# Query Classification with Multi-objective Backoff Optimization

Hang Yu
hang.yu@trademe.co.nz
Trade Me Ltd.
Wellington, New Zealand

Lester Litchfield*
lester.litchfield@volparasolutions.com
Volpara Solutions Ltd.
Wellington, New Zealand

## ABSTRACT

E-commerce platforms greatly benefit from high-quality search that retrieves relevant search results in response to search terms. For the sake of search relevance, Query Classification (QC) has been widely adopted to make search engines robust against low text quality and complex category hierarchy. Generally, QC solutions categorize search queries and direct users to the suggested categories whereby the search results are then retrieved. In this way, the search scope is contextually constrained to increase search relevance. However, such operations might risk deteriorating e-commerce metrics when irrelevant categories are suggested. Thus, QC solutions are expected to demonstrate high accuracy. Unfortunately, existing QC methods mainly focus on the intrinsic performance of classifiers whereas fail to consider post-inference optimization that could further improve reliability. To fill up the research gap, we propose the Query Classification with Multi-objective Backoff (QCMB). The proposed solution consists of two steps: 1) hierarchical text classification that classifies search queries into multi-level categories; and 2) multi-objective backoff that substitutes potentially misclassified leaf categories with appropriate ancestors that optimize the trade-off between accuracy and depth. The proposed QCMB is evaluated using the real-world search data of Trade Me that is the largest e-commerce platform in New Zealand. Compared with the benchmarks, QCMB delivers superior solutions with flexible tuning to satisfy different users' demands. To the best of our knowledge, this work is the first attempt to enhance QC with multi-objective optimization.

## CCS CONCEPTS

• **Information systems → Clustering and classification**.

## KEYWORDS

Query classification; multi-objective optimization; e-com search

**ACM Reference Format:**
Hang Yu and Lester Litchfield. 2020. Query Classification with Multi-objective Backoff Optimization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401320

## 1 INTRODUCTION

In recent years, Query Classification (QC) that classifies queries into pre-defined categories has become a promising technique to enhance e-commerce search engines. For a search query, QC increases search relevance by limiting the search scope to the suggested leaf categories wherein related products are retrieved and rendered.

Compared with normal text classification, e-commerce QC is challenged by the following factors: 1) search queries are usually short and ambiguous; and 2) the taxonomy to be classified could be hierarchical with an monotonically increased quantity but decreased discriminability for lower-level nodes close to the bottom. To address those challenges, in recent years, many solutions have been proposed to build high-quality QC from different perspectives including data augmentation and modeling that occur prior to inference [1, 2, 9–15, 18]. However, few existing solutions have considered applying post-inference optimizations to further mitigate misclassified leaf categories that might extremely deteriorate user experience and hinder the real-world applicability of QC.

To this end, we propose the Query Classification with Multi-objective Backoff Optimization (QCMB). The proposed QCMB includes two steps that are hierarchical QC and multi-objective Backoff that activate prior to and after inference, respectively. In detail, firstly, a multi-label classifier is trained in a supervised manner using implicit user feedback [2, 9, 11] to generate multi-level inferences. Next, potentially misclassified leaves are identified and substituted with appropriate ancestor categories, which are usually more discriminative due to the semantic aggregation at higher levels, using the backoff thresholds that optimize both accuracy and depth. From the evaluation using real-world search data, our approach successfully improves the accuracy and usability for real-world deployment of QC solutions.

The rest of the paper is organized as follows. First, the existing solutions for QC are reviewed and summarized in Section 2. Next, Section 3 presents the details of the proposed QCMB followed by the evaluation in Section 4. Finally, conclusions are in Section 5.

## 2 BACKGROUND AND RELATED WORK

Query classification aims to classify search queries into the predefined taxonomy. To overcome the challenges of QC, corresponding solutions usually include data augmentation and text classification.

