International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

A Novel Approach with Feed Back Sessions for Inferring User Search Goals

Ratna Raju Mukiri [1]& Dr. G. Manoj Someswar [2]

[1] Associate Professor, Department of CSE, Eswar College of Engineering.
mukiriratnaraju001@gmail.com

[2] Professor, Department of CSE. Anwarul-uloom College of Engineering & Technology.
manojgelli@yahoo.co.in

ABSTRACT

For a broad-topic and ambiguous query, different users may have different search goals when they submit it to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. In this paper, we propose a novel approach to infer user search goals by analyzing search engine query logs. First, we propose a framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. Second, we propose a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. Finally, we propose a new criterion "Classified Average Precision (CAP)" to evaluate the performance of inferring user search goals. Experimental results are presented using user click-through logs from a commercial search engine to validate the effectiveness of our proposed methods.

KEYWORDS: User Search goals; feedback sessions; pseudo-documents; classified average precision

I. INTRODUCTION

Web mining is one of the applications of data mining techniques to discover knowledge from the web. In web search, users are submitted queries to the search engines to get relevant information. But many search engines results are not informative and failed to produce results according to the user search goals. Users are usually giving some vague keywords representing their interests in their minds. Such keywords do not match with the results produced by the search engines. Many works about user search goals analysis should be carried out.

Some users give ambiguous queries to the search engines (e.g. Apple, jaguar, the sun etc.) they get mostly the irrelevant results. User search goals are classified as Navigational and Informational, the queries that seek a single website or webpage and queries that reflect the intent of the user to perform a particular transaction respectively. Many related works have been carried out according to the web search applications and the user search goals. In previous works, clustering is done on a set of top ranked results.

The user search logs information is not analyzed and the feedback sessions are not considered. Analyzing the clicked URLS only from the web search logs. They only identify whether a pair of queries belong to the same goal or mission and does not care about what the goal is in detail. Semantic based web search for a particular query and the similarity between the words are carried out. Various algorithms such as star clustering algorithm, k-means clustering algorithm are used for clustering the pseudo documents but it also does not cluster the relevant information according to the user search goals.

In clustering the cluster labels discovered are also of informative. User search goal is the information on different aspects of a query that users wants to obtain. Information need is a user's wish/desire to obtain the relevant information to satisfy his need. To cluster web search results, the URLs are analyzed by extracting the titles and snippets. But all those works produced noisy results and does not obtain the user search goals precisely. When more irrelevant and relevant results are produced by the search engines it is time consuming to browse. In this paper, the user submits the query into the browser.

The search engine searches the relevant information according to the user query. The user actions are stored in the user click through logs. From the user click through logs each and every session is analyzed and generates the feedback session. The feedback session contains both the clicked and unclicked URLs and the last clicked URL in a single session. The feedback session contains the URLs and the click sequence. By analyzing the feedback sessions, the pseudo documents are generated. The pseudo documents contains the keywords that are most clicked in a session.

Likewise the pseudo documents are clustered using the clustering algorithm. The user search goals are obtained according to the feedback sessions. The restructure result is produced for the user query based on the user search goal. The CAP evaluation can be done for each user search goal and the clustering can be done to find the optimal number of users.

II. FRAMEWORK

Fig 2 shows the framework of our approach. Queries are submitted to search engines to represent the information needs of users. Ambiguous queries contain one or several polysemous terms. Query ambiguity is one of the main reasons for poor retrieval results (difficult queries are often ambiguous). User Click-through data log contains data about the interactions between users and Web search engines. It is one of the most extensive (yet indirect) surveys of user experience.

The user search information's are stored in the user click trough logs. . From the user click through logs the feedback sessions are constructed. The proposed feedback session consists of both clicked and unclicked URLs and ends with the last URL that was clicked in a single session. The feedback sessions is based on a single session, although it can be extended to the whole session.
The feedback session contains the URLs with the click sequence. A novel optimization method to combine the enriched URLs in a feedback session to form a pseudo-document. This can effectively reflect the information need of a user. The URLs are enriched from the feedback sessions based on the click sequence. The enriched URLs with more value in click sequence are mapped to pseudo-documents.

The pseudo documents are depicted with some keywords based on the URL. At last, cluster these pseudo documents to infer user search goals and depict them with some keywords. For clustering the pseudo-documents the fuzzy c-means clustering algorithm is used. The clustered pseudo documents are stored in the user search goals. From the user search goals the restructured results are produced.

A new criterion CAP to evaluate the performance of user search goal inference based on restructuring web search results. Thus the restructured web search result is produced. This proposed novel criterion "Classified Average Precision" to evaluate the restructure results.

