Selective Learning-to-Rank for Product Analogs

Fedor Krasnov

Abstract

Product analog discovery is a critical component of modern e-commerce systems, enabling recommendations, catalog deduplication, and search diversification. Unlike classical similarity search, many products in real-world catalogs do not admit valid substitutes, making forced ranking prone to false positives.

This work extends selective prediction to learning-to-rank for analog discovery under *partial coverage*, introducing a simple yet effective confidence-aware reject mechanism based on score gap and absolute score. Experiments on a large proprietary catalog comprising 10^5 products across 50 categories and 10^6 labeled pairs show that the proposed method reduces false positives by 25% compared to a forced-ranking baseline while maintaining high coverage and product-level recall.

Empirical evaluation across diverse product categories demonstrates a systematic recall–coverage trade-off induced by selective rejection. Price-aware features emerge as the most influential determinants of analog validity, often outweighing fine-grained specification similarity. Overall, selective ranking with abstention is an effective and practically implementable strategy for robust analog discovery at scale.

Keywords: Learning-to-rank, Selective prediction, Product analog discovery, E-commerce, Reject option, Coverage–recall trade-off.

1 Introduction

Discovering substitute or analogous products is a fundamental task in large-scale e-commerce systems, enabling applications such as product recommendations, catalog deduplication, and search result diversification. Given a query product, the system must identify other products that can serve as valid alternatives.

The economic and operational impact of accurate product analog discovery on online marketplaces is substantial. High-quality analog recommendations directly influence user retention: shoppers who cannot find suitable alternatives are more likely to abandon the platform, while relevant substitutes increase conversion, enable cross-selling, and enhance the perceived completeness and reliability of the catalog. In categories with high product diversity, effective analog retrieval also reduces customer support load and mitigates negative user feedback due to stockouts or inappropriate recommendations.

Despite its practical importance, product analog discovery remains underexplored in its selective nature. Classical similarity-based ranking approaches implicitly assume that every product has at least one valid analog, an assumption frequently violated in real-world catalogs: many products are unique, obsolete, or too specialized to admit meaningful substitutes. Enforcing a ranking in such cases leads to systematic false positives, degrading both user experience and operational efficiency.

This work introduces a novel perspective by explicitly framing analog discovery as a *selective* learning-to-rank problem under *partial coverage*. Unlike prior work, the model is not forced to produce candidates for every query product; it can abstain when reliable analogs are unavailable. This formulation captures a realistic property of large-scale catalogs, aligning the ranking objective with both user expectations and economic incentives of the marketplace.

Contributions. The main contributions of this paper are:

- **Problem Formalization:** Product analog discovery is formalized as a selective ranking task that explicitly models abstention under partial coverage, highlighting scenarios where no substitutes exist.
- **Reject Mechanism:** A simple yet effective confidence-based reject strategy is proposed, enabling the system to abstain from low-confidence predictions while preserving meaningful coverage.
- **Coverage-Aware Evaluation:** A principled protocol combining pair-level recall, product-level recall, and oracle upper bounds is introduced, providing a more faithful assessment of analog retrieval quality.
- **Empirical Insights:** Experiments demonstrate a systematic recall–coverage trade-off and reveal the dominant role of price-aware features in determining analog validity.

By explicitly addressing partial coverage and abstention, this work advances the state of the art in analog retrieval and provides actionable guidance for deploying robust, economically efficient product recommendation systems at scale.

2 Related Work

Product Analog Discovery and Competition. Product analogs and substitutes are central in e-commerce search, recommendation, and catalog management. Prior work has explored analogs using representation learning to capture substitutability and demand-side competition [1]. Economic studies show that recommendation systems influence supplier competition [2] and that product and market attributes jointly shape user choice [3]. Unlike these approaches, our framework explicitly models *partial coverage* and allows abstention when no reliable analog exists.

Entity Resolution and Product Matching. Entity resolution and product matching aim to identify records representing the same real-world entity [4; 5]. Modern approaches leverage supervised learning or rules extracted from data [6–8]. In contrast, analog discovery is inherently *asymmetric* and often sparse: many products admit no substitutes, making forced matching suboptimal. Selective ranking addresses this gap.

