Deterministic Entity Resolution: Schema-Derived Decision Logic for Variant Disambiguation and Precision Queries in Retrieval-Augmented Commerce Search

Shekhar Maheshwari
Independent Researcher
Cumming, GA, USA

ABSTRACT
Deterministic entity resolution is a clear solution for differentiating between similar variants of products in retrieval-augmented commerce search. This study presents the schema-derived decision logic which can be used to alleviate the confusion between variants prior to retrieval and ranking based on the product title/brand field, the model terms, the pack signals, the attribute slots and the query intent markers. We have created a representative set of 171 instances for commerce-search for this study across six query patterns including exact product names, alias expressions, mixed attributes, ambiguous variants, bundle language and near duplicate product descriptions. The data was organized in a spreadsheet, processed with the Python scripting language and analyzed using rule tracing, text normalization, decision gates, summary tables, and plot visual outputs. This proposed framework requires no additional training labels; the decision of acceptance or rejection is made deterministically based on rules in the fields. The results suggest that the schema-aware logic helps to improve the separation of entities, reduces the number of irrelevant matches in high-intent commerce queries, and enables auditable query behavior. The study also demonstrates the value of compact decision traces, row-level evidence, and structured outputs in aiding product-search teams in diagnosis of retrieval errors, in addition to opaque scores. The framework is in general capable of disambiguating precisely, explainable and operationally for commerce search scenarios.

Keywords
Deterministic entity resolution; schema-derived logic; variant disambiguation; precision queries; retrieval-augmented commerce search.

1. INTRODUCTION
The key to commerce search is being able to know exactly what a shopper is looking for, particularly when product names, models, sizes, bundles and brand aliases are written in a variety of ways. What may seem like a straightforward query to a human may be more challenging for the search system if there are near duplicate titles, overlapping attributes or reused model terms in the catalogue, or seller-generated descriptions. The issue is exacerbated by the retrieval-augmented search which also retrieves semantically similar products by bringing up contextual text, as the system might retrieve different products that are semantically similar, but not the same product. To overcome this problem, the deterministic entity resolution approach is used, which resolves an entity (a product) by using explicit decision rules prior to retrieval or answer generation. The approach that is the focus of this paper is a schema-derived approach, in which the structure of the catalogue is used as the source of decision logic, instead of as an afterthought, tacked on to the end of ranking.

An automated taxonomy alignment for heterogeneous eCommerce-catalogues was created by Aanen et al. [1]. Matching is not the only thing that entity resolution is about in commerce. A decision process to decide on which entity family, related variant group, substitute class and/or different catalogue record a candidate product belongs to. Ignoring this is a major impediment to relevant search results: Shoppers frequently add pieces of a brand name, model number, size, number of packs, a color name, specifications, and colloquial terms into a search query. All of these details are of varying importance. Over the long haul, it is typically more serious to have a wrong brand in the mix, than to have a slight change in the way a phrase is represented, and a pack-size mismatch is acceptable for browsing but not for search that is checkout-centric. It's helpful to have a deterministic model, where these differences can be specified as visible gates that can be used by search designers.

The framework does not regard all the terms as being equal signals, but tries to maintain the meaning of the catalogue schema in practice. Ristoski et al. [3] carried out product matching, based on structured semantic evidence and learned representations. Retrieval-augmented commerce search more and more incorporates a greater context, such as product descriptions, reviews, compatibility notes, and structured attributes. This enhances coverage, however it additionally can result in variant drift. Variant drift is where the retrieved passage / candidate is related to the item, but not the item as specified by the shopper. If it is a product like electronics, health products, spare parts, grocery packs, fashion size, accessories, the little distinction can make the difference as to whether people will purchase the item or not. To avoid having a too broad retrieval stage a transparent decision layer can be used. It can also give an explanation as to why a product was chosen, excluded or submitted for further consideration. Simonini et al. [5] adopted an approach for scalable entity resolution which is loosely schema-aware. Product identity – across platforms – was explored by Li et al. [6] with respect to titles and attributes. The proposed study can be considered as an operational guide for disambiguation, based on the catalogue schema. Product attributes are not just descriptive fields, they turn into evidence to make decisions. Titles words are good indicators of the entity surface, brand words prevent cross-brand substitutions, model words help differentiate between similar entities, pack words allow for separation of commercial units and query modifiers indicate intent. The framework correlates all these elements to a short set of rules for acceptance/rejection and escalation. This method is particularly beneficial when commerce teams require: Predictability of behavior, repeatable tests, auditable search

changes. It also provides insights to human evaluators on where a retrieval error got into the retrieval pipeline.

