Volume 3, Issue 9, September 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Scoring Method that Decouples Document Scoring from the Inverted List Evaluation Strategy, Allowing free Optimization

Mrs. K V Kiranmai[#1], Mr. Dr. Ch. GVN Prasad[#2], Mr.P.Appala Naidu[#3]

Hyderabad, (INDIA)

Abstract: :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. Taken alone, term-independent bounds can be employed in the wand algorithm, also reducing the number of score evaluations. Further, in conjunction with the adaption of max-score, this novel heuristic is able to short-circuit the scoring of documents.

Keywords: Boolean queries, significant, Composability, rehabilitation, score aggregation functions, hierarchical query specification

I. Introduction

Search service providers have an interest in delivering competitive effectiveness levels within the smallest possible resource cost. This is no less true in specialized search services dedicated to medical and legal literature, which are called upon to support complex queries by professional searchers, possibly with significant commercial or societal outcomes resting on the results of the search. In particular, although the number of queries submitted per day to biomedical search engines is orders of magnitude less than the number submitted to web-scale search systems (millions per day for pubmed,1 rather than billions per day for free web search), such services are typically funded as public services rather than by advertising; the queries are often much more complex, involving dozens or even hundreds of terms; there is a great deal of reformulation and reevaluation; and the user evaluation process typically involves hundreds or thousands of answer documents rather than a mere handful. Ranked rehabilitation has been successfully deployed in a wide range of applications. The main advantages of ranking are the simplicity of querying, and that results are ordered by estimated relevance, so that query quality can quickly be assessed once the top few results have been inspected. Having the answers returned as a ranked list also gives users the ability to consciously choose the amount of effort they are willing (or able) to invest in inspecting result documents. However, Boolean rehabilitation has not been superseded, and is still the preferred method in domains such as legal and medical search. Advantages of Boolean rehabilitation include: Complex information need descriptions: Boolean queries can be used to express complex concepts; Composability & Re-use: Boolean filters and concepts can be recombined into larger query tree structures; Reproducibility: Scoring of a document only depends on the document itself, not statistics of the whole collection, and can be reproduced with knowledge of the query; Referendum: Properties of retrieved documents can be understood simply by inspection of the query; and Strictness: Strict inclusion and exclusion criteria are inherently supported, for instance, based on metadata. For these reasons, Boolean rehabilitation – and the elongate Boolean variant of it that we pursue in this paper remains a critically important rehabilitation mechanism. For carefully formulated information needs, particularly when there are exclusion criteria as well as inclusion criteria, ranking over bags of words are not appropriate. As one particular example, recent results suggest that ranked keyword queries are not able to outperform complex Boolean queries in the medical domain. Boolean queries have the disadvantage of being harder to formulate than ranked queries, and, regardless of the level of expertise of the user, have the drawback of generating answer lists of unpredictable length. In particular, changes in the query that appear to be small might result in disproportionately large changes in the size of the result set. This is a problem that even expert searchers struggle with, adding and removing terms and operators until a reasonably sized answer set is retrieved, potentially even at the expense of rehabilitation effectiveness. Only when the answer set is of a manageable size can the searcher begin to invest time in examining its contents.

Elongate Boolean rehabilitation (EBR) models, such as the p-norm model, seek to rank on the basis of Boolean query specifications. They generate a list of top-k answers that can be elongate if required, without necessarily sacrificing detailed control over inclusion and exclusion of terms and concepts. But EBR queries are slow to evaluate, because of their

complex scoring functions; and none of the computational optimizations available for ranked keyword rehabilitation have been applied to EBR. In particular, approaches that involve non-exact methods, such as quantized impact-ordered indexes or index pruning, do not satisfy all of the requirements listed above. Our contributions in this paper are threefold. 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. Taken alone, term-independent bounds can be employed in the wand algorithm, also reducing the number of score evaluations. Further, in conjunction with the adaption of max-score, this novel heuristic is able to short-circuit the scoring of documents. We evaluate the efficiency of these methods on a large collection of biomedical literature using queries and results derived from real searches. Taken together, the optimizations greatly reduce query evaluation times for the p-norm EBR model on both short and complex queries, making EBR a competitive and viable choice for rehabilitation situations where such models are required. The results generalize to other models with hierarchical query specification and monotonic score functions.

