Grenze International Journal of Engineering and Technology, June Issue

Overview of Boolean Search Queries Optimization in Information Retrieval Systems

Pratibha Waghale[1], Unnati Thakre[2], Triveni Kotgule[3], Disha Rajgadkar[4], Aishwarya Rode[5] and Shriti Nandanwar[6]
[1-6]Yeshwantrao Chavan College of Engineering, Computer Science and Engineering, Nagpur, India
Email: pratibha201986@gmail.com, thakreunnati304, kotguletriveni, disharajgadkar, aishwaryarode9, shritinandanwar}@gmail.com

***Abstract*— Generating Boolean Search queries for patent retrieval integrates Machine Learning, Sophisticated computational approaches and informational retrieval techniques. With the goal to improve patent searches and refine queries, this paper determines recent improvements in the Boolean logic, optimization techniques, and AI-driven methods. Innovative approaches like SetBERT demonstrate that fine-tuning the neural models can improve the retrieval performance of Boolean queries like AND, OR, and NOT, which are necessary for patent search. By inversed-contrastive loss, SetBERT gives great recall performance and outperforms traditional BERT-based methods, with applications in patent information retrieval. Similarly, reinforcement learning-based pipelines in systematic literature reviews (SLRs) illustrate that an approach wherein query refinements are suggested based on real-time feedback shows the possibility of being extended to the optimization of Boolean queries for patent datasets. Optimization techniques are very important in patent searching in relation to both scalability and complexity. Thus, Quantum**
Approximate Optimization Algorithm (QAOA) offers scalable solutions to the problem of multi-query optimization for more effectively searching across large-sized patent databases. Advanced indexing methods such as BE-Tree and the graph-based elemental indexes enable efficient processing of high-dimensional Boolean queries, which is critical for exploration of nuanced semantic relationships in patent datasets. This would be very ideal for such applications as patent retrieval, where large data sets can be computationally intensive.

***Index Terms*— Boolean search optimization, patent information retrieval, machine learning in patent search, SetBERT, inversed-contrastive loss.**

I. INTRODUCTION

Relevant documents are fetched by the Information Retrieval Systems in response to the user queries. Sometimes irrelevant documents in the result sets are fetched due to improper structure of the query and the motive behind the retrieval of the particular document is adequately unexplainable. Optimization of the queries is essential for reducing slowdowns, irrelevant results, errors and computational cost. Mixed results were produced when the Natural Language Processing techniques are applied on Information Retrieval Systems. IR system which are great at detecting textual similarity and locating relevant documents, attempts of including linguistically complex techniques have typically produced negligible gains or degraded output.

Grenze ID: 01.GIJET.11.2.282_20

The success of IR systems that treat text as "bag of words" highlights how retrieval approaches [22] achieve a compromise between practical effectiveness and complexity. Genetic programming can optimize the Boolean queries to a good extent. Precision and recall is enhanced by enforcing weights for Boolean operatorsand due to the application of genetic algorithms to refine Boolean queries. Harmonic mean measures are employed to balance precision and recall [17] Innovative approaches have addressed the limitations of the traditional system. Use of AND, NOT, OR which are intersection, Difference and Union respectively prove to be helpful when the user requirements are specific. Traditional retrieval systems and neural models, such as BERT and its variant pose significant hurdles in processing Boolean queries that involve AND, OR and NOT [1]. Lexicon-based methods failed to handle negations effectively. Sometimes "and" and "not" are treated as similar words by the BERT and T5. Enhanced query embedding techniques and optimization strategies, as demonstrated by the SetBERT model prove to be essential to overcome these limitations. Synthesis based techniques like SlabCity prove to be helpful in overcoming the limitations of Rule based rewriting methods such as Apache Calcite and WeTune which rely on predefined patterns. Synthesis based techniques offer whole-query optimization that's why there is no need for predefined rules thus helping program synthesis to generate equivalent but faster queries [5]. In this paper, advanced methodologies for optimizing Boolean queries and improving retrieval performance, evolutionary algorithms, program synthesis and indexing techniques are explored. Insights from neural models are drawn. Limitations of traditional and existing retrieval systems are addressed and highlighted in various studies in this paper. Research from early works on Boolean logic optimization to modern advancements in vector database management is synthesized, efficient and accurate information retrieval solutions are given in this paper.