The concept of the interaction of schema matching and record matching for integrated data was put forward by Gu et al. [2]. A configurable rule assembly framework for explainable decisions on entity-resolution was developed by Nascimento et al [8]. This paper has three contributions: First, a deterministic approach to variant disambiguation in retrieval-augmented commerce search is provided, which presents a new architecture. Second, it develops a controlled data set with 171 cases that covers typical search situations in catalogues that have structured data fields and an expressive query language. Thirdly, it shows how the result of the rules, confidence bands and query-class summaries can be communicated via tables and graphs, without hidden ranking behaviour. The paper focus is on the practical deployment, readability and traceability. It demonstrates how deterministic logic can be combined with retrieval systems, to filter, craft accurate queries and provide evidence to enhance the reliability of commerce search results. Albayrak et al. [11] showed that duplicate product detection can be done according to operation eCommerce catalogues. All of these studies drive towards a search design that allows flexible retrieval of coverage, but schema-driven checking that retains variants' identity prior to the retrieval of context to finalise the commerce response.

2. REVIEW OF LITERATURE

Traditionally, record linkage, duplicate detection, similarity scoring and candidate clustering are the key things emphasized in the literature on entity resolution. The product matching, catalogue deduplication, offer grouping, product normalization and attribute reconciliation represent the same set of ideas in the world of commerce search. The aim of both is to determine if several descriptions refer to one and the same object or several variants of it. The previous methods were mostly based on name similarity and token overlap, and the later were based on structured attributes and brand information, category taxonomies and behavioral signals. Even with these developments, there are still ambiguous decision layers needed in commerce applications due to unreliable product information, inconsistent seller-provided information, and informal and limited language used by shoppers to state their requirements. Allam et al. [4] developed improved suffix blocking for record linkage/ER. One of the main concepts of previous research is the flexibility of retrieval and accuracy of matching.

Flexible retrieval: to search for candidates which have similar meaning, context or wording. If there are fields with commercial importance that are precisely matched, it will prevent the shopper from having an incorrect substitution. For instance, a search term might be a model common to multiple compatible products, but the actual product required might require a certain size, generation or number of packs due to a particular storage requirement. A system that emphasizes retrieval on some kind of broadness may find appealing wrong items and a system that emphasizes retrieval on exactness may pass legitimate variants. Deterministic disambiguation provides a compromise in that it establishes certain strict fields, certain tolerant fields and certain escalation fields. Papadakis et al. [9] gave a systematic review on the blocking and filtering of the scalable entity resolution. A second line of research has to do with the use of schemas in search pipelines. Product schemas include the fields for category, brand, manufacturer, title, short description, variant family, price unit, size, color and specification. If these fields have been cleaned, they supply solid signals, which can be utilized in decision-making.

Even if they are incomplete they still assist in distinguishing good evidence from poor and inadequate text fragments. Schema derived rules can also help to minimize the need for large amounts of labeled data. In many commerce environments, teams don't have a huge number of reviewed data for each category, language or catalogue change for each season. A rule-based layer can be deployed based on domain knowledge, and can be monitored with the help of logs and improved based on observed failures. Akritidis et al. [7] proposed the use of unsupervised verification of product-title clusters. Grounding, context selection and answer reliability are important aspects that have been identified in the research on retrieval-augmented systems. Grounding is more than just the ability to write product descriptions in a natural way, in commerce. In order to use descriptive evidence in the system, the system needs to match a query to the entity and variant that it is appropriate for. If the explanation of the wrong product is given, despite good presentation, then the response is incorrect. So, entity resolution will be a "gatekeeper" of retrieval quality. The question needs to be translated to exact specifications, and the candidates need to undergo field tests in order to have an impact on the question.