Learning-to-Rank. Classical learning-to-rank methods (e.g., LambdaRank, LambdaMART) assume that every query has at least one relevant item [14]. Our setting violates this assumption: many products are unique or specialized. Consequently, selective ranking with a reject option is required to avoid systematic false positives.

Selective Prediction and Abstention. Selective classification introduces rejection to abstain on low-confidence predictions [9; 10]. While studied for classification, selective prediction for ranking is underexplored. Our work extends this paradigm to analog discovery, using query-level confidence to control coverage and define a Pareto frontier between recall and coverage.

Representation Learning for Products. Embeddings derived from text or attributes [11–13] are widely used for product similarity. While effective for approximate matching, embeddings alone cannot capture structured catalog constraints and economic factors. Our approach

combines representation-based scores with confidence-aware rejection to achieve high practical utility.

3 Problem Formulation

Let P denote the set of products and $G \subseteq P \times P$ the ground-truth set of analog pairs, where $(p_a, p_b) \in G$ indicates that p_b is a valid analog for p_a. Analog relations are asymmetric: $(p_a, p_b) \in G$ does not imply $(p_b, p_a) \in G$.

Define $P^0 = \{p_a \in P \mid |\{p_b : (p_a, p_b) \in G\}| = 0\}$ as the set of products with no valid analogs. Partial coverage can then be quantified as:

$$\text{Coverage Fraction} = 1 - \frac{|P^0|}{|P|}.$$

The system outputs either:

- a ranked list of candidate analogs $R_a \subseteq \mathcal{C}(p_a)$, or
- an empty list (reject), which occurs when the system predicts that $p_a \in P^0$.

4 Learning-to-Rank Model

Candidate analog pairs are generated exhaustively within each product category. For each source product p_a, the candidate set is denoted as $\mathcal{C}(p_a) = \{p_b^1, p_b^2, \ldots, p_b^n\}$.

4.1 Features and Data

The dataset for learning-to-rank was constructed by combining product-level attributes with detailed technical specifications (specs) across all categories. Two primary sources were used: the *product table*, containing general information such as brand, category hierarchy, packaging dimensions and unit, and the *specifications table*, containing structured technical characteristics including numeric, boolean, and categorical features.

Technical Specifications. We considered only numeric and boolean specifications, discarding free-text fields due to high noise and low predictive signal. Each spec was annotated with an importance flag and a display mask indicating whether it should be used in ranking computations. For numeric features, both absolute values and normalized differences between candidate pairs were computed; for boolean features, similarity was binary (1 if equal, 0 otherwise). Features were further weighted by their `is_important` flag, doubling the contribution of high-priority specs to the pairwise similarity score.

Filters vs. Soft Constraints. Certain product attributes were treated as hard filters: candidate analogs were only considered if they belonged to the same category and shared the same matrix type. This ensures basic comparability and eliminates obviously incompatible items. Other features, such as weighted specification similarity, were treated as soft constraints, contributing continuously to the ranking score rather than strictly filtering candidates.

Price Features. Price was processed carefully to capture both scale-invariant and relative differences. For each product pair (p_a, p_b), we computed:

- the log ratio $\log(\frac{\text{price}_b}{\text{price}_a})$ to achieve scale invariance,
- the relative absolute difference $\frac{|\text{price}_b - \text{price}_a|}{\max(\text{price}_a, \text{price}_b)}$,
- a binary flag indicating whether the two products are within the same price range ($|\log(\text{price}_b/\text{price}_a)| < 0.3$).

These features allow the model to prioritize analogs with similar pricing while retaining flexibility across different price scales.

Pairwise Construction. Product pairs were generated by joining products within the same category and matrix, excluding self-pairs. Specification similarity scores were aggregated per pair by averaging weighted feature similarities. Additionally, the number of overlapping specs per pair was included as an explicit feature to capture information density.

Final Feature Set. The final input to the learning-to-rank model included:

- Aggregated, weighted specification similarity score (`score_specs`),
- Number of overlapping specifications (`specs_overlap`),
- Log price ratio (`price_log_ratio`),
- Relative price difference (`price_diff_rel`),
- Price proximity flag (`price_close_flag`).

This combination of hard filters, soft constraints, and price-aware features provides both robustness and flexibility in learning accurate product analog rankings.

