iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** II **Month of publication:** February 2022

DOI: https://doi.org/10.22214/ijraset.2022.40362

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue II Feb 2022- Available at www.ijraset.com

An Efficient Security Measure Extended Boolean Retrieval

S. Ramesh[1], V. Venkatesawara Rao[2]
[1, 2]*Student, M. Tech, NIET Guntur*

Abstract: ***The impact of Extended Boolean retrieval (EBR) has drastic changes over the decades and despite their significant advantages compared to either ranked keyword or pure Boolean retrieval. In particular, EBR models produce meaningful rankings; their query model allows the representation of complex concepts in an and-or format; and they are scriptable, in that the score assigned to a document depends solely on the content of that document, unaffected by any collection statistics or other external factors. Security is the major concern to make the queries retrieval and the entire data mining process. However, EBR is much more computationally expensive than the alternatives. We consider the implementation of the p-norm approach to EBR, and demonstrate that ideas used in the max-score and wand exact optimization techniques for ranked keyword retrieval can be adapted to allow selective by pass of documents via a low-cost screening process for this and similar retrieval models. We also propose term independent bounds that are able to further reduce the number of score calculations for short, simple queries under the extended Boolean retrieval model. Overall saving from 50 to 80 percent of the evaluation cost on test queries drawn from biomedical search. (Stefan Pohl, VOL. 24, NO. 6, JUNE 2012)***
Keywords: extended, boolen, model, EBR

I. INTRODUCTION

We present a scoring method for EBR models that decouples document scoring from the inverted list evaluation strategy, allowing free optimization of the latter. The method incurs partial sorting overhead, but, at the same time, reduces the number of query nodes that have to be considered in order to score a document. We show experimentally that overall the gains are greater than the costs. We adopt ideas from the max-score and wand algorithms and generalize them to be applicable in the context of models with hierarchical query specifications and monotonic score aggregation functions. Further, we show that the p-norm EBR model is an instance of such models and that performance gains can be attained that are similar to the ones available when evaluating ranked queries. Term-independent bounds are proposed, which complement the bounds obtained from max-score.After analysing the requirements of the task to be performed, the next step is to analyse the problem and understand its context. The first activity in the phase is studying the existing system and other is to understand the requirements and domain of the new system. Both the activities are equally important, but the first activity serves as a basis of giving the functional specifications and then successful design of the proposed system. Understanding the properties and requirements of a new system is more difficult and requires creative thinking and understanding of existing running system is also difficult, improper understanding of present system can lead diversion from solution. We also embedded the property of making the system secured and transparent.

II. SYSTEM ANALYSIS

A. SDLC Methdologies

This document play a vital role in the development of life cycle (SDLC) as it describes the complete requirement of the system. It means for use by developers and will be the basic during testing phase. Any changes made to the requirements in the future will have to go through formal change approval process.
SPIRAL MODEL was defined by Barry Boehm in his 1988 article, “A spiral Model of Software Development and Enhancement. This model was not the first model to discuss iterative development, but it was the first model to explain why the iteration models. As originally envisioned, the iterations were typically 6 months to 2 years long. Each phase starts with a design goal and ends with a client reviewing the progress thus far. Analysis and engineering efforts are applied at each phase.

The steps for Spiral Model can be generalized as follows:

1) The new system requirements are defined in as much details as possible. This usually involves interviewing a number of users representing all the external or internal users and other aspects of the existing system.
2) A preliminary design is created for the new system.
3) A first prototype of the new system is constructed from the preliminary design. This is usually a scaled-down system, and represents an approximation of the characteristics of the final product.

©IJRASET: All Rights are Reserved

4) A second prototype is evolved by a fourfold procedure:
a) Evaluating the first prototype in terms of its strengths, weakness, and risks.
b) Defining the requirements of the second prototype.
c) Planning an designing the second prototype.
d) Constructing and testing the second prototype.