II. Literature Survey

SetBERT, an enhanced BERT model increasing retrieval for Boolean logic queries (AND, OR, NOT) by using inversed-contrastive loss and a GPT-generated dataset [1]. It is superior to the BERT-base by 63% and achieves BERT-Large's performance with one-third the size. SetBERT, which is tested on the QUEST dataset, is specialized in solving the most involved queries effectively. The paper describes a tool that is engineered by means of fine-tuning of LLMs to generate Boolean search queries for systematic reviews. The model is mainly fine-tuned on PROSPERO and PubMed data also known as Mistral-Instruct-7b which brings the study certain sensitivity but on the other hand lower than optimal precision which makes it unasuitable for a non-audited application [2]. The paper describes iRangeGraph, which is a method of segment tree, a method of index maintenance operation with low overhead that may be used to do the efficient Range-Filter Approximate Nearest Neighbor (RFANN) search, and shows an on-the-fly graph index construction using it. In detail, it shows the maximum, accuracy as well as memory efficiency with the support of multi-attribute RFANN queries [3]. All writing faces problems when it comes to how data is fed into vector databases, such as how to compare unequal or similar entities and derive hybrid queries. It then comes up with solutions like compression, indexing, and query optimization. It assists systems that use vectors to select databases and includes topics for further investigations that would upgrade their viability [4]. The paper introduces SlabCity, a synthesis-based SQL optimizing method that does not follow predefined rewrite rules. Instead, it employs program synthesis and dataflows for the whole query optimization. It outperforms rule-based systems by executing different queries more efficiently and therefore getting better response times [5]. A new paper advocates the employment of gate-model quantum computers that would use the Quantum Approximate Optimization Algorithm (QAOA) for the solution of the Multiple Query Optimization problem that was found to be NP-hard. One of the other reasons aimed at is that of higher qubit usage efficiency and ease of scaling than quantum annealers [6]. The paper describes a reinforcement learning-based framework for developing and modifying search queries in systematic literature reviews (SLRs). It revises queries depending on the feedback received to maximize recall and to reduce workload, producing performance that is almost the same as human-generated queries [7]. Light spatial data indexing method for Hadoop enhance the search for geographical data in a very large system. The completion of spatial and text data criteria is expedited by the joint process. The outcomes point to a successful performance of ranked queries based on the geographical location that are similar to text-based queries, even of this type of case at considerable datasets [8]. In this paper, an effective decentralized Boolean querying scheme for document recovering at edge computing is suggested, with the reduction of data transmission and the query optimization being the main contributions. The published results demonstrate the drastic improvement of the decentralized method on the larger scale of asavahame the traditional one in some cases [9]. In order to enhance artificial neural networks, the paper makes a comparison of optimization techniques such as PSO, GA, and ABC to select the best possible combination of weights and structures [10]. A new method of the wolf pack is tried out using