Due to the insufficiency of training data, many early solutions leverage information obtained from external search engines to either generate or enrich training data. Typical examples are $Q^2C@UST$ [13] and the feature-free QC [10] that utilize external search engines. Based on $Q^2C@UST$, the retraining-free taxonomy-Bridging algorithm [14] and the hybrid model [12] with adaptive query switch are proposed.

[*] Work was done while the author was at Trade Me Ltd.

*SIGIR '20, July 25–30, 2020, Virtual Event, China*

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401320

Recent solutions utilize either explicit or implicit training data. Beitzel, S. M. et al. [1] attempt various methods using human-labeled data. Besides the queries, H. Cao et al. [2] adopt contextual features to build an context-aware classifier. Likewise, D.T. Le and R. Bernardi [9] leverage textual features of the clicked item to enrich the query and train the SVM classifier assisted by topic modeling. Y. Lin et al. [11] propose to construct session-based data using implicit feedback. Lately, transfer learning is applied to leverage product titles and corresponding category labels as the source data [15].

Unfortunately, most of the proposed approaches fail to consider using multi-objective backoff to enhance QC with complex taxonomy. Although very limited single-objective optimization has been done for backoff [1, 5], they lack tuning flexibility and their greedily ranked predictions fail to utilize the hierarchy. Thus, we introduce QCMB that is discussed in the following section.

## 3 PROPOSED SOLUTION

In this section, we present the key components of QCMB including the text classification and multi-objective backoff optimization.

### 3.1 Overview

The proposed QCMB is composed of two components that are the multi-level text classification and multi-objective backoff optimization. First, the multi-level text classifier is trained to predict the category for each level of the category tree for a given query. Next, the backoff thresholds for each level are determined by jointly optimizing accuracy and depth. During inference, multi-level predictions with their scores are firstly generated by the text classifier. Afterwards, the prediction scores are compared against the optimized thresholds from bottom to top and the deepest category with a prediction score above its threshold is selected as the final prediction.

It is worth noticing QCMB is not restrained to specific solutions for each individual component; thus it owns high flexibility to include different classification and optimization techniques. Compared with taxonomy pruning [11], training loss enhancement [17], and single-objective optimization [1, 5], QCMB addresses hierarchical accuracy-depth tradeoff with the following advantages: 1) complete taxonomy; 2) flexible combinations of QC classifiers and optimization algorithms; 3) flexible tradeoff tuning enabled by the diverse candidate solutions of multi-objective optimization. Next, the details of the two major components of QCMB are presented.

### 3.2 Multi-level Hierarchical Classifier

The first component of QCMB is the multi-level hierarchical classifier $f_\Theta$ predicting the category for each level $l$ given a query $q$:

$$f_\Theta(q) = \{c_l^*, p_{c_l^*} | c_l^* \in C, 1 \le l \le N\}, \tag{1}$$

where $c_l^*$ is the predicted category with the highest prediction score $p_{c_l^*}$ among all the predictions for level $l$, $C$ and $N$ denote all the nodes and the height of the category tree, respectively.

Potential candidate models for this task could be either a multi-label classifier taking $C$ as the target or $N$ single-label classifiers generating inferences for individual levels [16]. In this work, the multi-label FastText [8] is selected as the hierarchical classifier for QCMB considering the overall model size and inference time. Briefly, FastText represents queries with averaged word embeddings and applies the sigmoid function with binary cross entropy to each of the target class for muli-label classification. It's efficiency and accuracy for text classification have been shown in many research works [16, 19]. To train the multi-level classifier, the original label, which is the leaf category for a given query, is enriched by including all its ancestors retrieved from the category tree.

### 3.3 Multi-objective Backoff Optimization

Following the multi-level query classifier, the second component of QCMB is the multi-objective optimization that provides backoff thresholds optimizing both accuracy and depth simultaneously.

