Critique of The next 50 Years in Database Indexing or: The Case for Automatically Generated Index Structures

The paper [1] presents a new framework for generating index structures in computer science based on genetic algorithms. The authors argue that the traditional approach of inventing index structures is flawed and propose a new framework that mimics existing index structures and can automatically generate new structures given a workload and optimization goal. The framework introduces a generic logical indexing framework to clearly distinguish between the logical and physical aspects of an index, allowing for more flexible and scalable index design and implementation. The process of specifying the index structure involves deciding which search algorithm to use, specifying the data layout, and deciding whether to use a nested physical index. The authors present extensive experimental evaluation of their approach and conclude with future research directions.

The description of the experiment describes the use of a genetic algorithm to optimize an index structure. The performance of the algorithm is compared to various baseline index structures and the results show that the algorithm was successful in reproducing the performance of the baseline index structures. The largest dataset and workload were used to evaluate the results and it was found that the genetic algorithm produced similar results to the expected baseline for each workload.

In a comparison with other prevalent heuristic index types, the GENE index outperforms the other structures in certain scenarios, such as real-world skewed and sparse datasets, with an average index lookup time of around 350 ns. The physical structure of the GENE index is bulk loaded with hash nodes for the first and third partition, showing its effectiveness in optimizing index structures for different workloads.

However, there are limitations in the text that could be improved. The text does not provide any information on the evaluation metric used to compare the model's performance, which makes it difficult to assess the validity of the conclusions. Additionally, the text mentions that the population size has a limited influence, but does not provide any numerical evidence to support this claim. The text could also benefit from a more detailed explanation of the population insertion criterion, as it is a key aspect of the experiment.

The fitness function in the experiment is described well, although some minor improvements could be made. The description could benefit from providing more context and explanation of the optimization goal and why two runs were chosen. Additionally, the description could be improved by providing context or examples of other optimization goals and elaborating on the mention of incentivizing the filling grade of leaves. Overall, the description of the fitness function is clear and comprehensive, with a few areas for improvement.

References:
[1] Jens Dittrich et al. The next 50 Years in Database Indexing or: The Case for Automatically Generated Index Structures. VLDB 2022