III. ANALYZING USER CLICK THROUGH LOGS

The user click through logs is analyzed for each session to propose a feedback sessions. The feedback session is the better representation for the user click through logs. It is more efficient than analyzing the user click through logs directly. For a single query each and every session is analyzed and represents the feedback session. The feedback session is based on a single session although it can be extended to the whole session. An ambiguous query is that it gives more than one meaning. So the precise results according to the user search goal are difficult to obtain.

USER CLICK THROUGH LOGS

User Click-through data log contains data about the interactions between users and Web search engines. It is one of the most extensive (yet indirect) surveys of user experience. For researchers it helps to understand human interaction with IR results. The user click through logs contains all the user actions. It contains the session id, query term, position of the URL, click sequence and the URL.

FEEDBACK SESSIONS

The feedback sessions is discovered from each and every session from the user click through logs. The feedback sessions consists of the URLs that users visited and unvisited.

Using the click sequence, the order in which the URLs are visited by the users the feedback sessions are generated. The feedback sessions consists of URLs that contains the URLs from first URL and up to the last visited URL. The feedback session is based on the users browsing actions that are stored in user click through logs according to the particular query.

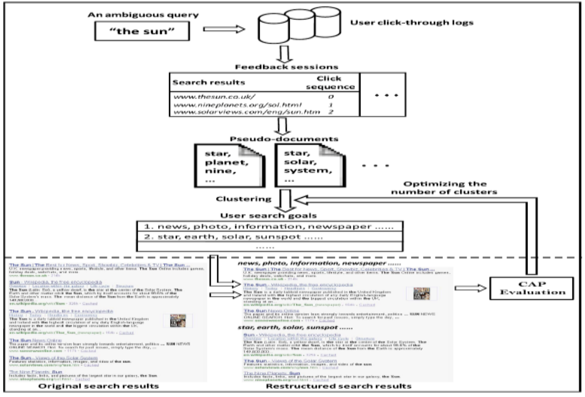

An ambiguous query
"the sun"
User click-through logs
Feedback sessions
Search results
Click sequence
www.thesun.co.uk/ 0
www.nineplanets.org/sol.html 1
www.solarviews.com/eng/sun.htm 2
. . .
Pseudo-documents
star, planet, nine, ...
star, solar, system, ...
. . .
Optimizing the number of clusters
Clustering
User search goals
1. news, photo, information, newspaper ......
2. star, earth, solar, sunspot ......
......
news, photo, information, newspaper ......
star, earth, solar, sunspot ......
CAP Evaluation
Original search results
Restructured search results

Fig. 2 Framework of our approach

IV. EXISTING SYSTEM

We define user search goals as the information on different aspects of a query that user groups want to obtain. Information need is a user's particular desire to obtain information to satisfy his/her need. User search goals can be considered as the clusters of information needs for a query. The inference and analysis of user search goals can have a lot of advantages in improving search engine relevance and user experience.

DISADVANTAGES OF EXISTING SYSTEM:

What users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical.

Analyzing the clicked URLs directly from user click-through logs to organize search results. However, this method has limitations since the number of different clicked URLs of a query may be small. Since user feedback is not considered, many noisy search results that are not clicked by any users may be analyzed as well. Therefore, this kind of methods cannot infer user search goals precisely.

Only identifies whether a pair of queries belongs to the same goal or mission and does not care what the goal is in detail.

V. PROPOSED SYSTEM

In this paper, we aim at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically.

We first propose a novel approach to infer user search goals for a query by clustering our proposed feedback sessions. Then, we propose a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. At last, we cluster these pseudo documents to infer user search goals and depict them with some keywords.

ADVANTAGES OF PROPOSED SYSTEM:

To sum up, our work has three major contributions as follows:

We propose a framework to infer different user search goals for a query by clustering feedback sessions. We demonstrate that clustering feedback sessions is more efficient than clustering search results or clicked URLs directly. Moreover, the distributions of different user search goals can be obtained conveniently after feedback sessions are clustered.

We propose a novel optimization method to combine the enriched URLs in a feedback session to form a pseudo-document, which can effectively reflect the information need of a user. Thus, we can tell what the user search goals are in detail.

We propose a new criterion CAP to evaluate the performance of user search goal inference based on restructuring web search results. Thus, we can determine the number of user search goals for a query.

PSEUDO DOCUMENTS

The pseudo documents are not the legitimate documents. The URLs in the feedback sessions are enriched by some format. The URLs are formatted by removing the stop words and the stemming words. It is the icon of showing the information about the whole document by some keywords.

The documents are created by the number of occurrences of the keywords. The keywords which are having the more frequency are grouped together. The pseudo documents contain the keywords that are retrieved from the URLs in the feedback sessions.

Using the Meta tag information the URLs are enriched. The Meta tag contains the most important keywords about the entire document information.