5) At the customer option, the entire project can be aborted if the risk is deemed too great. Risk factors might involved development cost overruns, operating-cost miscalculation, or any other factor that could, in the customer's judgment, result in a less-than-satisfactory final product.
6) The existing prototype is evaluated in the same manner as was the previous prototype, and if necessary, another prototype is developed from it according to the fourfold procedure outlined above.
7) The preceding steps are iterated until the customer is satisfied that the refined prototype represents the final product desired.
8) The final system is constructed, based on the refined prototype.
9) The final system is thoroughly evaluated and tested. Routine maintenance is carried on a continuing basis to prevent large scale failures and to minimize down time.

The following diagram shows how a spiral model acts like:

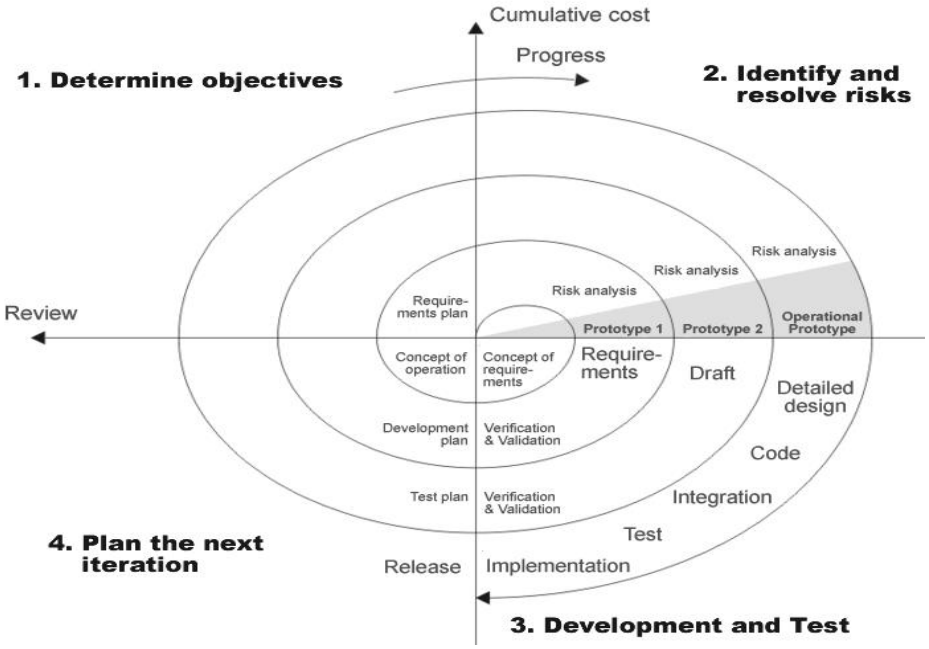


Fig 1-Spiral Model

III.STUDY ANALYSIS

A number of observations can be made from the results. First, although the CalcScore() scoring method requires partial sorting of tree node identifiers during document scoring, it has comparable execution times to a straightforward implementation that recursively iterates the query tree and computes scores for the next document in the ORset of that subtree (Tree Iteration), often being faster. However, this advantage might be due to localized access patterns and optimization of the scoring method. Second, the scoring method in the adaptation of maxscore yields significant reductions in the number of candidate documents scored, postings processed, and execution times, for all query sets. Third, nonunit p values lead to increased computational costs due to exponentiation. The more documents are to be retrieved, the less effective the top-k optimizations are, but even so, execution times are significantly below the baselines. Fourth, the TIB method for short-circuiting candidate document scoring is indeed able to reduce the number of score calculations, particularly on simpler queries and when combined with max-score. Indeed, for the "PUBMED Structured" queries the number of score calculations actually performed drops to within a factor of two of the minimal number that must inevitably be performed.