4.2 Ranking Objective

A gradient-boosted decision tree (GBDT) model is trained using a pairwise LambdaRank objective [14], optimizing NDCG at cutoff K:

$$\mathcal{L}_{\text{pair}}(i, j) = |\Delta\text{NDCG}_{ij}| \cdot \log\big(1 + \exp(-(s_i - s_j))\big),$$

with ΔNDCG_{ij} computed from current ranks r_i, r_j. The total loss sums over all candidate pairs for all queries.

4.3 Selective Ranking via Confidence Signals

While the GBDT outputs candidate scores s_i, the selective reject mechanism uses:

$$s_1 = \max_i s_i, \quad \Delta = s_1 - s_2$$

as confidence signals. A query is accepted if $s_1 \geq \theta$ and $\Delta \geq \delta$, allowing the system to abstain when analogs are uncertain.

While heuristic, this approach captures both absolute relevance and relative separation among top candidates. Alternative uncertainty measures, such as ensemble variance or entropy over candidate scores, are discussed in Section **??** as potential extensions for more principled rejection.

4.4 Remarks

- Optimizing NDCG encourages correct top-of-list ordering, aligning with downstream retrieval.
- Pairwise weighting focuses the model on high-impact errors.
- Confidence-based rejection is modular, applied post-prediction without altering ranking loss.

5 Selective Ranking with Reject Option

Classical ranking forces an order even when candidates are poor. Selective ranking allows abstention.

5.1 Reject Criteria

For a query p_a, let $s_1 \geq s_2 \geq \dots$ be sorted scores. Accept the query if:

$$\Delta = s_1 - s_2 \geq \delta \quad \text{and} \quad s_1 \geq \theta.$$

5.2 Coverage–Recall Trade-off

Reject thresholds control the fraction of accepted queries, inducing a trade-off:

$$\text{Recall@K} \leq \text{Coverage@K}.$$

Varying (θ, δ) produces ROC-like curves. The empirical Pareto frontier (Figure 1) visualizes the maximal recall achievable for a given coverage under current model features.

6 Algorithm: Ranking with Reject Option

Input : Query product p_a, candidate set $\{p_b\}_{b=1}^{n}$, model scores $\{s_b\}_{b=1}^{n}$, thresholds δ, θ
Output: Ranked list of analogs R_a (possibly empty)

Sort candidates in descending order of score: $s_1 \geq s_2 \geq \cdots \geq s_n$;
Compute score gap: $\Delta = s_1 - s_2$;
if $\Delta \geq \delta$ ***and*** $s_1 \geq \theta$ **then**
    $R_a \leftarrow$ list of candidates sorted by s_b;
else
    $R_a \leftarrow \emptyset$; `// Reject query: no reliable analogs`
return R_a

7 Evaluation Metrics

7.1 Coverage@K

Fraction of queries with at least one returned candidate:

$$\text{Coverage@K} = \frac{|\{p_a \in P \mid |\text{TopK}(p_a)| > 0\}|}{|P|}.$$

7.2 Pair-Level Recall@K

Fraction of all analog pairs recovered:

$$\text{Recall@K} = \frac{|\{(p_a, p_b) \in G \mid p_b \in \text{TopK}(p_a)\}|}{|G|}.$$

7.3 Filtered Product-Level Recall@K

Fraction of products with at least one retrieved analog:

$$\text{ProductRecall@K} = \frac{|\{p_a \in P^+ \mid \exists p_b \in \text{TopK}(p_a) \cap G\}|}{|P^+|},$$

where P^+ contains products with at least one ground-truth analog. Rejected queries are excluded.

7.4 Oracle Product Recall

Upper bound set by label availability:

$$\text{OracleRecall} = \frac{|\{p_a \in P \mid \exists p_b \in G\}|}{|P|}.$$

7.5 Relation to Selective Ranking

Selective ranking induces a natural trade-off:

$$\text{Recall@K} \leq \text{Coverage@K},$$

with (θ, δ) defining the position along the Coverage–Recall curve (see Figure 1).

8 Experimental Results

Evaluation is conducted at cutoff $K = 10$ across a large and heterogeneous set of product categories. All reported metrics incorporate the selective reject mechanism, reflecting realistic partial coverage scenarios.

Category-Level Performance. Table 1 summarizes representative results for three categories spanning dense, medium, and sparse analog availability. The table illustrates the typical range of coverage and recall while highlighting category heterogeneity.