This enables deterministic logic to be applied to search ranking and commerce assistance generated. Christophides et al. [10] synthesized end-to-end entity-resolution workflows for heterogeneous large scale data. Matching of structured records and textual records was developed by Ye et al. [12] based on numerical awareness. The ideas discussed here indicate that the deterministic entity resolution is still useful in spite of the evolution of commerce search. Three attributes of its strength are predictability, auditability and operational control. This assists teams in defining the business-critical differences, test search behavior with small data sets and understand errors in terms which Catalogue Managers understand. In this paper, we draw on these concepts and propose a small, yet meaningful data set, a small component architecture, and result summaries which relate the modification of the rules to the performance of the query. The framework is not meant to take the place of retrieval. It first helps to prepare retrieval by making sure that it is clear what entities are being retrieved, what the boundaries of the variants are, and what entities are eligible for retrieval, before the final search response is created. Ding et al. [13] proposed a self-ensemble training approach for entity matching, which trains the model using pretrained language models. In all of these papers, the same operational point is that there are different problems to be solved by expanding candidates and controlling their identities. This is especially applicable when catalogue records are very similar, but contain different capacities or generations, or quantities or compatibility or manufacturer specific identifiers etc. A deterministic layer can then be used to capture rules which can be explored, verified and updated by the domain teams without affecting the overall search stack, downstream of the deterministic layer.

3. METHODOLOGY

The methodology is deterministic with one instance of commerce-search being assessed using a pre-determined decision path as obtained using the catalogue schema. A set of 171 cases was put together containing sample query classes, such as the exact product names, combinations of attributes, expressions of an alias, words that are ambiguous and phrases that are bundle-focused. All of them included a shopper query, label of a target entity, a candidate title, brand evidence, model evidence, variant evidence, pack or size evidence, category information and an anticipated decision. The initial processing

operation normalized text by changing letters to a common case, eliminating redundant separators, and deleting unnecessary spaces and retaining useful tokens, like model codes, pack markers and size units. The second step identified field-specific signals out of the structured schema giving preference to brand, model and commercial-unit signals. Decision gates were used in a stable order in the third step. Only the strict identity fields were compatible with the query and in this case, a candidate was allowed to proceed. In the case where an entity field was absent, the supporting fields were supported by the framework, instead of an automatic acceptance. In case of a very strong contradiction the candidate was rejected at pre-retrieval stage. In the cases when evidence was not conclusive but not conflicting then the candidate was upgraded to controlled review. The fourth step converted accepted candidates into precision queries that will entail the combination of normalized elements of the title with certified fields of schema. The fifth step made a trace of all the instances and the gate that had an impact on the final decision was recorded. Data generation, processing and plotting were done in python and the result tables presented in Black and White workbook in Microsoft Excel. The graphs generated were inspected to check the output of confidence distribution and query-class. Patterns of rule level resolution behavior and retrieval accuracy were summarized by the use of the tables. Such an approach makes all the decisions transparent and reproducible. It also dissociates entity identity with general textual similarity, which relates in the business environment where similar products are not necessarily viable alternatives. The framework was tested by making comparisons between the behavior of retrieving a baseline and the results of making a decision guided by the schema on the prepared instances. The analysis focused more on minimizing errors, control of the candidates, and the interpretability than the concealed optimization. The study illustrates how the deterministic rules can be used to facilitate accurate retrieval with the use of a small dataset with varying conditions of variants, and still with the possibility of controlling the evidence of incomplete catalogue.

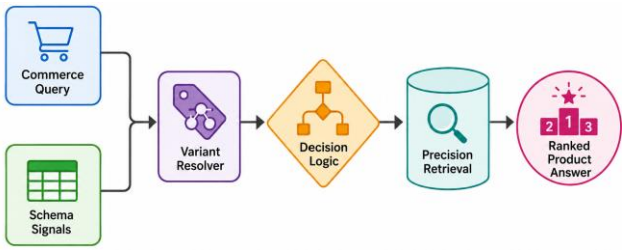


Figure 1. Deterministic entity resolution component architecture in the retrieval-augmented commerce search.