Given the multi-level classification $f_\Theta(q)$, the backoff process of QCMB selects the deepest category whose prediction score exceeds its corresponding threshold to be the final prediction $\hat{y}_q$. In this work, a unique backoff threshold $t_{vl}$ is defined for each business vertical $v$ (e.g. Marketplace, Motors, Property, etc.) and category level $l$ that can be retrieved from the predicted category. Mathematically, $\hat{y}_q$ is given by

$$\hat{y}_q = \{c_l^* | p_{c_l^*} > t_{vl},\ l = \hat{l}_q,\ \forall c_l^*, p_{c_l^*} \in f_\Theta(q)\}, \tag{2}$$

where $\hat{l}_q$ denotes the deepest level satisfying the threshold conditions, which is defined as

$$\hat{l}_q = \max\{l | p_{c_l^*} > t_{vl},\ \forall c_l^*,\ p_{c_l^*} \in f_\Theta(q)\}. \tag{3}$$

With the final prediction after backoff, in Eqn. (4), we define hit ratio $H$ and depth $D$ as the objectives to be optimized.

$$H = \frac{\sum_q I(|\hat{y}_q \cap Y_q|)}{\sum_q |\hat{y}_q|},\ D = \frac{\sum_q \hat{l}_q}{\sum_q |\hat{l}_q|}, \tag{4}$$

where $Y_q$ denotes the true category and all its ancestors, $I(x)$ is an indicator function such that $I(x) = 1,\ \forall x > 0$ and $I(x) = 0,\ \forall x \le 0$. Briefly, hit ratio is a measure of accuracy that a prediction hits the path between the root and the true leaf category. Those two objectives are chosen considering both the effectiveness and interpretability for the accuracy-depth tradeoff.

Overall, the multi-objective optimization problem, which aims to find $t_{vl}$ to maximize $H$ and $D$, is formulated as follows:

$$\begin{aligned} \max_{t_{vl}} \quad & (H, D) \\ \text{s.t.} \quad & 0 \le t_{vl} \le 1. \end{aligned} \tag{5}$$

It is worth noting that the values of $t_{vl}$ are bounded between 0 and 1 due to the sigmoid functions adopted in the FastText classifier. In this work, the widely-used genetic algorithms, namely NSGA-II [4] and NSGA-III [3], are adopted to solve the multi-objective optimization problem. The basic idea of NSGA-II and NSGA-III is to find the Pareto front, which contains the Pareto optimal solutions that dominate others, via genetic operations such as crossover, mutation, and selection over generations. Compared with NSGA-II, NSGA-III additionally adopts reference points to achieve a uniform distribution of solutions across the objective space. In our case, each candidate solution is encoded as a chromosome composed of 20 $t_{vl}$ representing thresholds for five verticals and four levels. For clarity, QCMB utilizing NSGA-II and NSGA-III are collectively called QCMB-II and QCMB-III, respectively.

# 4 EVALUATION

The proposed QCMB is measured using off-line evaluation with Trade Me search data. This section presents the datasets, parameters, metrics, benchmarks, and results of evaluation.

## 4.1 Data Preparation

In this work, the widely-used implicit user clicks [2, 9, 11, 12] are adopted to create the classification dataset. Specifically, 94,472,105 (query, category) pairs are collected from the one-month clicked searches occurred on Trade Me. The categories in the data set are the leaf categories of a four-level tree-based taxonomy that comprises 5484 leaves out of 6367 nodes as described in Table 1. The full data is then randomly partitioned into a training set, a validation set, and a test set with a 90:5:5 split, respectively. The training set is for training the text classifier whereas the validation set is utilized to optimize the backoff thresholds.

**Table 1: Statistics of Trade Me Taxonomy**

| Stats | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|
| # of all categories | 27 | 387 | 2466 | 3487 |
| # of leaf categories | 1 | 80 | 1916 | 3487 |

## 4.2 Parameters

Mostly, the default parameters for FastText, NSGA-II, and NSGA-III in their implementations are adopted. For the FastText model, the original Facebook implementation is used. The learning rate is set to 0.05 and training is done with 5 epochs. The dimension of word embeddings is 300 with the maximum length of word n-grams set to 2. For multi-objective optimization, the Platypus [6] implementations of NSGA-II and NSGA-III algorithms are adopted. Specifically, the population size and number of generations are both set to 100. The number of outer divisions for NSGA-III is set to 99 to obtain sufficient solutions for comparison.