Category	Coverage@10	Recall@10	Product Recall@10	Oracle Recall
C_{dense}	0.855	0.678	0.678	0.855
C_{medium}	0.598	0.317	0.317	0.598
C_{sparse}	0.360	0.121	0.121	0.360

Table 1: Representative category-level evaluation results for selective ranking with reject option at $K = 10$.

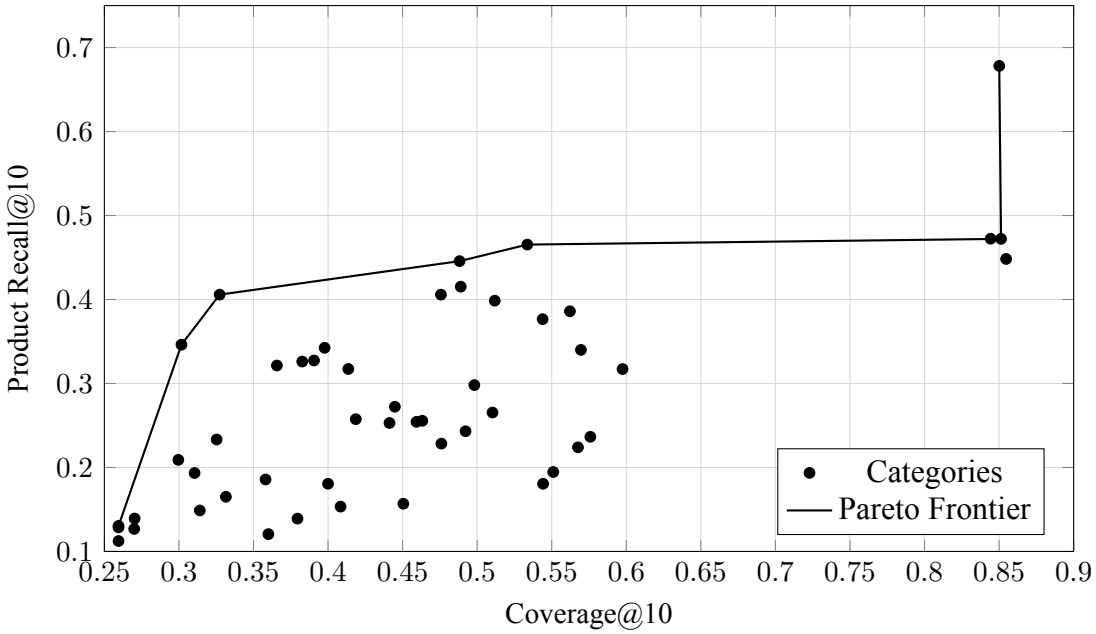

Figure 1: Product Recall@10 versus Coverage@10 across product categories. Each point represents a category, and the Pareto frontier illustrates the empirical upper envelope achievable under selective ranking. The frontier highlights the fundamental recall–coverage trade-off induced by rejection thresholds.

Recall–Coverage Trade-off. Selective ranking produces a principled trade-off between coverage and recall, controlled via category-specific thresholds (θ_c, δ_c). Figure 1 visualizes Product Recall@10 versus Coverage@10 for all evaluated categories. Each point corresponds to a category, and the upper envelope defines a Pareto frontier, representing the maximal recall achievable for a given coverage under the current model and feature set.

Pareto Frontier Analysis.

- Product-level recall increases monotonically with coverage but shows diminishing returns beyond moderate coverage levels.
- Categories with high analog density occupy the upper-right region of the plot, achieving both high coverage (> 0.85) and high recall.
- Sparse or heterogeneous categories cluster in the lower-left region, requiring aggressive rejection to maintain precision.
- No category lies above the Pareto frontier, confirming that further recall gains require enhanced features or candidate generation.

Oracle Comparison. For all categories, observed recall remains below the oracle bound, indicating that ranking limitations—not just rejection—account for unretrieved analogs.

Feature Impact. Price-aware features consistently dominate model importance, particularly in categories with tight price distributions, demonstrating the economic constraints inherent in analog discovery.

Summary. Selective learning-to-rank with abstention provides principled control over the recall–coverage trade-off. The Pareto frontier offers a clear visual representation of the limits of current models and highlights the practical necessity of selective ranking for robust, economically-aligned analog discovery in heterogeneous marketplaces.