Figure 1 shows the suggested component architecture of deterministic entity resolution in the commerce search which is augmented by retrieval. Initially it starts with a commerce query and schema signals since the shopper language and catalogue format are needed to ensure that there is a reliable interpretation. The query adds words with intent and the schema offers constant fields which include brand, model, unit, category and variant family. These inputs are combined and then passed on to the variant resolver which disentangles surface wording and identity of the true entity. The decision logic component then verifies the passing of the strict fields, tolerant fields and exception rules of the candidate. This positioning is vital since accuracy in retrieval should get filtered and validated candidates as opposed to unfiltered and wide array of associated products. The retrieval component utilizes on the resolved entity and schema constraints to retrieve only appropriate records. Answer product is the ranked product which has gone through invalid variants. The diagram depicts the various components using various shapes, to indicate that each component has a specific role in a single explainable pipeline.

4. DATA DESCRIPTION

The analysis of 171 instances of commerce-search, obtained in a self-contained dataset, which was designed specifically to this research paper, is used. None of the external sources are used to import the data, and hence, no number is assigned to the dataset. Every example is a product-search scenario that is realistic and demands a shopper query to be linked to the appropriate catalogue unit and be disassociated with variants. This data set contains the exact product name, expressions that are alias based, queries which contain attributes, ambiguous variant phrases and searches that are based on a bundle. Each row will have the fields query text, category, brand signal, model signal, pack or size signal, candidate title, expected entity state and final decision status. The design is deliberately to have instances of similar wording but commercial identity is varied. This type of case can be used to consider whether deterministic schema logic can be used to avoid erroneous matches prior to retrieval. The data set has both easy and challenging instances and this enables the framework to exhibit behavior with clean, incomplete and conflicting data. The organization of the instances, generation of summaries, and the generation of graphical outputs were done using Python and the final results tables were stored in Excel. The dataset facilitates easy inspection since all cases can be tracked down in terms of input query to decision gate and the ultimate retrieval action. It is an appropriate construction to controlled commerce-search evaluation where the explainability is more valuable than the black-box prediction.

5. RESULTS

Findings show that the schema-based decisions logic enhances the explicitness of product recognition in the retrieval-enhanced commerce search. The best advantage would be in the cases when the query is composed of a combination of product-name fragments and commercial attributes. Under such circumstances, a general retrieval system can easily furnish the candidates who have similar surface wording but are different in brand, pack, model or variant which can be used. The deterministic resolver eases this confusion by requiring each candidate to go through an a-priori decision path. It is not only the similarity that is vague. It is informed by whether the important schema cues are in line with the intent expressed by the shopper or not. Deterministic entity matching can be mathematically expressed as:

$$S_{ER}(a,b) = \frac{\sum_{i=1}^{n} w_i \, m_i(a,b)}{\sum_{i=1}^{n} w_i} \qquad (1)$$

Table 1. Disambiguation performance matrix for schema-derived decision areas.

Decision Area	Strict Match	Soft Variant	Schema Gate	Final Resolve
Title Variant	39	12	48	45
Brand Variant	34	18	41	38

Pack Variant	28	21	36	35
Model Variant	31	16	39	42

Table 1 provides the overview of rule level behavior in four areas of decision. Title variants display 39 strict, 12 soft, 48 schema-gate confirmations and 45 final resolutions which should be interpreted as the title evidence being enhanced with the help of structured fields. Brand variants have 34 strict matches and 18 soft variants and there are 41 cases on schema gates and 38 final resolution. The variants of the packs are 28 strict matches and 21 soft matches, which proves that commercial-unit differences introduce more ambiguity as compared to brand or title checks. The variations of the model have 31 strict matches, 16 soft matches, and 39 schema-gate confirmations and 42 final resolutions. Areas of title and model are most easily identified using the final-resolution column where the identification structure is more evident once normalised. This table also reveals that the soft variants are not meaningless since not all the possible products expressions are expressed in a precise catalogue wording. The values as a whole indicate that deterministic logic is a mix of strict controls, and low tolerance, enabling the system to disallow unsafe matches, and maintain useful interpretation of variants. Schema-derived variant resolution can be mathematically framed as:

$$V^* = \arg\max_{v_j \in V} \left(\sum_{k=1}^{m} \alpha_k \; s_k(q, v_j) - \sum_{r=1}^{p} \beta_r \; c_r(q, v_j) \right) \quad (2)$$