the Grey Wolf Optimization algorithm and the Boolean operations where the task of the itemset mining of the real High Utility Itemset Mining (HUIM) is solved more quickly and more efficiently. The paper actually performs better than the rest of the produced models when dealing with a fake or real dataset at the level of accuracy and computational efficiency [11]. The study tackles the algorithm selection problem; the main subject is on enhancing combinatorial search problems through machine learning-based mappings of problem instances to algorithms. Of course, it needs to be explicitly expressed that the methods are SATzilla and that the algorithm portfolios play the major role of performance enhancement [12]. The research documents a modified Particle Swarm Optimization (PSO) algorithm and Hill Climbing alliance seeking to discover balanced Boolean functions having good cryptographic properties. It shows competitive numbers in nonlinear optimizations against linear algorithms and their corresponding criteria of security protection [13]. The research documents a modified Particle Swarm Optimization (PSO) algorithm and Hill Climbing alliance seeking to discover balanced Boolean functions having good cryptographic properties. It shows competitive numbers in nonlinear optimizations against linear algorithms and their corresponding criteria of security protection [14]. The paper describes a faster method to the k-nearest neighbor search in 3D point clouds. It adopts a new data structure and cubic grid-based space partitioning to improve the processing speed. This method excels at speed among the other ones whereas working on big data. The experimental results are consistent to the fact that millions of points can be worked with [15]. The journal suggests optimization of the boolean queries in the information retrieval systems by using genetic programming and fuzzy logic. It includes the introduction of advanced Boolean operators (e.g., XOR and OF) and the assignment of weights to terms and operators to an increase of the query precision and recall. Two model is compared: the classic Boolean model and the extended version of the fuzzy model. The results of the experiment showed the accuracy of the retrieval progressively improved with the help of evolutionary algorithms that helped the user to optimize the queries [16]. The study is dedicated to exposing a novel optimization of the already known heuristic hill-climbing algorithm. It is done by careful candidate selection, which ensures both efficiency and accuracy. The preliminary findings provide the methodology of preventing similarity "brother nodes" selection, and it illustrates their usage to AI tasks. Experiments on the UCI machine learning repository demonstrate the developed solutions. The piece is devoted to beam search, and among other things, the process of improving the k-path selection strategy is required [17]. Extremal optimization is a technique that is unique in getting after certain optimization problems found in systems via the random changing of parameters. It avoids parameter tuning and it is a typical example of other techniques simulations use. It is good at finding the relatively small feasible solutions to standard problems like the bipartitioning and satisfiability ones [18].

III. EXISTING METHODOLOGY

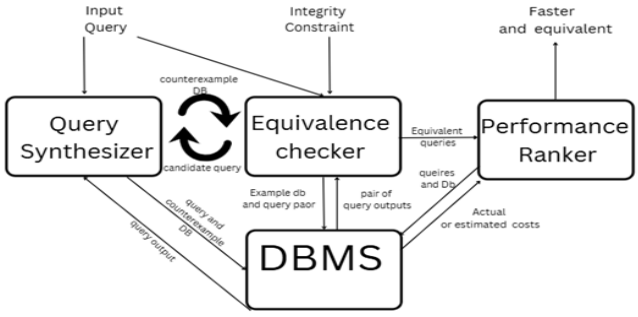

Figure 1. SlabCity: A framework for efficient query rewriting with synthesis, equivalence checking, and ranking

According to the methodology illustrated in Figure 1, dependency on the rewrite rules is reduced due to the introduction of SlabCity, a synthesis-based query rewriting framework [5]. It integrates three components such as the query synthesizer which enumerated candidate queries based on a novel dataflow-based which is the scoring function; Equivalence Checker which combines SMT-based verification and heuristic based testing for performing validation of semantic equivalence of queries; and the Performance Ranker, which incorporate either DBMS-executed EXPLAIN cost estimates or actual query execution matric to rank queries which are optimised. Queries are constructed by analyzing and prioritizing query dataflows, which indicates semantic operations. CEGIS loop iteratively synthesizes, test, and perform verification of candidate queries with the help of equivalence testing through bounded and full verification techniques. Search space is organized into layers of decreasing relevance based on monotonic scoring and optimization constraints is enforced. Substantial latency reductions across large-scale real-world workload is achieved with the help of this methodology.

