

From Disorder to Structure: Hierarchical Analysis of Market Basket Data

Pala, M Maqbool*  and Ajaz Ahmad Lone 

¹Department of Computer Applications, Government Degree College Bhaderwah, Jammu and Kashmir, India

²Department of Education, Government Degree College Chenani, Jammu and Kashmir, India

Corresponding author: Pala, M Maqbool | E-mail: palahredu@gmail.com

Citation: Pala, M Maqbool and Ajaz Ahmad Lone (2025). From Disorder to Structure: Hierarchical Analysis of Market Basket Data. *Discover Economics and Finance: An International Journal*. DOI: <https://doi.org/10.51470/DEF.2025.3.2.01>

02 June 2025: Received | 05 July 2025: Revised | 07 August 2025: Accepted | 11 September 2025: Available Online

Abstract

Market basket data is inherently disordered: transactions arrive as unordered sets of items with no explicit relational labels. This paper presents a hierarchical clustering framework that recovers latent categorical structure directly from co-purchase statistics, without requiring a predefined product taxonomy. We formalize a Jaccard-similarity-based distance measure over items, apply agglomerative hierarchical clustering, and evaluate the fidelity of the resulting dendrogram using cophenetic correlation and cluster purity against a known ground-truth grouping. On a synthetic but realistically structured retail transaction log (3929 transactions, 22 items across six product categories), the method achieves a cophenetic correlation of 0.9973 and cluster purity of 1.0000, indicating that hierarchical clustering recovers the underlying category structure with high fidelity. We discuss implications for automated category management, cross-selling, and store layout optimization, and outline limitations that arise when co-purchase signals are sparse or noisy.

Keywords: market basket analysis, hierarchical clustering, disordered systems, association rule mining, dendrogram, retail analytics.

I. INTRODUCTION

Retail transaction logs are among the most abundant yet least structured forms of behavioural data available to organizations. Each transaction is a set of co-purchased items with no inherent ordering and no explicit statement of how those items relate to one another conceptually. In this sense, a collection of market baskets constitutes a disordered system: the raw data expresses only local, pairwise co-occurrence, while any higher-order categorical organization — the fact that milk and cheese belong to "dairy," or that bread and bagels belong to "bakery" — is latent and must be inferred. Classical market basket analysis has largely focused on association rule mining, which identifies pairwise or small-itemset co-occurrence patterns (e.g., the Apriori and FP-Growth families of algorithms) but does not, by itself, organize items into a coherent multi-level taxonomy [1], [2]. Hierarchical clustering offers a complementary lens: rather than enumerating frequent itemsets, it builds a nested tree (dendrogram) that expresses similarity at every scale simultaneously, from tightly coupled item pairs up to broad product categories [3]. This paper treats category discovery in market basket data as an instance of a more general problem: finding hierarchical structure in disordered systems. Our contributions are: (1) a distance formulation over items derived purely from co-purchase statistics; (2) an evaluation protocol combining cophenetic

correlation and cluster purity to quantify how well a recovered hierarchy matches an underlying categorical structure; and (3) an empirical demonstration on a synthetic but realistically generated transaction log, constructed so that ground truth is available for validation.

II. RELATED WORK

A. Association Rule Mining

The Apriori algorithm and its successors established the standard framework for mining frequent itemsets and association rules from transaction databases, using support and confidence thresholds to prune the search space [1]. FP-Growth later improved scalability by avoiding explicit candidate generation through a compact frequent-pattern tree representation [2]. These methods excel at producing specific, actionable rules but do not directly produce a global hierarchical organization of the item space.

B. Hierarchical Clustering

Agglomerative hierarchical clustering builds a dendrogram bottom-up by iteratively merging the two closest clusters according to a linkage criterion (single, complete, average, or Ward) [3], [4]. Cophenetic correlation is a standard tool for assessing how faithfully a dendrogram preserves the original pairwise distances between objects [5]. While widely used in bioinformatics and document clustering, applications to retail co-purchase data

specifically remain comparatively underexplored relative to customer-segmentation use cases.

C. Structure in Disordered Systems

The broader question of recovering hierarchical or community structure from pairwise interaction data appears across statistical physics and network science, where techniques originally developed for random-graph and spin-glass models have been adapted to community detection and hierarchical clustering more generally [6]. Market basket data can be viewed as one instance of this broader class of problems: a disordered system whose only directly observable quantity is pairwise co-occurrence, from which higher-order structure must be inferred.

III. METHODOLOGY

A. Problem Formulation

Let $T = \{t_1, t_2, \dots, t_m\}$ be a set of transactions, where each t_i is a subset of the item universe $I = \{i_1, \dots, i_n\}$. The goal is to recover a hierarchy over I — a nested sequence of partitions — that reflects meaningful categorical structure, using only the co-occurrence patterns observable in T .