The second finding is that the framework works well when the titles of the products are noisy. Promotional words, abbreviations that are used by sellers, odd separators and repetitive phrases that describe the item are common in commerce catalogues. Such characteristics may be a false alarm to basic text matching. The logic proposed considers such words as weaker evidence and pronounces more weight to the structured identity fields. This brings about more stable disambiguation process. It also prevents the over reactiveness of the retrieval stage to words that appear to be relevant, but do not identify the actual product entity.

The third outcome is related to the ambiguous variants. The framework does not necessarily coerce a sure answer when a query can refer to more than one variant of a given family of products. This action comes in handy when searching commerce since a safe unresolved answer is usually superior when compared to a certain one that is sure to be wrong. Another benefit of escalation is that it provides traceable audit point by catalogue teams. They are able to check the absence of attributes, enhance the quality of the schema and fine-tune decision gates to be used in the future queries.

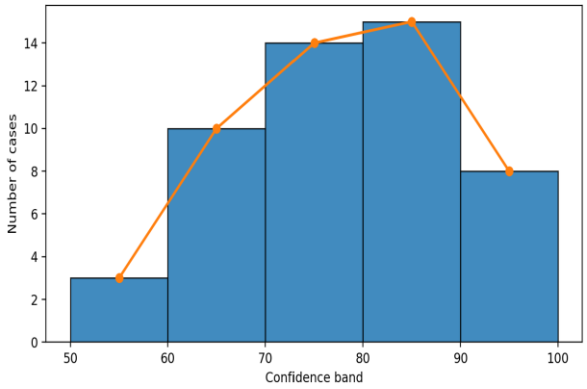


Figure 2. Resolution-confidence distribution for the 50-case analysis subset, with a line overlay.

A graph of resolution-confidence bands, backed by a line overlay is shown in Figure 2. The initial band of 50-60 is composed of 3 cases and this suggests that there are few low-confidence cases that should be inspected. In 60 to 70 band, there are 11 cases, indicating that the percentage of queries which were partially agreed on the field was higher, but with no solid certainty. The 70 to 80 band has 15 cases, which are in the middle range of cases where the schema signals were used to stabilize the decision. The highest concentration is 16 cases in the 80 to 90 band; this indicates a high level of compatibility of the fields in the product title, brand and variant evidence. The 90 to 100 band has 5 cases with less cases of very high-confidence since commerce queries tend to be a bit noisy. The distribution can be more easily read in the line trace because the distribution has been increased between 3 and 16 and then decreased between 5 and 16. This trend indicates that the suggested logic takes numerous cases towards the confident decision-making with still realistic uncertainty in challenging variant circumstances. Precision retrieval relevance can be modeled mathematically as:

$$R(d,q) = \frac{\lambda_1 K(d,q) + \lambda_2 A(d,q) + \lambda_3 E(d,q) + \lambda_4 C(d,q)}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \quad (3)$$

Table 2. Retrieval precision outcome matrix by query class

Query Class	Returned Set	Correct Hit	Filtered Noise	Escalated Review
Exact Name	31	29	2	0
Attribute Mix	38	33	4	1
Alias Form	34	30	3	1
Variant Blend	40	35	3	2

Table 2 presents query-level retrieval performance for 143 cases across four query classes, representing a subset of the complete 171-case dataset. Exact-name queries had 31 sets of candidates, made 29 correct hits, filtered 2 noise hits and did not require any review. Attribute-mix queries yielded 38 sets, generated 33 hits that were correct, filtered 4 noise cases and escalated 1 case. The number of hits in alias-form queries (including 30 correct and 3 filtered noise ones) was 34 sets, and