## 4.3 Metrics

The two major metrics adopted for evaluation are the hit ratio and depth defined in Eqn. (4). Besides, we also report hierarchical precision $hP$, recall $hR$, and F1 score $hF_1$ that are specifically designed to measure the classifications for hierarchical taxonomy [16]. Compared with hit ratio and depth, false predictions for sibling leaves are less penalized by the hierarchical metrics, which is demonstrated in the definitions below:

$$hP = \frac{\sum_q |\hat{Y}_q \cap Y_q|}{\sum_q |\hat{Y}_q|},\; hR = \frac{\sum_q |\hat{Y}_q \cap Y_q|}{\sum_q |Y_q|},\; hF_1 = \frac{2 \cdot hP \cdot hR}{hP + hR}, \quad (6)$$

where $\hat{Y}_q$ is the predicted category and all its ancestors.

In this paper, we adopt the @1 performance for all the metrics considering the real-world application at Trade Me.

## 4.4 Benchmark

The proposed QCMB is compared against two benchmark solutions that are: 1) the single-class flat classifier, viz. Flat for short, that classifies a query into one leaf category with no backoff; and 2) the Query Classification with Single-objective Backoff (QCSB) that is the multi-level hierarchical classifier with backoff thresholds optimized for the hierarchical F1 score [5]. Similar to QCMB, we have QCSB-II and QCSB-III that represent QCSB adopting NSGA-II and NSGA-III, respectively. For all the benchmarks, FastText with the same parameters used by QCMB is adopted as the base model.

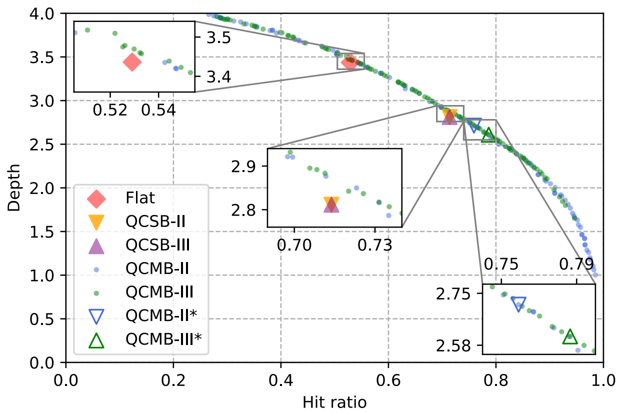

**Figure 1: Evaluation results for hit ratio and depth**

## 4.5 Results and Discussions

First, the results for hit ratio and depth are presented in Fig. 1. It is clear that the solutions provided by QCMB-II and QCMB-III, which are represented by the blue and green dots, form the Pareto fronts that dominate all other solutions given by the flat model and QCSB. In other words, neither the flat model nor QCSB is able to surpass QCMB for both objectives at the same time. Especially, the potential improvement to hit ratio compared with the flat model validates the success of the backoff process. However, due to the task-specific performance variance [7], the solutions generated by NSGA-II in our case demonstrate better diversity than NSGA-III as shown by the solutions with hit ratios above 0.8.

Besides the Pareto optimality, QCMB additionally enables flexible, convenient, and comprehensive tuning between hit ratio and depth after training the query classifier thanks to the high coverage of optimal solutions over the objective space. In other words, users have a high degree of freedom to select the candidate solutions based on their own criteria without retraining the classifier. This feature could significantly increase the applicability of QC models when the performance demand cannot be satisfied by the single-solution approaches. Moreover, multiple optimal solutions could get involved in the online test so as to find the final solution that optimizes the online metric such as Click Through Rate (CTR).