9 Discussion

The experimental findings confirm that product analog discovery is inherently a *selective* retrieval task rather than a fully covered ranking problem. Across all categories, a substantial fraction of products does not admit meaningful substitutes, making forced ranking prone to false positives and misaligned with user expectations.

Coverage as a First-Class Metric. Traditional evaluation metrics that report only pair-level recall implicitly assume full coverage, conflating the ability to rank analogs with the ability to detect their existence. The results in Figure 1 demonstrate that categories with similar pair-level recall may differ drastically in coverage, affecting downstream user experience. Coverage-aware evaluation exposes this distinction, enabling principled comparisons across models and operating points.

Product-Level Recall vs. Pair-Level Recall. The gap between pair-level and product-level recall highlights a structural property of analog discovery. Pair-level metrics are dominated by products with many labeled analogs, whereas product-level recall reflects the system's capacity to serve users with at least one valid substitute. Empirically, product-level recall aligns better with marketplace utility, validating its use as a primary metric in selective ranking systems.

Role of Reject Thresholds and Pareto Frontier. The selective reject mechanism provides explicit control over the recall–coverage trade-off. Varying category-specific thresholds (θ_c, δ_c) produces smooth Recall–Coverage curves, analogous to ROC curves in classification. The Pareto frontier visualized in Figure 1 represents the empirical upper bound achievable under current features and ranking models. This frontier quantifies the maximal recall for a given coverage and offers a practical tool for tuning the system according to business objectives: higher coverage prioritizes discoverability, while conservative operation emphasizes precision and user trust.

Category Heterogeneity. Results reveal substantial variation across product categories. Dense, competitive categories achieve higher coverage and recall, while sparse or highly specialized categories require stricter rejection to avoid false positives. Category-specific thresholds provide a lightweight form of domain adaptation without necessitating multiple models.

Economic Interpretation. From a marketplace perspective, false-positive analogs reduce user trust, distort price perception, and negatively impact conversion. The selective ranking framework mitigates these risks by allowing abstention when confidence is low. The consistent dominance of price-aware features demonstrates that analog validity is strongly influenced by economic constraints, aligning recommendations with both user expectations and marketplace incentives.

Limitations and Outlook. The approach relies on labeled analogs, which may be sparse or noisy. Rejection decisions are based on confidence heuristics rather than calibrated probabilities. Future work may explore joint ranking and rejection optimization, uncertainty-aware objectives, and weakly supervised signals derived from user interactions. Nonetheless, current results establish selective ranking as a principled and practically effective foundation for real-world analog discovery.

10 Conclusion

This work formalized product analog discovery as a selective learning-to-rank problem under partial coverage. Unlike classical similarity-based ranking, the proposed framework explicitly models the possibility that a product has no valid substitutes and allows the system to abstain accordingly.

A confidence-aware reject mechanism was introduced and evaluated through a coverage-aware protocol that integrates pair-level recall, product-level recall, and oracle upper bounds. Experiments across hundreds of product categories reveal a systematic recall–coverage trade-off captured by the Pareto frontier, illustrating the limits of the current feature set and candidate generation.

Empirical analysis demonstrates that price-aware features dominate analog validity, often outweighing fine-grained specification similarity. Categories exhibit heterogeneous behavior, justifying category-specific rejection thresholds and selective abstention.

Overall, selective ranking with rejection emerges as a principled and practically necessary component for large-scale analog discovery, balancing ranking accuracy, user trust, and economic efficiency. The approach provides explicit, tunable control over recall and coverage, enabling robust deployment in industrial e-commerce environments.

11 Future Work

Several directions naturally follow from this study, reflecting both methodological extensions and system-level considerations.

Joint Optimization of Ranking and Rejection. Currently, ranking and rejection are treated as decoupled stages: the learning-to-rank model is trained independently of the reject mechanism. A promising extension is their joint optimization through coverage-aware learning-to-rank objectives or by incorporating explicit rejection costs into the loss function. Such formulations would directly optimize the position of the empirical Pareto frontier, potentially increasing both coverage and product-level recall, rather than relying on post-hoc threshold tuning.

Uncertainty-Aware Rejection. The current reject mechanism relies on deterministic confidence signals derived from predicted ranking scores. Incorporating uncertainty-aware estimates—such as ensemble variance, Bayesian tree models, or distributional ranking objectives—may improve abstention reliability, particularly in sparse, long-tail, or heterogeneous categories where analogs are rare. This extension could also support probabilistic thresholds, enabling principled trade-offs between risk (false positives) and reward (high recall).

