Diva: Dynamic Range Filter for Var-Length Keys and Queries

Navid Eslami
Department of Computer Science
University of Toronto
navideslami@cs.toronto.edu

Range Filters and Applications. Range filters are compact probabilistic data structures that accept two keys as the end-points of a range and determine if at least one key in a given set is in-between them. These filters are used in many domains, e.g., in key-value stores [4], B-Tree indices [2], and relational databases [1], to quickly rule out the existence of keys in a given query range and avoid having to search for them in storage.

Range Filtering Goals. The ideal general-purpose range filter must simultaneously support (G1) the lowest possible false positive rate under a stringent memory budget, (G2) range queries of any length, (G3) variable-length keys (e.g., strings), (G4) dynamic modification operations such as insertions, deletions, expansions, and contractions, and the best possible (G5) query and (G6) construction performance.

Design Contentions. Yet, every existing range filter only fulfills a subset of these goals [1, 4, 2]. In fact, almost none of the existing ones attain (G3) or (G4). Moreover, Goswami et. al. have proven an information-theoretic lower bound on the memory footprint of range filters, stating that it is impossible for any range filter to achieve (G1) and (G2) at the same time [3]. However, this lower bound only holds for worst-case datasets, meaning that it may still be possible to achieve (G1) and (G2) for "common" datasets. As such, we pose the following research question: Is it possible to design a range filter that simultaneously fulfills all six goals for common datasets? We present an affirmative answer to this question.

Diva. We introduce Diva, the first range filter to support dynamic operations, variable-length queries and keys, and high performance, all at the same time. Diva learns the dataset's distribution by sampling keys and storing them in a cache-efficient trie. For all keys in-between two samples of the trie, Diva removes the longest common prefix. It also truncates their suffixes while keeping enough bits in the middle of each key (i.e., an infix) to differentiate them in most cases, thus achieving (G1). At the same time, the trie separates dense and sparse regions of the key space. This allows for handling short range queries over densely populated regions and long range queries over sparse regions, thus meeting (G2). By discretizing all keys into fixed-length infixes without the use of hashing, Diva achieves (G3). This discretization also allows for storing the infixes of adjacent groups of keys within a constant time data structure called an Infix Store, fulfilling (G5). As Diva derives infixes without hashing and stores them in the original sorted order of the keys, it does a single efficient sequential pass over the dataset during construction, making it the fastest range filter to construct. It therefore attains (G6). Diva handles dynamic updates by splitting Infix Stores, thereby meeting (G4).

We show, theoretically and empirically, that Diva achieves a false positive rate on par with the state of the art on real-world datasets while supporting dynamicity and variable-length queries and keys.

Diva was endowed with the best research paper award at VLDB 2026, held in the United Kingdom, London.

References

[1] Karolina Alexiou, Donald Kossmann, and Per-Åke Larson. Adaptive range filters for cold data: avoiding trips to siberia. *Proc. VLDB Endow.*, 6(14):1714–1725, sep 2013. URL: https://doi.org/10.14778/2556549.2556556.

[2] Navid Eslami and Niv Dayan. Memento filter: A fast, dynamic, and robust range filter. *Proc. ACM Manag. Data*, 2(6), December 2024. URL: https://doi.org/10.1145/3698820.

[3] Mayank Goswami, Allan Grønlund, Kasper Green Larsen, and Rasmus Pagh. Approximate range emptiness in constant time and optimal space. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, page 769–775, USA, 2015. Society for Industrial and Applied Mathematics.

[4] Huanchen Zhang, Hyeontaek Lim, Viktor Leis, David G. Andersen, Michael Kaminsky, Kimberly Keeton, and Andrew Pavlo. Surf: Practical range query filtering with fast succinct tries. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 323–336, New York, NY, USA, 2018. Association for Computing Machinery. URL: https://doi.org/10.1145/3183713.3196931.