

Hierarchical Query Classification in E-commerce Search

Bing He
bhe46@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

Suhang Wang
szw494@psu.edu
The Pennsylvania State University
University Park, PA, USA

Zhen Li
amzzhn@amazon.com
Amazon
Palo Alto, CA, USA

Sreyashi Nag
sreyanag@amazon.com
Amazon
Palo Alto, CA, USA

Zheng Li
amzzhe@amazon.com
Amazon
Palo Alto, CA, USA

Haiyang Zhang
hhaiz@amazon.com
Amazon
Palo Alto, CA, USA

Limeng Cui
culimeng@amazon.com
Amazon
Palo Alto, CA, USA

Rahul Goutam
rgoutam@amazon.com
Amazon
Palo Alto, CA, USA

ABSTRACT

E-commerce platforms typically store and structure product information and search data in a hierarchy. Efficiently categorizing user search queries into a similar hierarchical structure is paramount in enhancing user experience on e-commerce platforms as well as news curation and academic research. The significance of this task is amplified when dealing with sensitive query categorization or critical information dissemination, where inaccuracies can lead to considerable negative impacts. The inherent complexity of hierarchical query classification is compounded by two primary challenges: (1) the pronounced class imbalance that skews towards dominant categories, and (2) the inherent brevity and ambiguity of search queries that hinder accurate classification.

To address these challenges, we introduce a novel framework that leverages hierarchical information through (i) enhanced representation learning that utilizes the contrastive loss to discern fine-grained instance relationships within the hierarchy, called “instance hierarchy”, and (ii) a nuanced hierarchical classification loss that attends to the intrinsic label taxonomy, named “label hierarchy”. Additionally, based on our observation that certain unlabeled queries share typographical similarities with labeled queries, we propose a neighborhood-aware sampling technique to intelligently select these unlabeled queries to boost the classification performance. Extensive experiments demonstrate that our proposed method is better than state-of-the-art (SOTA) on the proprietary Amazon dataset, and comparable to SOTA on the public datasets of Web of Science and RCV1-V2. These results underscore the efficacy of our proposed solution, and pave the path toward the next generation of hierarchy-aware query classification systems.

CCS CONCEPTS

• **Computing methodologies** → **Semi-supervised learning settings**; **Natural language processing**; • **Information systems** → **Web mining**.

KEYWORDS

Query Classification, Hierarchical Text Classification, Semi-supervised Learning

ACM Reference Format:

Bing He, Sreyashi Nag, Limeng Cui, Suhang Wang, Zheng Li, Rahul Goutam, Zhen Li, and Haiyang Zhang. 2024. Hierarchical Query Classification in E-commerce Search. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion)*, May 13–17, 2024, Singapore, Singapore. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3589335.3648332>

1 INTRODUCTION

Hierarchical query classification is a vital task in the domain of e-commerce and search, playing a crucial role in driving customer obsession [8]. As users interact with online services, they input various queries to search for products, services, or information. Accurately classifying these queries is pivotal in ensuring that users are presented with the most relevant and valuable results. One significant application of the hierarchical query classifier in industry is categorizing sensitive queries that follow a predefined hierarchy in e-commerce. For example, given a query, it can be classified as harmful, adult-oriented, or non-sensitive products (Here, for illustration, we define these categories as parent categories). Furthermore, for harmful products, there are two child categories: self-harm and harm to others. The child categories for the adult-oriented category can be adult products and adult content. Since these queries contain offensive content or pertain to unregulated goods, and different categories need to be handled differently, mis-classification of such queries can lead to unpleasant or even detrimental user experiences, potentially damaging a service’s reputation and user trust. Moreover, presenting inappropriate or restricted content could lead to legal ramifications for the service. Hence, building an accurate hierarchical query classification framework is of paramount importance,



This work is licensed under a Creative Commons Attribution International 4.0 License.