II. System Overview

Existing System

In our Existing System, A significant amount of work has been devoted to the evaluation of top-k queries in databases. Provide a survey of the research on top-k queries on relational databases. This line of work typically handles the aggregation of attribute values of objects in the case where the attribute values lie in different sources or in a single source. For example, Bruno etc. Consider the problem of ordering a set of restaurants by distance and price. They present an optimal sequence of random or sequential accesses on the sources (e.g., Zagat for price and Mapquest for distance) in order to compute the top- k restaurants. Since the concept of top-k is typically a heuristic itself for locating the most interesting items in the database, Theobald et al.Describe a framework for generating an approximate top-k answer, with some probabilistic guarantees. In our work, we use the same idea; the main and crucial difference is that we only have "random access" to the underlying database (i.e., through querying), and no "sorted access." Theobald et al. assumed that at least one source provides "sorted access" to the underlying content.

Disadvantages of Existing System:

The queries are often much more complex, involving dozens or even hundreds of terms; there is a great deal of reformulation and reevaluation and the user evaluation process typically involves hundreds or thousands of answer documents rather than a mere handful.

PROPOSED SYSTEM

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. Taken alone, term-independent bounds can be employed in the wand algorithm, also reducing the number of score evaluations. Further, in conjunction with the adaption of max-score, this novel heuristic is able to short-circuit the scoring of documents.

Advantages of Proposed System:

Complex information need descriptions: Boolean queries can be used to express complex concepts.

Composability & Re-use: Boolean filters and concepts can be recombined into larger query tree structures.

Reproducibility: Scoring of a document only depends on the document itself, not statistics of the whole collection, and can be reproduced with knowledge of the query.

Referendum: Properties of retrieved documents can be understood simply by inspection of the query.

Strictness: Strict inclusion and exclusion criteria are inherently supported, for instance, based on metadata.

III. System Analysis

MODULES:

Using Boolean Condition (AND)
Using Boolean Condition (OR)
Using Boolean Condition (NOT)
Top k-Query Search

Modules Description:

Using AND Condition:

We define the novel problem of applying ranking on top of sources with no ranking capabilities by exploiting their query interface.

For instance, if the user query is Q= [anemia, diabetes, sclerosis], then we can submit to the data source queries q1 = [anemia AND diabetes AND sclerosis], q2 = [anemia AND diabetes AND NOT sclerosis], q3 = [diabetes OR sclerosis], and so on. The returned results are guaranteed to match the Boolean conditions but the documents are not expected to be ranked in any useful manner.

Using OR Condition:

We describe sampling strategies for estimating the relevance of the documents retrieved by different keyword queries. We present a static sampling approach and a dynamic sampling approach that simultaneously executes the query, estimates the parameters required for Profitable query execution, and compensates for the biases in the sampling process. For instance, if the user query is Q= [anemia, diabetes, sclerosis], then we can submit to the data source queries q1 = [anemia AND diabetes AND sclerosis], q2 = [anemia AND diabetes AND NOT sclerosis], q3 = [diabetes OR sclerosis], and so on. The returned results are guaranteed to match the Boolean conditions but the documents are not expected to be ranked in any useful manner.

Using NOT Condition:

We present algorithms that, given a user confidence input, retrieve a minimal number of results from the source through submitting high selectivity (conjunctive) queries, so that the user's confidence requirement is satisfied. For instance, if the user query is Q= [anemia, diabetes, sclerosis], then we can submit to the data source queries q1 = [anemia AND diabetes AND sclerosis], q2 = [anemia AND diabetes AND NOT sclerosis], q3 = [diabetes OR sclerosis], and so on.

The returned results are guaranteed to match the Boolean conditions but the documents are not expected to be ranked in any useful manner.

Top K-Query Search:

Our overall goal is to figure out during our querying process, how many of the top-k relevant documents we have retrieved and how many are still unretrieved in the database. Unfortunately, we cannot be absolutely certain about these numbers unless we retrieve and score all documents: an expensive operation. Alternatively, we can build a probabilistic model of score distributions and examine, probabilistically, how many good documents are still not retrieved. We describe our approach here.

