non-empty spreadsheets. We evaluate two aggregation strategies— Chamfer distance and Hausdorff distance which differ in their matching approaches. Formal definitions are provided in Appendix B.

4 Experiments

We evaluate our embedding framework and distance measure through three complementary sections: clustering analysis, relative importance of dimensions and computational scalability. Due to compute constraints, we restricted our analysis to 133 randomly selected spreadsheets across seven template families (catalog products, census, countries metadata, product manycols, and sport season, strategic focus, and triathlon) from the FUSTE real-world dataset [12], providing a reproducible benchmark for template discovery evaluation. All code can be found here `https://anonymous.4open.science/r/spreadsheet-similarity-E286/README.md`.

4.1 Clustering Analysis

We evaluate the effectiveness of our structural embeddings for unsupervised organization of spreadsheet collection by comparing our method with the Mondrian method proposed by Vitagliano et al.[12]. Using k-medoids clustering with k = 7 clusters (matching the number of template families in our dataset), we assess how well each method recovers the original template structure.

Results: Table 1 presents clustering performance across all distance measures. Our Chamfer-based method achieved perfect cluster recovery (ARI = 1.00), substantially outperforming the Mondrian baseline (ARI = 0.90) in partition quality. While Chamfer's silhouette coefficient (0.64) is lower than Mondrian's (0.83), the perfect ARI demonstrates superior recovery of the true cluster structure. Our Hausdorff-based approach underperformed both methods with ARI = 0.61 and silhouette = 0.49, suggesting the measure's sensitivity to outliers and extreme points makes it less suitable for this template discovery task. These results demonstrate that Chamfer distance combined with our hybrid similarity metric provides superior template discrimination. The perfect ARI confirms that incorporating semantic and type dimensions alongside spatial features enables the model to capture the essential structural characteristics that define document templates.

Table 1: Clustering quality metrics using k-medoids with $k = 7$ clusters. Initial values are $w_{\text{semantic}} = 0.3$, $w_{\text{type}} = 0.5$ and $w_{\text{spatial}} = 0.2$. In 4.2, we vary them to study relative importance.

Distance Measure	**ARI** (Adjusted Rand Index) ↑	**Silhouette Coeff.**
Chamfer	**1.00**	0.64
Mondrian (benchmark)	0.90	0.83
Hausdorff	0.61	0.49

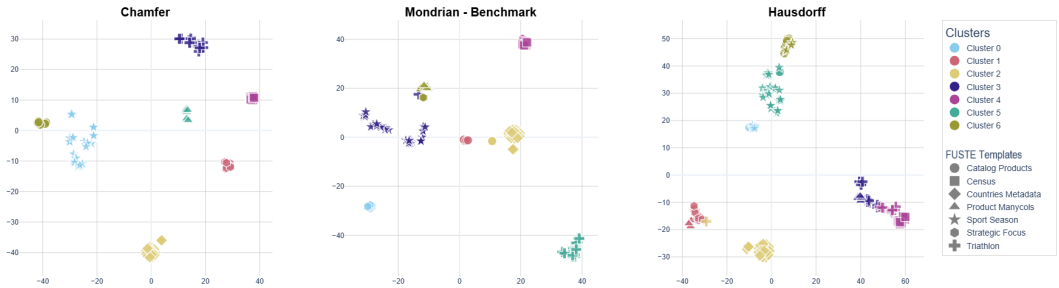
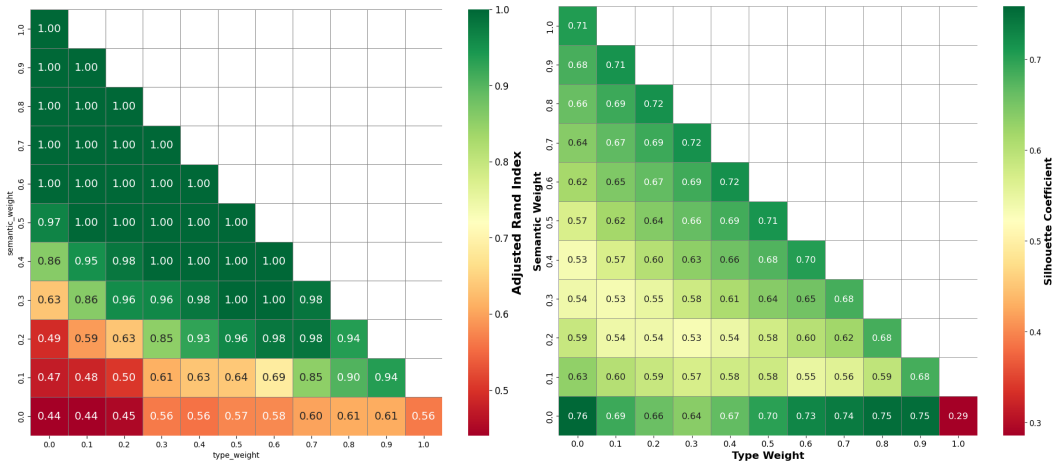


Figure 1: t-SNE projections (perplexity=9) of spreadsheet collections colored by k-medoids cluster assignments. Each subplot shows clustering results for a different distance measure, with point shapes indicating ground truth template families. Clear visual separation indicates successful cluster recovery.