To be precise, the numeric values of all the performance metrics are presented in Table 2. For QCMB, we select the solutions with the highest hierarchical F1 scores, viz. QCMB-II* and QCMB-III*, as representatives. As expected, QCMB-II* and QCMB-III* successfully increase the hit ratio so that false predictions could be avoided to protect search experience. The F1 scores of QCMB are close but inferior to those of QCSB, however, we argue that the hierarchical metrics might fail to reflect our focus as they tolerate the sibling leaves that should be avoided.

**Table 2: Evaluation results**

| | hit ratio | depth | HP@1 | HR@1 | HF1@1 |
|---|---|---|---|---|---|
| Flat | 0.5291 | 3.4363 | 0.7659 | 0.7606 | 0.7633 |
| QCSB-II/III | 0.7138 | 2.8113 | 0.8661 | 0.7389 | 0.7975 |
| QCMB-II* | 0.7591 | 2.7127 | 0.8717 | 0.7244 | 0.7912 |
| QCMB-III* | 0.7865 | 2.6090 | 0.8852 | 0.7151 | 0.7911 |

Next, the breakdown of backoff hops, which summarizes the percentages of different hop counts of backoff for the correct predictions, is presented in Table 3. The statistics show that mostly the unnecessary backoff is avoided for the sake of depth. In addition, the backoff of QCMB-II* and QCMB-III* are more proactive than that of QCSB-II/III to trade off depth against hit ratio.

**Table 3: Breakdown of backoff hops**

| | 0 hop | 1 hop | 2 hops | 3 hops | 4 hops |
|---|---|---|---|---|---|
| QCSB-II/III | 62.30% | 19.59% | 10.04% | 4.99% | 3.08% |
| QCMB-II* | 58.55% | 15.30% | 9.00% | 9.13% | 8.02% |
| QCMB-III* | 56.09% | 15.34% | 8.03% | 10.8% | 9.74% |

Correspondingly, the accuracy for categories at different levels are shown in Table 4. It can be seen that QCMB-II* and QCMB-III* have increasing accuracy values for categories at higher levels closer to the root whereas QCSB-II/III demonstrates the opposite accuracy distribution. This is because the hierarchical F1 score in Eqn. (6) has a larger depth penalty contributed by both hierarchical precision and recall. As a result, for QCSB, higher levels would have lower thresholds that are conservative to backoff the predictions with low confidence scores.

**Table 4: Breakdown of level-based accuracy**

| | level 1 | level 2 | level 3 | level 4 |
|---|---|---|---|---|
| QCSB-II/III | 0.6010 | 0.6536 | 0.7320 | 0.7083 |
| QCMB-II* | 0.8758 | 0.7555 | 0.7869 | 0.6312 |
| QCMB-III* | 0.9237 | 0.7892 | 0.7535 | 0.6931 |

Considering the above, it can be concluded that QCMB is capable of providing optimal solutions with high tuning flexibility. Moreover, the concept can easily be extended to support other query classifiers and optimization techniques.

## 5 CONCLUSIONS

Query classification plays an important role in improving the relevance of search engines. Most existing solutions focus on data processing and model enhancement that are prior to inference. Thus, in this paper, we propose QCMB that adopts multi-objective post-inference threshold optimization to further increase the applicability of QC models. Through performance evaluation using the real-world search data, it is validated that QCMB could deliver superior solutions with high tuning flexibility compared with the benchmarks.

## ACKNOWLEDGMENTS

This work supported by Trade Me Ltd., New Zealand.