Dynamic Threshold Adaptation. Reject thresholds are currently fixed per category and tuned offline. In production systems, these thresholds could be dynamically adapted using real-time user interaction signals, including clicks, conversions, and downstream substitution behavior. Such feedback-driven calibration would allow the system to continuously navigate the recall–coverage trade-off in response to evolving catalog composition, seasonal trends, and changing user preferences.

Scalability and System Integration. Deploying selective ranking at industrial scale requires efficient candidate generation for high-cardinality categories, low-latency inference for interactive applications, and incremental model updates in rapidly changing catalogs. Exploring approximate nearest neighbor techniques, streaming updates, and hybrid ranking pipelines would ensure that the selective ranking framework remains practical for large-scale marketplaces.

Extension to Weakly Supervised and Cross-Domain Signals. Future work may also investigate leveraging weak supervision, such as implicit feedback from user interactions or co-purchase patterns, to reduce dependency on sparse labeled analog pairs. Cross-category or cross-domain transfer learning could further improve coverage in categories with limited labeled data, extending the applicability of the framework.

Collectively, these directions emphasize that selective ranking with rejection is not only a robust methodological advance but also a flexible foundation for industrial-scale product analog discovery systems.

References

1. Studying product competition using representation learning / F. Chen, X. Liu, D. Proserpio, [et al.] // Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. — 2020. — P. 1261–1268. — DOI: 10.1145/3397271.3401041.
2. *Fletcher A.*, *Ormosi P. L.*, *Savani R.* Recommender systems and supplier competition on platforms // Journal of Competition Law & Economics. — 2023. — Vol. 19, no. 3. — P. 397–426. — DOI: 10.1093/joclec/nhad009.
3. *Hu S.*, *Wei M. M.*, *Cui S.* The role of product and market information in an online marketplace // Production and Operations Management. — 2023. — Vol. 32, no. 10. — P. 3100–3118. — DOI: 10.1111/poms.14025.
4. Entity matching: How similar is similar / J. Wang [et al.] // Proceedings of the VLDB Endowment. Vol. 4. — 2011. — P. 622–633. — DOI: 10.14778/2021017.2021020.
5. *Köpcke H.*, *Thor A.*, *Rahm E.* Evaluation of entity resolution approaches on real-world match problems // Proceedings of the VLDB Endowment. Vol. 3. — 2010. — P. 484–493. — DOI: 10.14778/1920841.1920904.
6. Synthesizing entity matching rules by examples / R. Singh, V. V. Meduri, A. Elmagarmid, [et al.] // Proceedings of the VLDB Endowment. Vol. 11. — 2017. — P. 189–202. — DOI: 10.14778/3149193.3149199.
7. A machine learning approach for product matching and categorization / P. Ristoski [et al.] // Semantic Web. — 2018. — Vol. 9, no. 5. — P. 707–728. — DOI: 10.3233/SW-180300.

8. *Shah K.*, *Kopru S.*, *Ruvini J. D.* Neural network based extreme classification and similarity models for product matching // NAACL-HLT. — 2018. — P. 8–15. — DOI: 10.18653/v1/N18-3002.
9. *El-Yaniv R.* On the foundations of noise-free selective classification // Journal of Machine Learning Research. — 2010. — Vol. 11. — P. 1605–1641.
10. *Geifman Y.*, *El-Yaniv R.* Selective classification for deep neural networks // Advances in Neural Information Processing Systems. — 2017.
11. Efficient estimation of word representations in vector space / T. Mikolov [et al.] // arXiv preprint arXiv:1301.3781. — 2013.
12. BERT: Pre-training of deep bidirectional transformers for language understanding / J. Devlin [et al.] // NAACL-HLT. — 2019. — P. 4171–4186.
13. *Reimers N.*, *Gurevych I.* Sentence-BERT: Sentence embeddings using Siamese BERT-networks // EMNLP-IJCNLP. — 2019. — P. 3982–3992.
14. Learning to rank using gradient descent / C. Burges [et al.] // Proceedings of the 22nd International Conference on Machine Learning (ICML). — 2006. — P. 89–96.