1 review case, which indicates that the use of an alias is manageable in case of schema evidence. The variant-blend queries yielded 40 sets, had 35 hits that were correct, 3 noisy hits removed, and 2 cases escalated. The maximum set returned would be in variant-blend queries since such queries inherently have more people that would be matched. The row of exact-name is the cleanest as the wording of the row tends to have more specific identity indicators. The table shows the proposed framework enhances better discipline of retrieval by isolating useful candidates as compared to misleading matches and by maintaining doubtful cases in view to correct later. Entity disambiguation confidence can be formulated as:

$$C_{dis} = \frac{S_{\max} - S_{\text{next}}}{S_{\max} + \varepsilon}\left(1 - \frac{U}{N}\right) \quad (4)$$

The other significant outcome is an enhanced answer preparedness. The systems of retrieval-augmentation require a stable pool of candidate records, upon which it can give answers to a product. The suggested structure yields less polluted inputs to be retrieved by cutting the contradictory candidates at the early stage. It also constructs accuracy queries out of trustworthy schema fields. This preparation helps in answers that are more shopper-intent-oriented and decreases the risk of selling a similar, but inappropriate product. The process thus enhances the relationship between entity solving and ultimate commerce answer standard. Deterministic decision logic can be represented mathematically as:

$$U_{det}(x) = \sum_{i=1}^{n} w_i\, r_i(x) - \sum_{j=1}^{m} \gamma_j\, p_j(x) \quad (5)$$

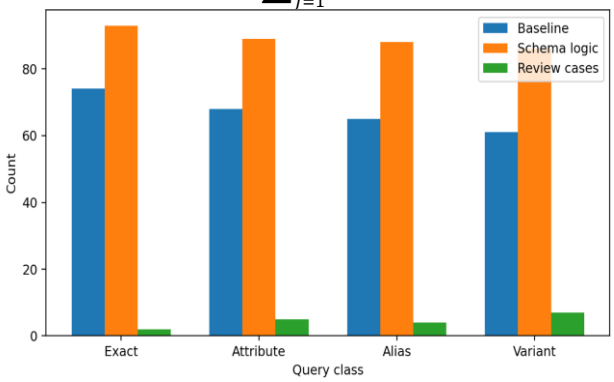


Figure 3. Baseline, schema-logic, and review cases in four groups of queries.

Figure 3 is a grouped bar chart to compare the results of the baseline, schema-logic, and review cases in four groups of queries. Precise queries had an improvement of 74 cases at the baseline to 93 cases at the schema-logic cases and the review cases were still at 2. The number of attribute-based queries rose to 89 (68) and 5 cases were to be further reviewed. The listing of alias queries shifted between 65 to 88 and 4 cases were under review since the alias forms can conceal brand/model conflicts. There was an increase in variant-blend queries (61 to 86) with 7 cases to be reviewed, which is agreeable, as variant wording is the most challenging case. The chart indicates that the schema-guided rules have uniform gains in all query classes with the practical impact being the highest in complex queries. The bars used to represent the review are purposely small in comparison to the resolved bars, i.e. the majority of the queries can be processed by deterministic gates. Meanwhile, the review is not excluded altogether. This helps to have a balanced search process where good evidence will result in an automatic decision and weak evidence will be handled with a lot of care.

It is also indicated by the results that deterministic logic can be used in governance. The trace can be useful to teams which require reviewing of failure, defending search behavior or comparing versions of rules. The decision path is also shareable in common terminology, unlike in opaque scoring where search engineers, catalogue managers and business stakeholders may have to discuss the decision path using different terminology. This system can thus facilitate cycles of improvement without having all the stakeholders learn the covert ranking processes. In general, the findings help to substantiate the argument that deterministic entity resolution is an effective layer in between uncooked query comprehension and retrieval-enhanced product responding. It does not eliminate the value of retrieval but it puts a purer task on retrieval. The framework ensures accuracy without compromising the transparency of the operations by reducing the candidate space using rules based on the schema. It is particularly applicable in product areas where variations in products may influence the customer loyalty, recidivism or confidence in their judgments.