On the other hand, the TIB approach does not translate into significant computational time savings over the simpler adapted max-score. Only if nonbinary term weights are being used—when the computational cost associated with scoring is likely to be larger—the TIB methods can be anticipated to have a clear advantage. The time to compute the necessary TIB bounds rarely reached our limit of 50 ms, validating the feasibility of the approach. In experiments not reported here because of space constraints, conditional term-dependent bounds (TDB) turned out to be slightly better than TIB, but are very dependent on the choice of the term that is conditioned on. Fifth, while the query execution times of strict Boolean execution appear to be magnitudes faster, it must be remembered that the result of a Boolean query is of indeterminate size, and that a nontrivial number of these queries in fact return no results at all, while others return many thousands of documents. This uncertainty means that more Boolean queries are likely to be submitted by a user. Also, search service providers are free to decide which of the incoming queries they execute with the suggested EBR implementation, for instance, on basis of subscriptions by users that require the added benefits of the EBR model.
Finally, the gap between the disk-dominated and computational timings neatly correlates with the number of terms typically used in the query sets, and roughly reflects the number of term lookups necessary. It is also worth noting that "TIB only" sometimes has the best performance, but is not consistent in that regard.

IV. OUTPUTS

Below is the scrrenshot to retrive the Query Process

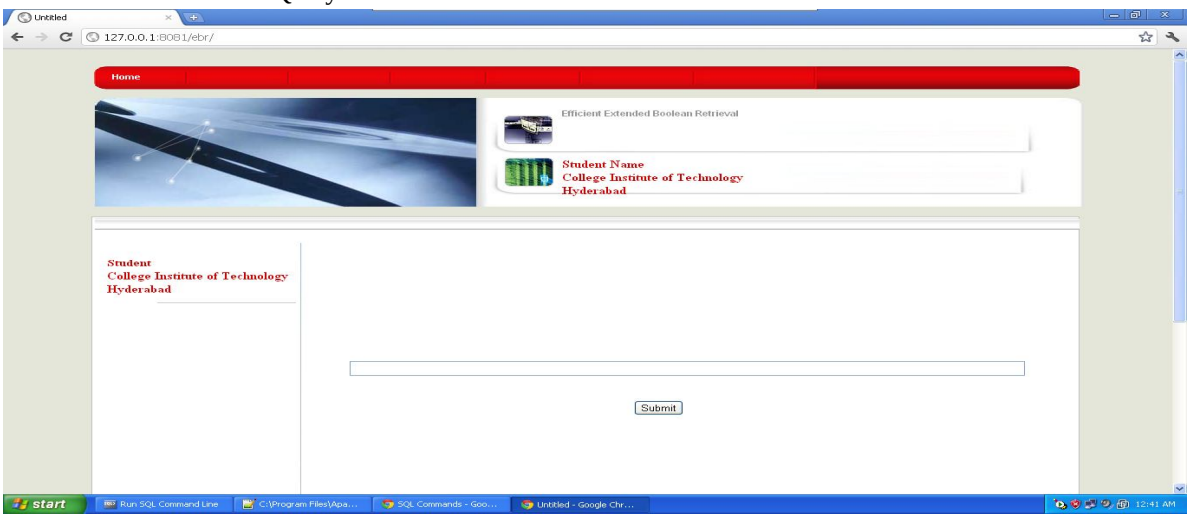


Fig 2:Query Process

The Query thus retrived can be selected through the selected the query which is retrived