## REFERENCES

[1] Steven M. Beitzel, Eric C. Jensen, Ophir Frieder, David Grossman, David D. Lewis, Abdur Chowdhury, and Aleksandr Kolcz. 2005. Automatic Web Query Classification Using Labeled and Unlabeled Training Data. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05).* ACM, New York, NY, USA, 581–582. https://doi.org/10.1145/1076034.1076138

[2] Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian-Tao Sun, Enhong Chen, and Qiang Yang. 2009. Context-aware Query Classification. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09).* ACM, New York, NY, USA, 3–10. https://doi.org/10.1145/1571941.1571945

[3] K. Deb and H. Jain. 2014. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Transactions on Evolutionary Computation* 18, 4 (Aug 2014), 577–601. https://doi.org/10.1109/TEVC.2013.2281535

[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (April 2002), 182–197. https://doi.org/10.1109/4235.996017

[5] K. Draszawka and J. SzymaÅĎski. 2013. Thresholding strategies for large scale multi-label text classifier. In *2013 6th International Conference on Human System Interactions (HSI).* 350–355. https://doi.org/10.1109/HSI.2013.6577846

[6] David Hadka. 2015. Platypus: A Free and Open Source Python Library for Multiobjective Optimization. https://github.com/Project-Platypus/Platypus.

[7] H. Ishibuchi, R. Imada, Y. Setoguchi, and Y. Nojima. 2016. Performance comparison of NSGA-II and NSGA-III on various many-objective test problems. In *2016 IEEE Congress on Evolutionary Computation (CEC).* 3045–3052. https://doi.org/10.1109/CEC.2016.7744174

[8] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.* Association for Computational Linguistics, 427–431.

[9] Dieu-Thu Le and Raffaella Bernardi. 2012. Query Classification Using Topic Models and Support Vector Machine. In *Proceedings of ACL 2012 Student Research Workshop (ACL '12).* Association for Computational Linguistics, Stroudsburg, PA, USA, 19–24. http://dl.acm.org/citation.cfm?id=2390331.2390335

[10] Lin Li, Luo Zhong, Guandong Xu, and Masaru Kitsuregawa. 2012. A feature-free search query classification approach using semantic distance. *Expert Systems with Applications* 39, 12 (2012), 10739 – 10748. https://doi.org/10.1016/j.eswa.2012.02.191

[11] Y. Lin, A. Datta, and G. D. Fabbrizio. 2018. E-commerce Product Query Classification Using Implicit UserâĂŹs Feedback from Clicks. In *2018 IEEE International Conference on Big Data (Big Data).* 1955–1959. https://doi.org/10.1109/BigData.2018.8622008

[12] Dou Shen, Ying Li, Xiao Li, and Dengyong Zhou. 2009. Product Query Classification. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09).* ACM, New York, NY, USA, 741–750. https://doi.org/10.1145/1645953.1646047

[13] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. 2005. Q2C@UST: Our Winning Solution to Query Classification in KDDCUP 2005. *SIGKDD Explor. Newsl.* 7, 2 (Dec. 2005), 100–110. https://doi.org/10.1145/1117454.1117467

[14] Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2006. Building Bridges for Web Query Classification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06).* ACM, New York, NY, USA, 131–138. https://doi.org/10.1145/1148170.1148196

[15] Michael Skinner and Surya Kallumadi. 2019. E-commerce Query Classification Using Product Taxonomy Mapping: A Transfer Learning Approach. In *Proceedings of the SIGIR 2019 Workshop on eCommerce.* ACM.

[16] Roger Alan Stein, Patricia A. Jaques, and JoÃČÂčo Francisco Valiati. 2019. An analysis of hierarchical text classification using word embeddings. *Information Sciences* 471 (2019), 216 – 232. https://doi.org/10.1016/j.ins.2018.09.001

[17] Cinna Wu, Mark Tygert, and Yann LeCun. 2017. Hierarchical loss for classification. *arXiv preprint arXiv:1709.01062* (2017).

[18] C. Xia and X. Wang. 2015. Graph-Based Web Query Classification. In *2015 12th Web Information System and Application Conference (WISA).* 241–244. https://doi.org/10.1109/WISA.2015.68

[19] Wenhu Yu, Zhiqiang Sun, Haifeng Liu, Zhipeng Li, and Zhitong Zheng. 2018. Multi-level Deep Learning based e-Commerce Product Categorization.. In *eCOM@SIGIR.*