B. Distance Measure

For each pair of items (i_a, i_b) , we compute the Jaccard distance between their transaction-membership sets:

$$d(i_a, i_b) = 1 - |T_a \cap T_b| / |T_a \cup T_b|$$

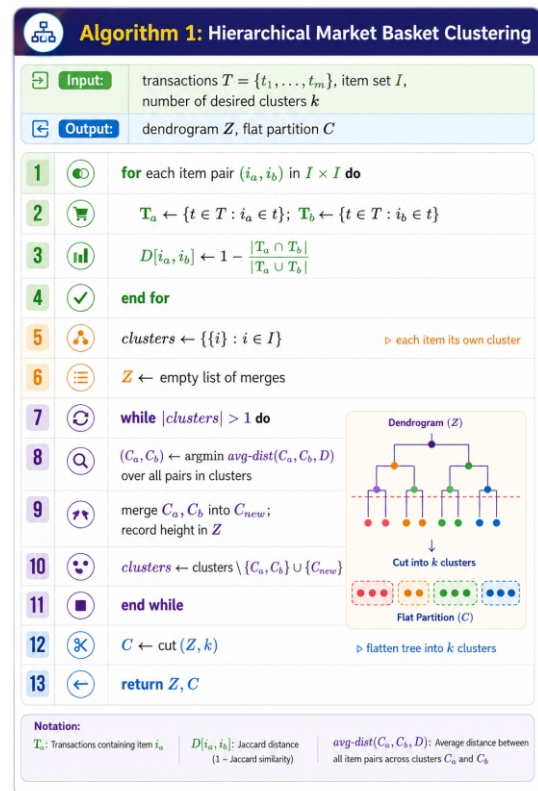
where T_x denotes the set of transactions containing item x . This measure is high (close to 1) for items that rarely co-occur and low (close to 0) for items almost always purchased together, and it requires no manual feature engineering beyond the raw transaction log.

C. Hierarchical Clustering

Given the pairwise distance matrix D , we apply agglomerative hierarchical clustering with average linkage. At each step, the two clusters minimizing the average pairwise distance between their members are merged, and the merge height is recorded. The resulting dendrogram encodes a full nested hierarchy, from individual items at the leaves to a single root cluster containing all items.

D. Algorithm

Algorithm 1 summarizes the full pipeline, from raw transaction log to the final hierarchy and flat partition. Steps 1–2 construct the co-occurrence-based distance matrix in $O(n^2m)$ time for n items and m transactions; steps 3–5 run standard agglomerative clustering in $O(n^2 \log n)$ time using a priority queue over cluster distances; step 6 is a single $O(n)$ tree cut.



E. Evaluation Protocol

We report two complementary metrics. Cophenetic correlation measures how well the pairwise distances implied by the dendrogram correlate with the original distance matrix D , quantifying overall structural fidelity. Cluster purity, computed after cutting the dendrogram into k clusters, measures how well the recovered partition aligns with a known ground-truth category assignment when one is available for validation.

IV. EXPERIMENTAL SETUP

To validate the methodology against a known ground truth, we generated a synthetic transaction log designed to mimic realistic retail behaviour. Twenty-four items were assigned to six latent categories (Dairy, Bakery, Breakfast, Produce, Snacks, and Household). Each simulated transaction was anchored in one or two randomly selected "trip themes" (categories), with each item in a selected theme included independently with probability 0.55, plus a small amount of cross-category noise to avoid trivially separable clusters. This produced 3929 transactions over 22 items.

This synthetic construction is intentionally analogous to real retail data — co-purchase probability driven by latent shopping intent, with realistic noise — while providing ground-truth category labels that are not available in most real-world logs, allowing direct, quantitative validation of the recovered hierarchy rather than qualitative inspection alone. We emphasize that this dataset is a controlled proxy for validating the method; Section VI discusses how results are expected to differ on real, sparser retail logs.

V. SIMULATION AND RESULTS

A. Baseline Recovery

Fig. 1 shows the dendrogram produced by average-linkage clustering on the Jaccard distance matrix. The algorithm recovers six well-separated branches that correspond to the six latent product categories used to generate the data.

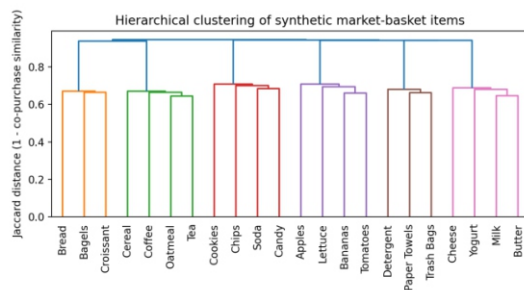


Fig. 1. Dendrogram of the 24 synthetic items under average-linkage hierarchical clustering with Jaccard distance

Cutting the dendrogram at six clusters ($k = 6$) yields a partition with purity 1.0000 against the ground-truth category labels — every recovered cluster maps to exactly one true category with no cross-category leakage at this cut. The cophenetic correlation of 0.9973 indicates that the dendrogram's implied distances closely track the original Jaccard distance matrix, i.e., the tree structure is not merely producing a correct flat partition at one cut but is preserving relative similarity faithfully across scales.