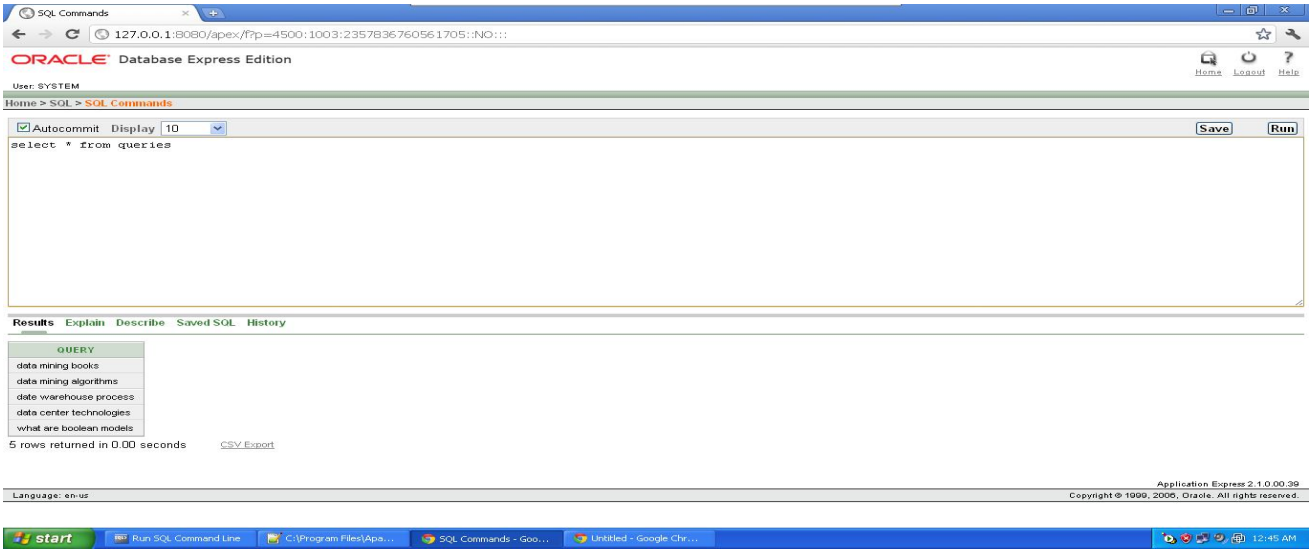


Fig 3:Query retrival after mining

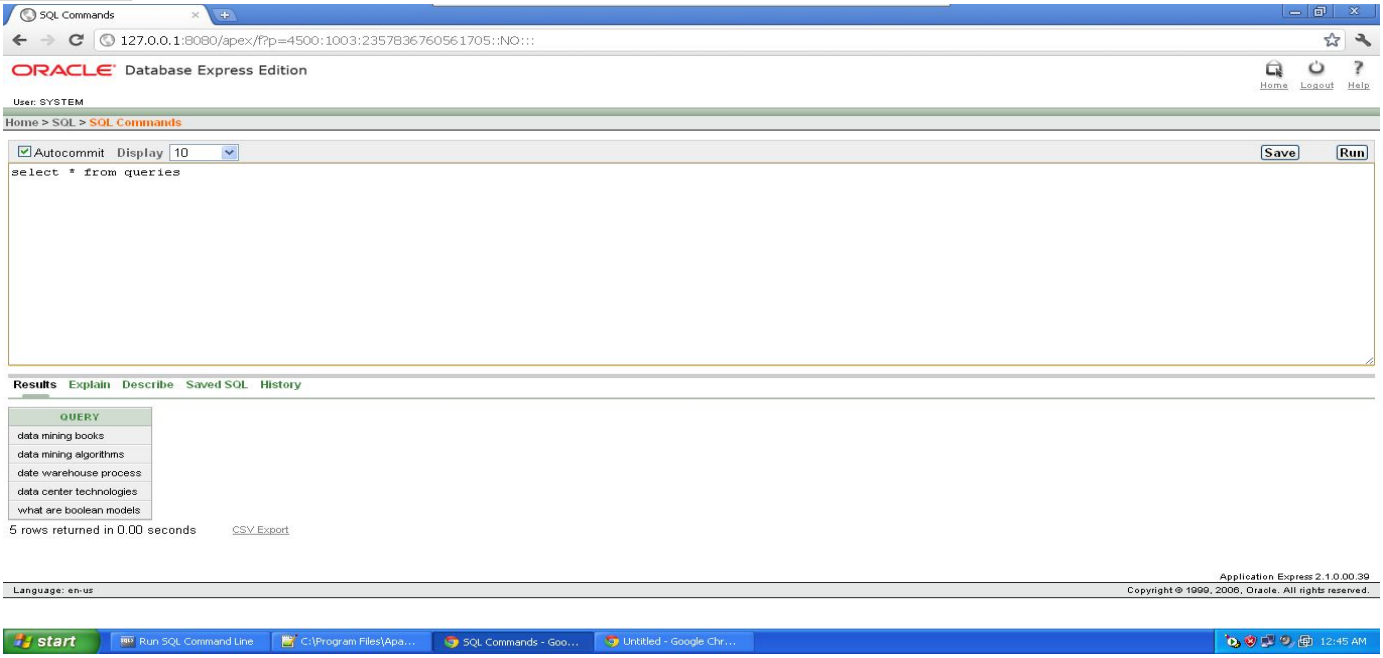


Fig 4: Verify the queries for processing

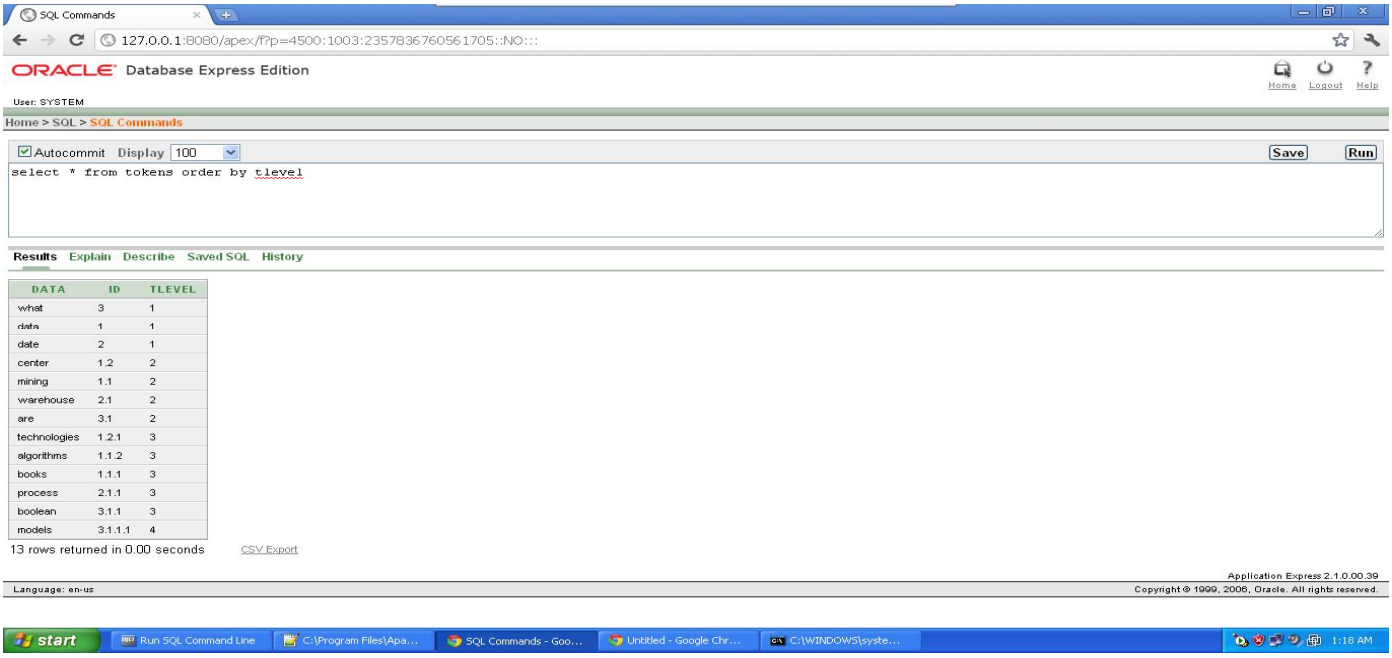


Fig 6: Table Retrieval

V. SYSTE SECURITY

1) SYSTEM SECURITY refers to the technical innovations and procedures applied to the hardware and operation systems to protect against deliberate or accidental damage from a defined threat.