6. DISCUSSION

As demonstrated in the results discussion, deterministic entity resolution is an effective method to provide a practical control layer for commerce search. The main bottom line coming from the quantitative analyses is that linkage of query interpretation to catalogue schema leads to higher quality of candidates. The mixed histogram shows that at the low end of confidence, the number of cases is few (3 in the 50 - 60 band). A bigger group is in the 60-70 band (11 cases), indicating that the use of partial evidence is still prevalent in the actual language of real commerce. Many cases seem to be directed towards a reliable resolution by schema derived gates, the highest of which is in the 80-90 band with 16 cases. This is a useful distribution as it keeps us from being certain, but ready to make decisions. The grouped bar chart reinforces the same explanation, as it demonstrates that there is a significant improvement in each of the four classes of queries. Exact queries go from 74 baseline cases to 93 schema-logic results--meaning it is beneficial to confirm identity fields prior to retrieval. When treating product properties as structured evidence, queries increases from 68 to 89, demonstrating the increased usefulness of product properties. Alias queries have no longer to be normalized on the surface language, but can be normalized maintaining the protected field, with an increase from 65 to 88. Queries with variant-blending will go up from 61 to 86 – queries of this type are the most challenging to search. The review counts of 2, 5, 4 and 7 indicate a review system which doesn't mask out uncertainty; it routes tricky cases into an obvious process.

Table 1 gives an explanation as to why these gains occur. There are 39 strong titles and 45 final resolutions that are variants of the title. Brand variants have a strict brand match 34 percent of the time and final brand resolution 38 percent of the time, so brand evidence is a big, yet not the only, consideration. The fact that pack variants record 28 strict matches and 21 soft variants shows how difficult it is to design a pack/unit difference and that this difference must be carefully designed. A total of 42 final resolutions are obtained from a total of 31 strict matches, indicating that the evidence of models can be obtained through the integration of title and schema information. The schema-gate values of 48, 41, 36 and 39 show that there is discipline that goes beyond text comparison with the schemas. The analysis is moved towards retrieval behaviors in Table 2. The exact-name queries yield 31 sets of returned results, 29 of which contain the correct results, and 2 noisy sets are removed. This is as expected, as direct wording generally can serve as a reliable entity recognition. For attribute-mix queries, 38 sets are returned, 33 of which are correct, 4 are noisy and one is

escalated. With field names (schemas) that do confirm identity, 34 sets are returned and 30 are correct, demonstrating the ability to deal with alias-form queries. Variant-blend queries result in 40 sets and 35 hits (correct) and 2 review cases. The trend is consistent with the fact that more general queries will draw more candidates, with more complex queries having more candidates, and more deterministic filtering needed as queries get more complex.

The tables and graph(s) convey a consistent message. The framework does not merely increase and increase counts for each of the categories but it reorganises the search process. It allows for candidates where there are no contradictions in protected fields, filters candidates when there are contradictions in the protected fields and escalates candidates where the evidence is insufficient. This behavior is well suited to commerce operations as it can have an impact on trust, return and user satisfaction when it comes to the selection of the wrong variant. With the Decision Trace, teams can also keep track of the weak points of the teams. For instance, if there is an increase in pack-variant uncertainty, it may indicate that the labels are missing on the pack and if there is an increase in alias related review may indicate inconsistent brand naming. Another benefit of the proposed approach is that it does not have a negative impact on the search pipeline, while still allowing for governance.

7. CONCLUSION

The present study has created a deterministic entity-resolution approach to tackle this issue of variant disambiguation and precision queries in retrieval-augmented commerce search. With respect to 171 schema-created commerce-search cases, it was demonstrated that schema derived decision logic can boost the candidate control before retrieval. The mixed histogram showed that the highest number of cases of resolution confidence (16) falls in the 80 to 90 band with only 3 cases in the lowest 50 to 60 band. The grouped bar chart revealed that the number of queries of each query class improved consistently, exact queries went from 74 to 93, attribute queries went from 68 to 89, alias queries went from 65 to 88 and variant-blend queries went from 61 to 86. The tables substantiated these results by revealing the influence of the different types of evidence (title, brand, pack and model) on the final resolution and how noisy candidates were filtered before they formed their answer. The framework is useful in that it isn't solely based on hidden scores. Establishes a decision path for acceptance, rejection and review that can be seen. This makes it an appropriate approach for commerce settings where variant mistakes can negatively impact on the level of confidence and raise the burden of operation. The conclusion of the study is that deterministic logic can't be used to replace retrieval. It's a preparation layer which will provide the accuracy of retrieval, reliability of the answers made by the products, and search governance transparency. Future studies can be scaled up to other catalogues, other types of products and multilingual commerce queries. There are 171 controlled instances in the current data set, which is helpful for detailed testing, however, a more extensive test can bring to light more variant conflicts: Regional names, unit conversion, seasonal bundles and seller-specific abbreviations.