Feasibility Report

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

Economical Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system Profitable ly. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

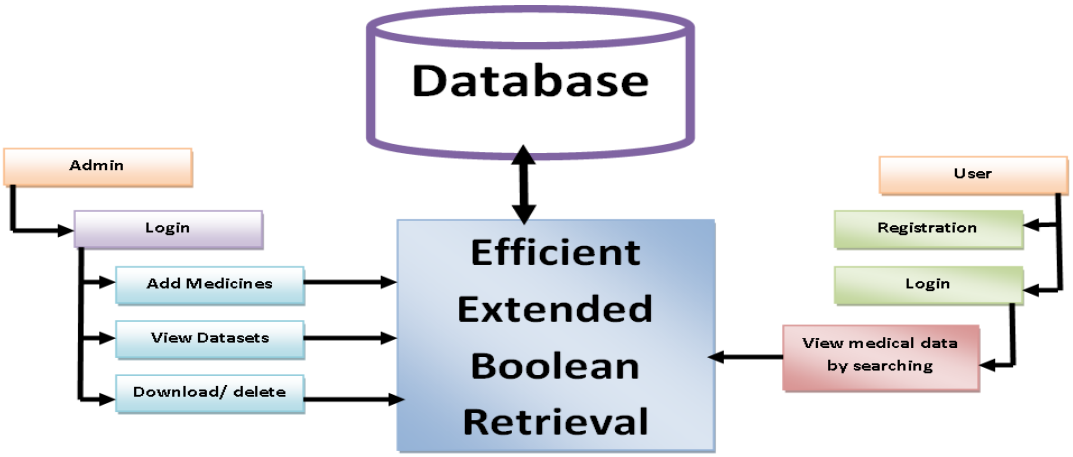
IV. System Design

Introduction To Design:

System design is transition from a user oriented document to programmers or data base personnel. The design is a solution, how to approach to the creation of a new system. This is composed of several steps. It provides the understanding and procedural details necessary for implementing the system recommended in the feasibility study. Designing goes through

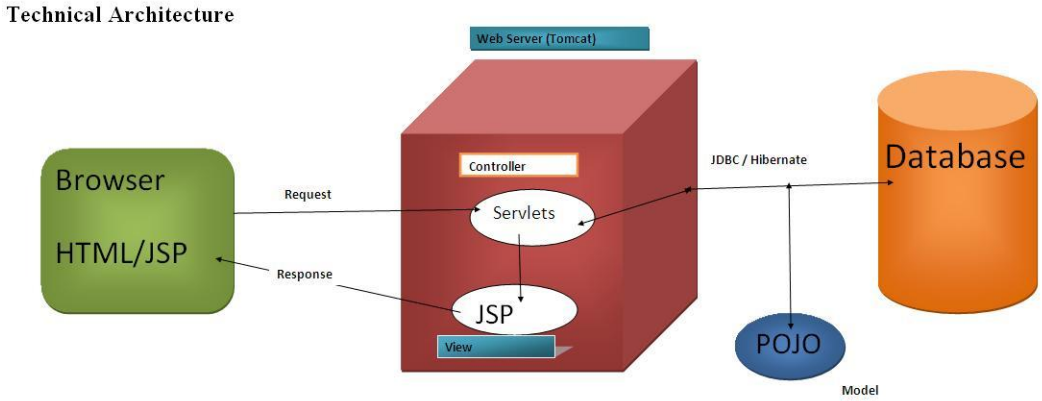
logical and physical stages of velopment, logical design reviews the present physical system, prepare input and output specification, details of mplementation plan and prepare a logical design walkthrough. The database tables are designed by analyzing functions involved in the system and format of the fields is also designed. The fields in the tabase tables should define their role in the system. The unnecessary fields should be avoided ecause it affects the storage areas of the system. Then in the input and output screen design, the design should be made user friendly. The menu should be precise and compact.

System Architecture:



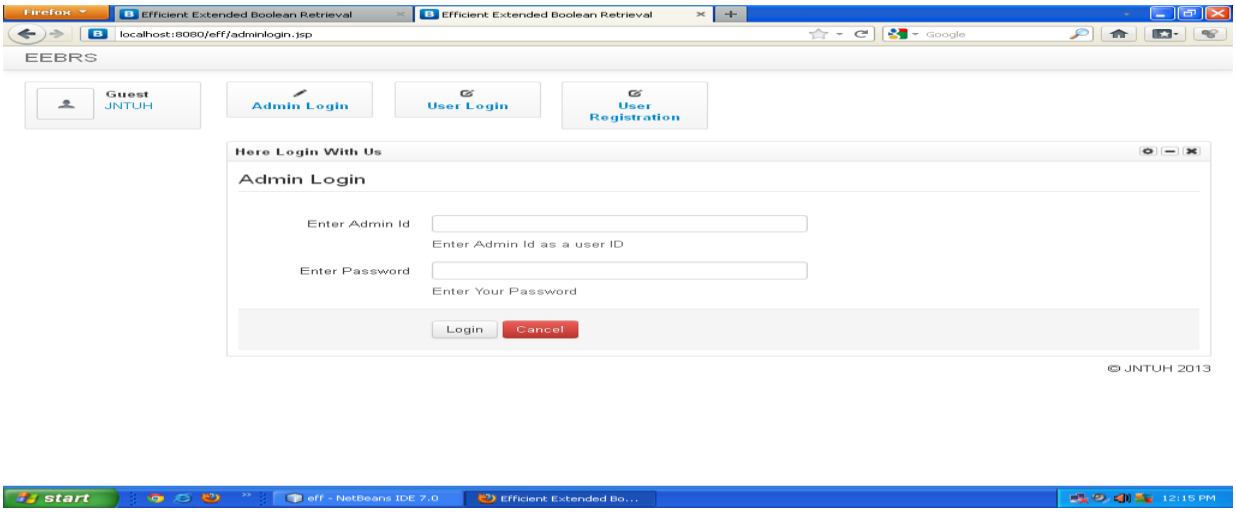
System Architecture

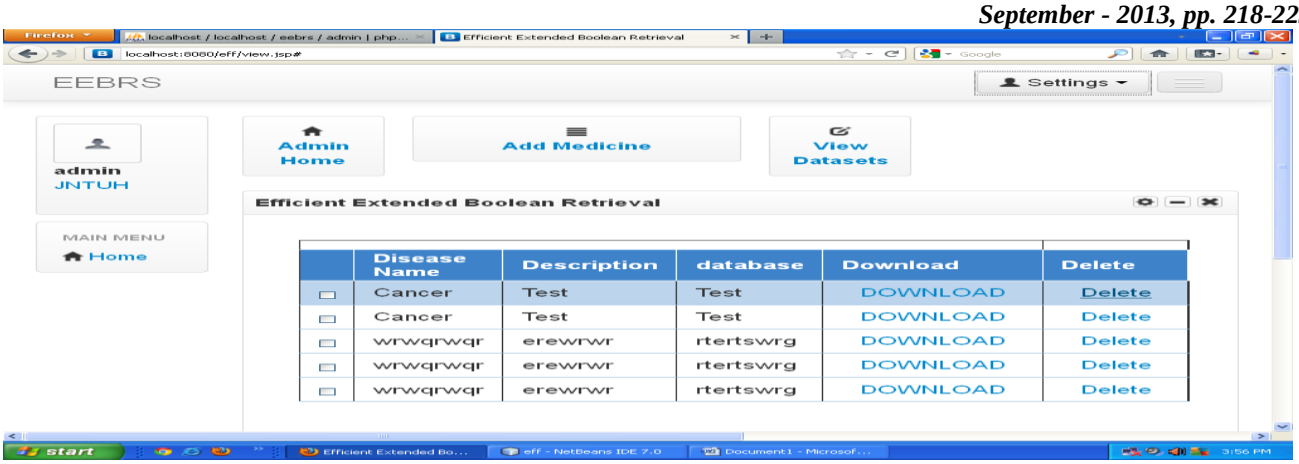
Technical Architecture:

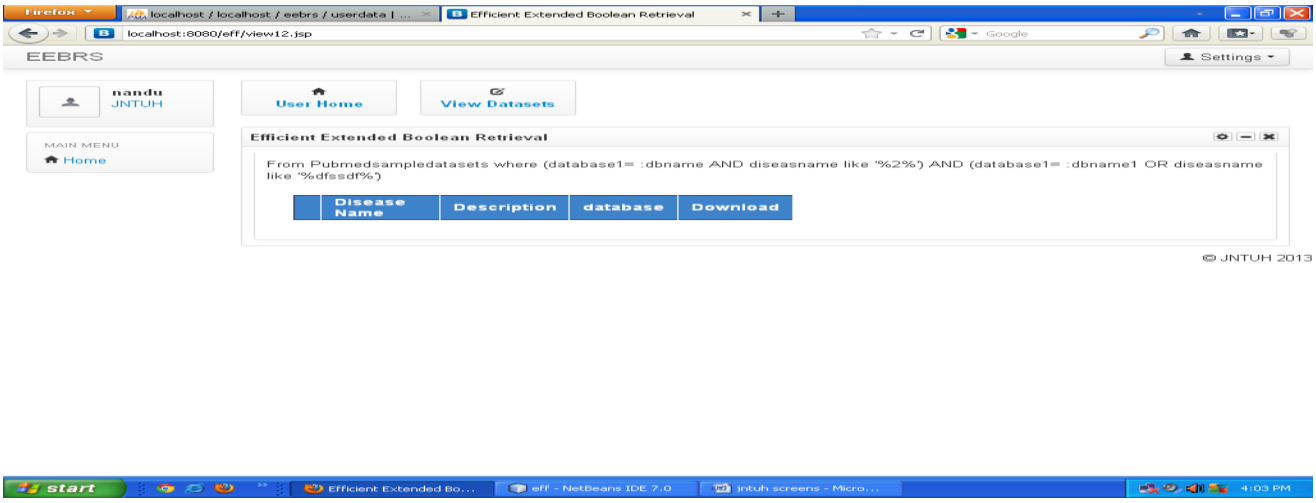


Technical Architecture

V. SYSTEM IMPLEMENTATION







VI. Conclusion

We have presented novel techniques for Profitable query evaluation of the p-norm elongate Boolean rehabilitation model, and applied them to document-at-a-time evaluation. We showed that optimization techniques developed for ranked keyword rehabilitation can be modified for EBR, and that they lead to considerable speedups. Further, 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. Secondly, to reduce the number of disk seeks for queries with many terms, it seems desirable to store additional inverted lists for term prefixes,instead of expanding queries to hundreds of terms; and this is also an area worth exploration.

Future Enhancements

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 referendum property that is so important .

References

[1] G. Salton, E. A. Fox, and H. Wu, "Elongate Boolean Information Rehabilitation," Commun. ACM, vol. 26, no. 11, pp. 1022–1036, Nov. 1983.

[2] J. H. Lee, W. Y. Kin, M. H. Kim, and Y. J. Lee, "On the evaluation of Boolean operators in the elongate Boolean rehabilitation framework," in Proc. Of the 16th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Rehabilitation. Pittsburgh, PA, USA: ACM, 1993, pp. 291–297.

[3] V. N. Anh and A. Moffat, "Pruned query evaluation using pre-computed impacts," in Proc. of the 29th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Rehabilitation. Seattle, WA, USA: ACM, 2006, pp. 372–379.

[4] J. P. T. Higgins and S. Green, Eds., Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2008.

[5] L. Zhang, I. Ajiferuke, and M. Sampson, "Optimizing search strategies to identify randomized controlled trials in MEDLINE," BMC Med. Res. Meth., vol. 6, no. 1, p. 23, May 2006.
[6] M. Sampson, J. McGowan, C. Lefebvre, D. Moher, and J. Grimshaw, "PRESS: Peer review of electronic search strategies," Ottawa: Canadian Agency for Drugs and Technologies in Health, Tech. Rep. 477, 2008.
[7] F. McLellan, "1966 and all that – when is a literature search done?" The Lancet, vol. 358, no. 9282, p. 646, Aug. 2001.

BIOGRAPHIES:

Mrs. K V Kiranmai ,Education Details: B.Tech-- CSE(computer science and engineering) 2006-2010,JNTUK. Studying M.Tech-- CSE(Computer Science And Engineering)2011-2013, Sri Indhu College Of Engineering And Technology, JNTUH ,Hyderabad.

Mr. Dr. Ch GVN Prasad, M.Tech,Ph.D(Experience-- 20 years ; 12 years IT industry (8 years in National Informatics Centre, Govt. of India, as Scientist and Software Analyst in AT&T in US) and 11 years Teaching as Professor and HOD of CSE dept). He Is Currently Working As Professor In Department Of Computer Science & Engineering In Sri Indu College of Engg & Tech. prasadch2042@gmail.com

P. APPALA NAIDU currently working as Assoc Prof. in department of computer science and engineering from Sri Indu College of Engineering and Technology, JNTU. He has obtained M. Tech (CSE) degree from Acharya Nagarjuna University. presently pursuing Ph. D in Computer Science and Engineering from Rayalaseema University, Kurnool. He has a teaching experience of 7 Yrs. He guided many UG and PG projects as supervisor. He has published several papers in international and national journals and conferences. Area of Specialization is Data Mining, DBMS,Automata Theory. psreenaidu@gmail.com