2) DATA SECURITY is the protection of data from loss, disclosure, modification and destruction.

3) SYSTEM INTEGRITY refers to the power functioning of hardware and programs, appropriate physical security and safety against external threats such as eavesdropping and wiretapping.

4) PRIVACY defines the rights of the user or organizations to determine what information they are willing to share with or accept from others and how the organization can be protected against unwelcome, unfair or excessive dissemination of information about it.

5) CONFIDENTIALITY is a special status given to sensitive information in a database to minimize the possible invasion of privacy. It is an attribute of information that characterizes its need for protection.

6) CLIENT SIDE VALIDATION Various client side validations are used to ensure on the client side that only valid data is entered. Client side validation saves server time and load to handle invalid data. Some checks imposed are:

- VBScript in used to ensure those required fields are filled with suitable data only. Maximum lengths of the fields of the forms are appropriately defined.
- Forms cannot be submitted without filling up the mandatory data so that manual mistakes of submitting empty fields that are mandatory can be sorted out at the client side to save the server time and load.
- Tab-indexes are set according to the need and taking into account the ease of user while working with the system.

7) SERVER SIDE VALIDATION Some checks cannot be applied at client side. Server side checks are necessary to save the system from failing and intimating the user that some invalid operation has been performed or the performed operation is restricted. Some of the server side checks imposed is:

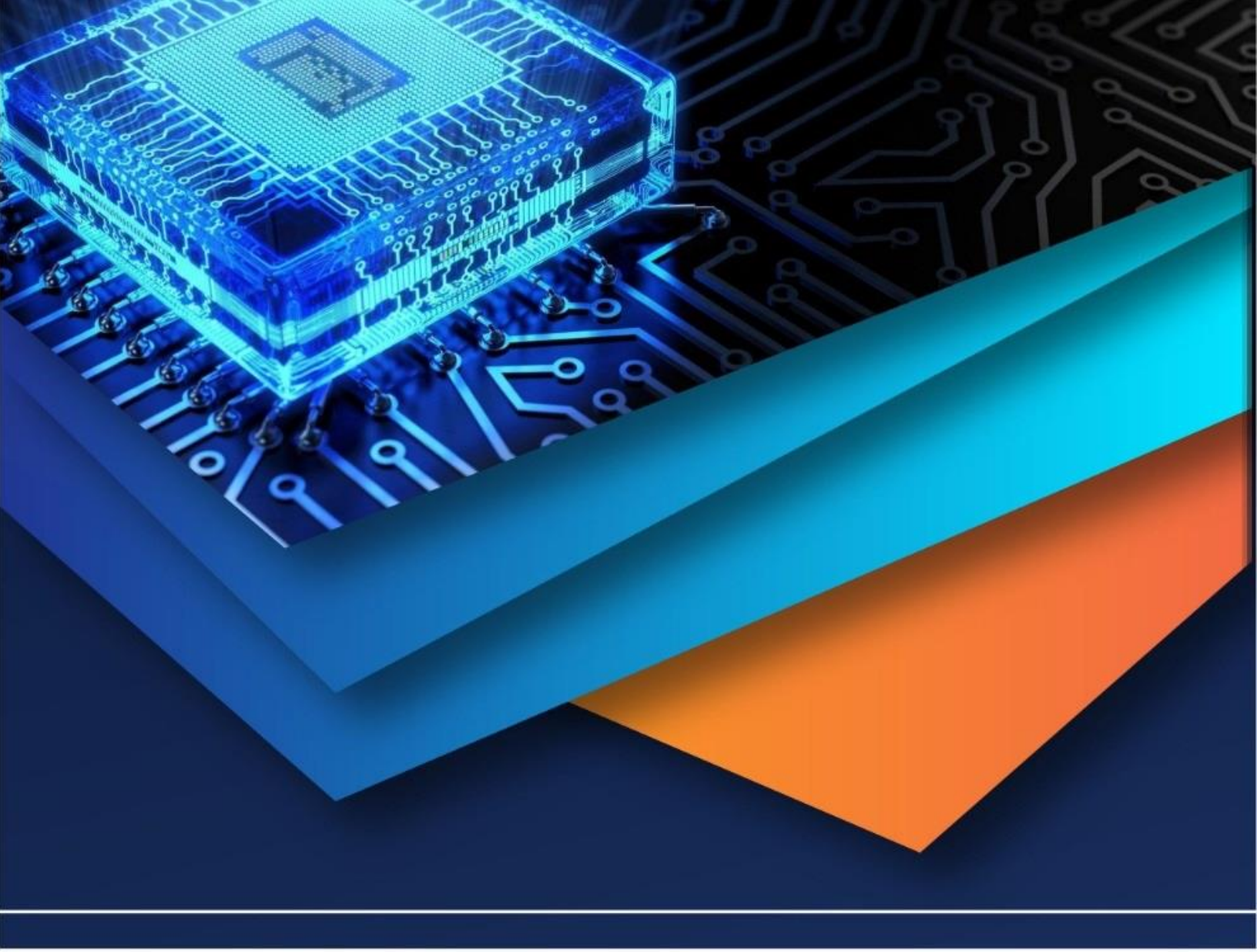
- Server side constraint has been imposed to check for the validity of primary key and foreign key. A primary key value cannot be duplicated. Any attempt to duplicate the primary value results into a message intimating the user about those values through the forms using foreign key can be updated only of the existing foreign key values.
- User is intimating through appropriate messages about the successful operations or exceptions occurring at server side.
- Various Access Control Mechanisms have been built so that one user may not agitate upon another. Access permissions to various types of users are controlled according to the organizational structure. Only permitted users can log on to the system and can have access according to their category. User- name, passwords and permissions are controlled o the server side.
- Using server side validation, constraints on several restricted operations are imposed.