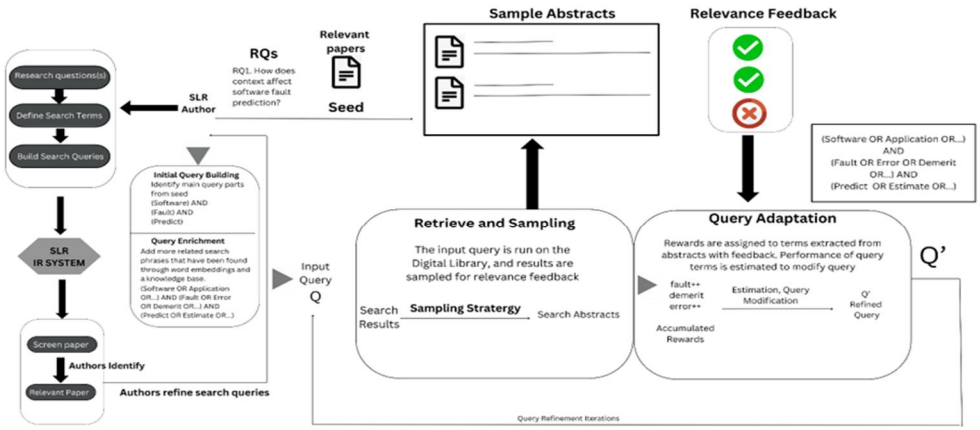

Figure 2. Query building and refinement pipeline starting with an SLR seed, generating an enriched query, and iteratively refining it through author feedback to enhance precision and relevance.

The methodology illustrated in Figure 2 describes an iterative and adaptive process of extraction, when developing and refining Boolean search queries [7] in systematic literature reviews (SLRs). It begins by stating research questions (RQs) and relevant search terms, which form the basis of making initial search queries. These queries are expanded with related semantic terms with the help of creating word embeddings and domain knowledge, thus enlarging the search space.

Expanded queries are put into the IR system with operations of Boolean, `AND`, and `OR` in search of relevant results from a digital library. After retrieving the results, a sampling strategy is employed to get abstracts for assessing their representative nature. The sampled abstracts are put under a relevance feedback loop; useful terms are rewarded while poor terms are penalized. This mechanism allows for iterated refinement of the query to reach an accommodation between precision and recall. From the dynamism represented in Figure 2, continuous query modification (`Q`), re-execution in the IR system, and improvement will happen until the results converge upon the predefined targets. This structured approach works to reduce the workload and, furthermore, ensure a wide and precise collection of relevant literature.

IV. Conclusion and Future Work

The state-of-the-art Boolean search optimization for patent retrieval is an integration of machine learning models, optimization algorithms, and efficient indexing techniques, making it very much scalable, relevance driven, and efficient in query performance. And examples include SetBERT with inversed-contrastive loss that enhances recall and reinforcement learning pipelines that enable real-time refinements of queries. Optimization techniques such as QAOA, PSO, GA, and GWO are used for handling high complexity to manage large patent datasets. Tree-like search engines extant, and using graph-based structures at least made the processing of the queries in question easier. Cloud-based or decentralized edge computing permits decentralized processing of queries, adding to its scalability. Besides ensuring more efficient memory usage and correctness in search results, vector database systems and graph-based systems, by working together, definitely enhance patent search system space efficiency and efficiency in general.

The future scope of the boolean query optimisation approach includes:

- Exploration of the applicability of SetBERT in domains such as video and image retrieval is included in the future work where boolean logic queries could also provide significant benefits. Final peer-reviewed queries are reflected by the PROSPERO dataset's search strategies.
- Models which are trained on final searches or PICO fields might perform better. The initial query restricts the set of the query plans that can be identified using ordinary query optimization approaches and query rebuilding techniques have been the subject of a lot of research.
- Still, a more investigation to determine how enhanced query rewriting affects the performance of traditional query tunning would be a fascinating domain of future work.

References

[1] Mai, Quan, Susan Gauch, and Douglas Adams. "SetBERT: Enhancing Retrieval Performance for Boolean Logic and Set Operation Queries." arXiv preprint arXiv:2406.17282, 2024.