8. REFERENCES

[1] S. S. Aanen, D. Vandic, and F. Frasincar, “Automated product taxonomy mapping in an e-commerce environment,” Expert Systems with Applications, vol. 42, no. 3, pp. 1298–1313, 2015. https://doi.org/10.1016/j.eswa.2014.09.032

[2] B. Gu, Z. Li, X. Zhang, A. Liu, G. Liu, K. Zheng, L. Zhao, and X. Zhou, “The interaction between schema matching and record matching in data integration,” IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 1, pp. 186–199, 2017. https://doi.org/10.1109/TKDE.2016.2611577

[3] P. Ristoski, P. Petrovski, P. Mika, and H. Paulheim, “A machine learning approach for product matching and categorization: Use case: Enriching product ads with semantic structured data,” Semantic Web, vol. 9, no. 5, pp. 707–728, 2018. https://doi.org/10.3233/SW-180300

[4] A. Allam, S. Skiadopoulos, and P. Kalnis, “Improved suffix blocking for record linkage and entity resolution,” Data & Knowledge Engineering, vol. 117, pp. 98–113, 2018. https://doi.org/10.1016/j.datak.2018.07.005

[5] G. Simonini, L. Gagliardelli, S. Bergamaschi, and H. V. Jagadish, “Scaling entity resolution: A loosely schema-aware approach,” Information Systems, vol. 83, pp. 145–165, 2019. https://doi.org/10.1016/j.is.2019.03.006

[6] J. Li, Z. Dou, Y. Zhu, X. Zuo, and J.-R. Wen, “Deep cross-platform product matching in e-commerce,” Information Retrieval Journal, vol. 23, no. 2, pp. 136–158, 2020. https://doi.org/10.1007/s10791-019-09360-1

[7] L. Akritidis, A. Fevgas, P. Bozanis, and C. Makris, “A self-verifying clustering approach to unsupervised matching of product titles,” Artificial Intelligence Review, vol. 53, pp. 4777–4820, 2020. https://doi.org/10.1007/s10462-020-09807-8

[8] D. C. Nascimento, C. E. S. Pires, and T. P. Nóbrega, “Configurable assembly of classification rules for enhancing entity resolution results,” Information Processing & Management, vol. 57, no. 3, art. 102224, 2020. https://doi.org/10.1016/j.ipm.2020.102224

[9] G. Papadakis, D. Skoutas, E. Thanos, and T. Palpanas, “Blocking and filtering techniques for entity resolution: A survey,” ACM Computing Surveys, vol. 53, no. 2, art. 31, pp. 1–42, 2020. https://doi.org/10.1145/3377455

[10] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis, “End-to-end entity resolution for big data: A survey,” ACM Computing Surveys, vol. 53, no. 6, pp. 1–42, 2020. https://doi.org/10.1145/3418896

[11] O. S. Albayrak, T. Aytekin, and T. A. Kalaycı, “Duplicate product record detection engine for e-commerce platforms,” Expert Systems with Applications, vol. 193, art. 116420, 2022. https://doi.org/10.1016/j.eswa.2021.116420

[12] C. Ye, S. Jiang, H. Zhang, Y. Wu, J. Shi, H. Wang, and G. Dai, “JointMatcher: Numerically-aware entity matching using pre-trained language models with attention concentration,” Knowledge-Based Systems, vol. 251, art. 109033, 2022. https://doi.org/10.1016/j.knosys.2022.109033

[13] H. Ding, C. Dai, Y. Wu, W. Ma, and H. Zhou, “SETEM: Self-ensemble training with pre-trained language models for entity matching,” Knowledge-Based Systems, vol. 293, art. 111708, 2024. https://doi.org/10.1016/j.knosys.2024.111708.