VI. CONCLUSION

We proposed term-independent bounds as a means to further short-circuit score calculations, and demonstrated that they provide added benefit when complex scoring functions are used. A number of future directions require investigation. Although presented in the context of document-at-a-time Evaluation, it may also be possible to apply variants of our methods to term-at-a-time evaluation. Second, to reduce the number of disk seeks for queries with many terms, it seems desirable to store additional inverted lists for term prefixes (see, for example, Bast and Weber [29]), instead of expanding queries to hundreds of terms; and this is also an area worth exploration. We also need to determine whether or not term-dependent bounds can be chosen to consistently give rise to further gains. As another possibility, the proposed methods could further be combined and
applied only to critical or complex parts of the query tree. Finally, there might be other ways to handle negations worthy of consideration. We also plan to evaluate the same implementation approaches in the context of the inference network and wand evaluation models. For example, it may be that for the data we are working with relatively simple choices of term weights in particular, strictly document-based ones that retain the scrutability property that is so important—can also offer good retrieval effectiveness in these important medical and legal applications. (Stefan Pohl, VOL. 24, NO. 6, JUNE 2012).

REFERENCES

[1] G. Salton, E. F. (Nov. 1983). “Extended Boolean Information Retrieval . Comm. ACM, vol. 26, no. 1.
[2] J.H. Lee, W. K. (1993). “On the Evaluation of Boolean Operators in the Extended Boolean Retrieval Framework. Proc. 16th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval.
[3] Lee, J. (Cornell Univ., 1995). Analyzing the Effectiveness of Extended Boolean Models in Information Retrieva. Technical Report TR95-150.
[4] S. Karimi, J. Z. (Nov. 2009.). The Challenge of High Recall in Biomedical Systematic Search. Proc. Third Int’l Workshop Data and Text Mining in Bioinformatics.
[5] Stefan Pohl, A. M. (VOL. 24, NO. 6, JUNE 2012). Efficient Extended Boolean Retrieval. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)