A Generic Inverted Index Framework for Similarity Search on the GPU

Björn Bebensee (2019–21343)
Topics in Artificial Intelligence

September 24, 2019

In high throughput applications there is often a need to match similarities of index structures or features. The Zhou et al. propose a framework called GENIE which reduces the programming complexity for this type of parallel similarity search on a GPU. GENIE can be used for similarity search on all data types that can be expressed in the *match-count model* which is defined as follows: given a query $Q = \{(d_i, [v_i^L, v_i^U]) \mid i \in [1, n]\}$ with query length n and an object $O = \{(d_1, v_1), \ldots, (d_l, v_l)\}$, the match-count model $MC(\cdot, \cdot)$ returns $MC(Q, O) = \sum_{r_i \in Q} C(r_i, O)$. Here C simply returns the number of elements $o \in O$ contained by r_i. Informally, MC returns the number of elements that are contained by at least one query item $r_i \in Q$. GENIE then aims to obtain the top-k candidates for the query Q using an inverted index on the GPU.

The inverted index is kept in the GPU's global memory. With modern GPUs having memory sizes of upwards of 12GB or even 24GB DDR5 memory and being able to use multiple GPUs simultaneously the memory size is on par with regulary (RAM) memory size.

GENIE divides the task into smaller task and runs all queries and all their query items in parallel to leverage the parallel computation power of GPUs. More specifically, one block on the GPU handles one query item and each thread corresponds to an object in the posting list.

In order to implement their framework, Zhou et al. propose a novel data structure *Count Priority Queue* (c-PQ) which helps reduce the time cost for similarity search and helps overcome the problem of top-k candidate selection while also reducing memory requirements for multiple queries. The idea behind c-PQ is to use a two-level data structure where only the top candidates are stored in the upper level structure so that during top-k candidate selection only the upper level needs to be searched while all objects in the lower level structure can be discarded.

While the authors admit that not all data types can be expressed within the MC model (and thus in the framework), they show that many can be. They prove that GENIE can support any similarity measure which satisfies the *Locality Sensitive Hashing* (LSH) scheme using their proposed Tolerance Approximate Nearest Neigbour (τ-ANN) search as well as more complex data types that can be transformed using *Shotgun and Assembly* (SA). Finally, Zhou et al. show the effiency of GENIE on multiple real-world datasets (OCR, SIFT, DBLP, Tweets, etc.). They find that their framework outperforms other algorithms by one or more orders of magnitude.

References

[1] Zhou, Jingbo, et al. "A generic inverted index framework for similarity search on the GPU." *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 2018.