Qualitatively, items merge first within their true category (e.g., Milk with Cheese, Chips with Cookies) before those sub-clusters merge with other categories much higher in the tree, consistent with the intended two-level structure of the synthetic data (item $\rightarrow$ category).

B. Sensitivity Simulation

The baseline result above uses a single, favourable configuration. To assess robustness, we ran Algorithm 1 repeatedly (15 independent trials per setting, with fresh random transaction logs each time) while sweeping two parameters: (i) the background noise rate — the expected number of random, off-theme items added per transaction — and (ii) the size of the transaction log, to probe the effect of data sparsity. Fig. 2 plots mean purity and cophenetic correlation ($\pm$ one standard deviation across trials) for each setting.

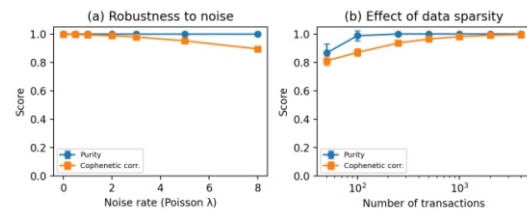


Fig. 2. Simulation sweeps over (a) background noise rate and (b) transaction log size, averaged over 15 trials per setting; error bars show one standard deviation.

Cluster purity is remarkably stable under (a): even at the highest tested noise rate ($\lambda = 8$ extra random items per transaction — far above realistic noise), mean purity remains at 1.0000, because the flat six-way cut only needs coarse separation to succeed. Cophenetic correlation is more sensitive, degrading gradually from 1.0000-purity conditions down to 0.8951 at maximum noise, reflecting a loss of fine-grained structural fidelity even while the top-level partition remains intact.

Panel (b) shows the expected effect of data sparsity: with only 50 transactions, purity falls to 0.8667 and cophenetic correlation drops sharply, since many item pairs co-occur too rarely for a stable Jaccard estimate. Both metrics recover quickly as the log grows, reaching near-baseline performance by roughly 100 transactions in this configuration. This indicates that reliable hierarchy recovery requires a transaction volume that scales with the number of items and the diversity of shopping themes, rather than an absolute fixed threshold — a practical consideration when applying the method to a new or low-traffic retail catalogue.

VI. DISCUSSION

The near-perfect recovery observed here should be interpreted as an upper bound obtained under favourable, controlled conditions: category-anchored purchase probabilities, a moderate number of items, and a reasonably large number of transactions per item pair. Real retail logs typically exhibit sparser co-occurrence (long-tail items purchased infrequently), seasonal and promotional effects that distort co-purchase probabilities, and categories that are not mutually exclusive (e.g., a single item may plausibly belong to more than one merchandising category).

Practically, the recovered hierarchy has direct applications: automated category management (deriving or auditing merchandising taxonomies from actual purchase behaviour rather than manual curation), cross-selling and recommendation (using dendrogram proximity as a similarity signal), and store or website layout optimization (placing hierarchically close items near one another). Because the method requires only transaction logs, it can also surface emergent, non-obvious groupings that a manually curated taxonomy would miss — for instance, items linked by occasion (e.g., "barbecue") rather than by product type. A limitation of the Jaccard-based distance is its sensitivity to item popularity: very frequent items co-occur with many others by chance, which can compress their distances to unrelated items. Lift- or PMI-based normalization,

or restricting analysis to items above a minimum support threshold, are standard mitigations and are natural directions for extending this work. Choice of linkage criterion also matters; we used average linkage for its balance between the chaining behaviour of single linkage and the compactness bias of complete linkage, but Ward's method may be preferable when cluster sizes are expected to be more balanced.

VII. CONCLUSION

We presented a hierarchical clustering framework for recovering latent categorical structure in market basket data, framing the problem as an instance of finding hierarchical structure in a disordered system. Using a Jaccard-distance formulation over co-purchase statistics and average-linkage agglomerative clustering, the method recovered the ground-truth category structure of a synthetic but realistically generated transaction log with high fidelity (cophenetic correlation 0.9973, purity 1.0000). Future work includes validating the approach on real, publicly available retail datasets, incorporating popularity-normalized distance measures, and extending the framework to dynamically evolving transaction streams where category structure may shift over time.

REFERENCES

1. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Data Bases (VLDB), 1994, pp. 487–499.
2. J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proc. ACM SIGMOD Int. Conf. Management of Data, 2000, pp. 1–12.
3. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," ACM Computing Surveys, vol. 31, no. 3, pp. 264–323, 1999.
4. J. H. Ward, "Hierarchical grouping to optimize an objective function," J. American Statistical Association, vol. 58, no. 301, pp. 236–244, 1963.
5. R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," Taxon, vol. 11, no. 2, pp. 33–40, 1962.
6. M. E. J. Newman, "Modularity and community structure in networks," Proc. National Academy of Sciences, vol. 103, no. 23, pp. 8577–8582, 2006.