4.2 Relative Importance of Dimensions

To understand relative importance of spatial, type and semantic information in template discovery, we vary $w_{type}, w_{semantic}$ across the feasible domain and compute ARI and silhouette coefficients.



(a) Adjusted Rand Index (b) Silhouette Coefficient

Figure 2: Clustering performance across weight combinations for type and semantic dimensions. White cells indicate invalid combinations where $w_{type} + w_{semantic} > 1.0$.

The heatmaps in Figure 2 reveal key insights. First, **both dimensions contribute independently**: performance improves along both axes, with type weight enhancing ARI from 0.44 to 0.61 at $w_{semantic} = 0.0$, and semantic weight driving ARI from 0.44 to 1.00 at $w_{type} = 0.0$. However, **semantic information is most important**: the rate of improvement is much steeper along the semantic axis, with ARI jumping from 0.49 to 0.85 when $w_{semantic}$ increases from 0.2 to 0.3 (at $w_{type} = 0.3$). Second, **the dimensions exhibit strong synergy at moderate weights**: type information amplifies the effect of semantic information (and vice versa). For instance, at $w_{type} = 0.2, w_{semantic} = 0.2$, ARI reaches 0.63, substantially higher than pure type (0.45) or pure semantic (0.49) at weight 0.2. The silhouette coefficient similarly shows complementary gains, increasing from 0.54 to 0.68 as type weight rises from 0.0 to 0.4 at $w_{semantic} = 0.3$. Third, **high semantic weight drives ARI performance**: once $w_{semantic} \geq 0.5$, ARI reaches near-optimal levels (ARI ≥ 0.97) regardless of type weight, demonstrating that semantic information alone is sufficient for accurate cluster recovery. However, **type weight enhances cluster cohesion**: the silhouette coefficient continues to improve with increasing type weight even at high semantic levels. For example, at $w_{semantic} = 0.7$, silhouette improves from 0.64 (pure semantic-spatial) to 0.72 (at $w_{type} = 0.3$), showing that type information contributes to tighter, more well-separated clusters. This reveals complementary roles: semantic information identifies the correct cluster assignments (external validity), while type information refines cluster quality and internal structure (internal validity).

5 Conclusion

This work presents a rigorous framework for quantifying similarity among spreadsheets through hybrid distance metrics that integrate spatial positioning with semantic-type information. Our primary contribution, **a novel hybrid distance metric**, showing superior clustering performance with ARI reaching 1.00, surpassing benchmark methods.

Limitations and Future Directions: Our current framework establishes a foundation for embedding structural and semantic information from spreadsheets. We plan to improve scalability by optimising implementation, and to explore extending the approach to additional structured documents such as presentation slides.

Broader Impact: This work provides a principled methodology applicable to automated document organization and retrieval, template discovery, and data format standardization across extensive document collections.

Our framework establishes that visual structural patterns intuitively recognized by humans can be systematically quantified and leveraged to enhance performance in clustering and classification tasks, creating new opportunities for applications where classifying spreadsheets plays a vital role.

References

[1] Christodoulakis et al. (2020) *Pytheas: Pattern-based Table Discovery in CSV Files*. *Proc. VLDB Endow.* doi:10.14778/3407790.3407810

[2] Deng, N., Sun, Z., He, R., Sikka, A., Chen, Y., Ma, L., Zhang, Y., & Mihalcea, R. (2024). *Tables as images? Exploring the strengths and limitations of LLMs on multimodal representations of tabular data*. arXiv preprint arXiv:2402.12424. `https://arxiv.org/abs/2402.12424`

[3] Dong, H., Liu, S., Han, S., Fu, Z., & Zhang, D. (2019). *TableSense: Spreadsheet table detection with convolutional neural networks*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 69–76). doi:10.1609/aaai.v33i01.330169

[4] Copul, R., Frost, N., Milo, T., & Razmadze, K. (2024). *TabEE: Tabular Embeddings Explanations*. In *Proceedings of the ACM on Management of Data* (Vol. 2, Issue 1). doi:10.1145/3654807