[2] Adam, Gaelen P., Jay DeYoung, Alice Paul, Ian J. Saldanha, Ethan M. Balk, Thomas A. Trikalinos, and Byron C. Wallace. "Literature search sandbox: a large language model that generates search queries for systematic reviews." JAMIA open 7, no. 3 ooae098, 2024.
[3] Xu, Y., Gao, J., Gou, Y., Long, C. and Jensen, C.S., “iRangeGraph: Improvising Range-dedicated Graphs for Range-filtering Nearest Neighbor Search”. Proceedings of the ACM on Management of Data, 2(6), pp.1-26,2024.
[4] Pan, J.J., Wang, J. and Li, G., “Survey of vector database management systems”. The VLDB Journal, 33(5), pp.1591-1615,2024.
[5] Dong, R., Liu, J., Zhu, Y., Yan, C., Mozafari, B. and Wang, X., “SlabCity: Whole-Query Optimization Using Program Synthesis. Proceedings of the VLDB Endowment”, 16(11), pp.3151-3164,2023.
[6] Fankhauser, T., Solèr, M.E., Füchslin, R.M. and Stockinger, K., 2023. “Multiple query optimization using a gate-based quantum computer”. IEEE Access, 11, pp.114031-114043,2023.
[7] Badami, M., Benatallah, B. and Baez, M., “Adaptive search query generation and refinement in systematic literature review. Information Systems”, 117, p.102231,2023.
[8] Dadheech, P., Goyal, D., Srivastava, S. and Kumar, A., “An optimal framework for spatial query optimization using hadoop in big data analytics”. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 13(6), pp.1188-1198,2020.
[9] Qiu, T., Xie, P., Xia, X., Zong, C. and Song, X.,” Aggregated boolean query processing for document retrieval in edge computing”. Electronics, 11(12), p.1908,2022.
[10] Abdolrasol, M.G., Hussain, S.S., Ustun, T.S., Sarker, M.R., Hannan, M.A., Mohamed, R., Ali, J.A., Mekhilef, S. and Milad, A., “Artificial neural networks-based optimization techniques: A review”. Electronics, 10(21), pp.2689,2021.
[11] Pazhaniraja, N., S. Sountharrajan, and B. Sathis Kumar. "High utility itemset mining: a Boolean operators-based modified grey wolf optimization algorithm." Soft Computing 24, no. 21: 16691-16704,2020.
[12] Kotthoff, Lars. "Algorithm selection for combinatorial search problems: A survey." In Data mining and constraint programming: Foundations of a cross-disciplinary approach”, Cham: Springer International Publishing, pp. 149-190,2016.
[13] Mariot, Luca, and Alberto Leporati. "Heuristic search by particle swarm optimization of boolean functions for cryptographic applications." In Proceedings of the companion publication of the 2015 annual conference on genetic and evolutionary computation, pp. 1425-1426. 2015.
[14] Sadoghi, M. and Jacobsen, H.A., “Analysis and optimization for boolean expression indexing”, ACM Transactions on Database Systems (TODS), 38(2), pp.1-47,2013.
[15] Nassar, M.O., Al Mashagba, F. and Al Mashagba, E., “Improving the user query for the boolean model using genetic algorithms". International Journal of Computer Science Issues (IJCSI), 8(5), p.66,2011.
[16] Owais, Suhail SJ. "Optimization of Boolean Queries in Information Retrieval Systems Using Genetic Algorithms-Genetic Programming and Fuzzy Logic." Journal of Digital Information Management 4, no. 4,2006
[17] Xu, Z.W., Liu, F. and Li, Y.X.,” The research on accuracy optimization of beam search algorithm”, In 7th International Conference on Computer-Aided Industrial Design and Conceptual Design pp. 1-4. IEEE, November 2006.
[18] Boettcher, S. and Percus, A.G., Extremal optimization: an evolutionary local-search algorithm. Computational Modeling and Problem Solving in the Networked World: Interfaces in Computer Science and Operations Research, pp.61-77, 2003.
[19] Voorhees, E.M., “Natural language processing and information retrieval”, In International summer school on information extraction (pp. 32-48). Berlin, Heidelberg: Springer Berlin Heidelberg,1999.
[20] Boettcher, Stefan, and Allon G. Percus. "Extremal optimization: an evolutionary local-search algorithm. "Computational Modeling and Problem Solving in the Networked World: Interfaces in Computer Science and Operations Research pp: 61-77,2003