[5] Barrow, H. G., Tenenbaum, J. M., Bolles, R. C., & Wolf, H. C. (1977). Parametric correspondence and chamfer matching: Two new techniques for image matching. *IJCAI*. doi:10.5555/1622943.1622971

[6] Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi:10.1109/34.232073

[7] Dong, H., Cheng, Z., He, X., Zhou, M., Zhou, A., Zhou, F., Liu, A., Han, S., & Zhang, D. (2022). *Table pretraining: A survey on model architectures, pretraining objectives, and downstream tasks*. arXiv preprint arXiv:2201.09745. `https://arxiv.org/abs/2201.09745`

[8] Gol, M. G., Pujara, J., & Szekely, P. (2019). *Tabular cell classification using pre-trained cell embeddings*. In 2019 IEEE International Conference on Data Mining (ICDM) (pp. 230–239). IEEE. doi:10.1109/ICDM.2019.00033

[9] Li, P., He, Y., Yashar, D., Cui, W., Ge, S., Zhang, H., Fainman, D. R., Zhang, D., & Chaudhuri, S. (2023). *Table-GPT: Table-tuned GPT for diverse table tasks*. arXiv preprint arXiv:2310.09263. `https://arxiv.org/abs/2310.09263`

[10] Nishida, K., Sadamitsu, K., Higashinaka, R., & Matsuo, Y. (2017). *Understanding the semantic structures of tables with a hybrid deep neural network architecture*. In Thirty-First AAAI Conference on Artificial Intelligence. doi:10.1609/aaai.v31i1.10484

[11] Sui, Y., Zhou, M., Zhou, M., Han, S., & Zhang, D. (2023). *GPT4Table: Can large language models understand structured table data? A benchmark and empirical study*. arXiv preprint arXiv:2305.13062. `https://arxiv.org/abs/2305.13062`

[12] Vitagliano, G., Reisener, L., Jiang, L., Hameed, M., & Naumann, F. (2022). *Mondrian: Spreadsheet layout detection*. In Proceedings of the 2022 International Conference on Management of Data (pp. 2361–2364). doi:10.1145/3514221.3520152

[13] Wang, Z., Dong, H., Jia, R., Li, J., Fu, Z., Han, S., & Zhang, D. (2021). *TUTA: Tree-based transformers for generally structured table pre-training*. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 1780–1790). doi:10.1145/3447548.3467434

[14] Zhang, T., Yue, X., Li, Y., & Sun, H. (2023). *TableLLaMA: Towards open large generalist models for tables*. arXiv preprint arXiv:2311.09206. `https://arxiv.org/abs/2311.09206`

[15] Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982–3992). doi:10.18653/v1/D19-1410

==========

A Data Type Mapping

For the structural embedding $\Phi(S)$ defined in Definition 1, we map each cell's data type to an integer encoding to enable metric computation. Table A1 presents the complete mapping used throughout our experiments.

Table A1: Data type to integer encoding mapping.

Data Type	Encoding
Integer	0
Float	0
Percentage	0
Scientific Notation	0
Currency	0
Date	1
Time	1
Email	2
Other	3
String	4

The type detection follows standard spreadsheet conventions: numerical formats are grouped together (e.g., scientific notation, currency symbols), temporal data by date/time patterns, and emails by the presence of the @ symbol. The "Other" category captures non-empty cells that do not match any specific type pattern. This encoding ensures that the type component $d_{\text{type}}(u, v)$ in 4 operates on discrete categorical values while maintaining the metric structure required for our distance computations.

B Aggregation Strategies

Given spreadsheets S_1, S_2 with embeddings $\Phi(S_1) = \{x_1, \ldots, x_m\}$ and $\Phi(S_2) = \{y_1, \ldots, y_n\}$, we define 2 strategies for aggregating cell-level distances d_c into spreadsheet-level distances:

- **Chamfer Distance:** Bidirectional average nearest-neighbor distance

$$D_{\text{Chamfer}}(S_1, S_2) = \frac{1}{m} \sum_{i=1}^{m} \min_j d_c(x_i, y_j) + \frac{1}{n} \sum_{j=1}^{n} \min_j d_c(x_i, y_j)$$

- **Hausdorff Distance:** Worst-case nearest-neighbor distance

$$D_{\text{Hausdorff}}(S_1, S_2) = \max \left\{ \max_i \min_j d_c(x_i, y_j), \max_j \min_i d_c(x_i, y_j) \right\}$$

C Theoretical Properties

Proposition 1 (Bounded Distance). For any spreadsheets S_1, S_2 and aggregation method: $D(S_1, S_2) \in [0, 1]$.

Proof. Since $d_c \in [0, 1]$, Chamfer, and Hausdorff are convex combinations of d_c values.