VI. CLUSTERING OF PSEUDO DOCUMENTS

Inferred user search goals from the pseudo documents by using clustering algorithm. The fuzzy self- constructing is used for the clustering of similar pseudo documents. The similarities of the keywords are grouped together and form the user search goals. Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" (all-or-nothing) but "fuzzy" in the same sense as fuzzy logic. Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible.

The FCM algorithm attempts to partition a finite collection of n elements into a collection of c fuzzy clusters with respect to some given criterion. Like the k- means algorithm, the FCM aims to minimize an objective function. For clustering of pseudo documents, the similarity of the documents is clustered using the fuzzy clustering. The same users in the same session have different goals at different times.

It is inappropriate to capture such overlapping interests of the users in crisp clusters. The fuzzy is used to discover different search goals. The similarity of the cluster is based on the centroid values. The search goals having least precision in one cluster may have to appear in another cluster with high precision. So discover different search goals for the users, the fuzzy clustering is used. The clusters are very informative and they are stored as the user search goals.

VII. LABELING THE CLUSTERS

A label will be generated to describe what each cluster is about. A user can then view the labels to decide which clusters to look into. The best cluster will have the high precision. Generate more meaning full cluster labels using the past keywords that are given by the users during the search. The keywords are derived from the user search logs.

Assuming that query words entered by users in the past that are associated with the current query can provide meaningful descriptions of the distinct aspects. Thus they can be better labels than those extracted from the ordinary contents of search results.

USER SEARCH GOALS

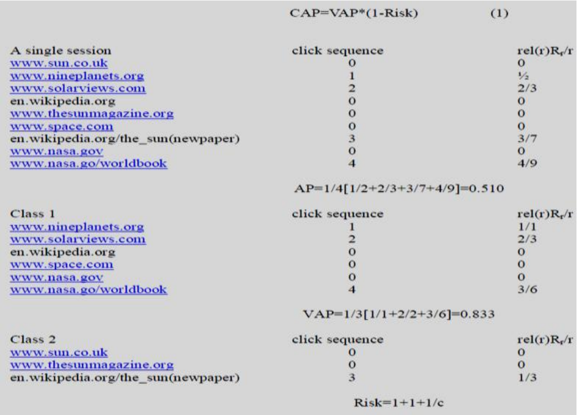
The clustering of pseudo-documents by fuzzy self-constructing clustering algorithm, which is simple and effective. Since we do not know the exact number of user search goals for each query. After clustering all the pseudo-documents,

VIII. EVALUATIONOF RESTRUCTURE SEARCH RESULTS

The evaluation of user search goals can be done using the CAP (CLASSIFIED AVERAGE PRECISION). The classified average precision is the calculation of precision of documents. Because from user click-through logs, we can get implicit relevance feedbacks, namely "clicked" means relevant and "unclicked" means irrelevant.

A possible evaluation criterion is the average precision (AP) which evaluates according to user implicit feedbacks. AP is the average of precisions computed at the point of each relevant document in the ranked sequence. VAP is the voted average precision which can be used for grouping the dissimilar documents for the particular user query search.

Risk is the mapping of similar and dissimilar documents for the particular user query. If there is a similarity then the mapping value is 0 and if there is no similarity between VAP and risk then the mapping value is 1.



CAP=VAP*(1-Risk) (1)

A single session	click sequence	$rel(r)R_r/r$
www.sun.co.uk	0	0
www.nineplanets.org	1	½
www.solarviews.com	2	2/3
en.wikipedia.org	0	0
www.thesunmagazine.org	0	0
www.space.com	0	0
en.wikipedia.org/the_sun(newpaper)	3	3/7
www.nasa.gov	0	0
www.nasa.go/worldbook	4	4/9

AP=1/4[1/2+2/3+3/7+4/9]=0.510

Class 1	click sequence	$rel(r)R_r/r$
www.nineplanets.org	1	1/1
www.solarviews.com	2	2/3
en.wikipedia.org	0	0
www.space.com	0	0
www.nasa.gov	0	0
www.nasa.go/worldbook	4	3/6

VAP=1/3[1/1+2/2+3/6]=0.833

Class 2	click sequence	$rel(r)R_r/r$
www.sun.co.uk	0	0
www.thesunmagazine.org	0	0
en.wikipedia.org/the_sun(newpaper)	3	1/3

Risk=1+1+1/c

IX. EXPERIMENTAL RESULTS

In this section, we will show experiments of our proposed algorithm. The data set that we used is based on the clickthroughlogs from a commercial search engine collected over a period of two months, including totally 2,300 different queries, 2.5 million single sessions and 2.93 million clicks. On average, each query has 1,087 single sessions and 1,274 clicks. However, these queries are chosen randomly and they have totally different click numbers. Excluding those queries with less than five different clicked URLs, we still have 1,720 queries. Before using the data sets, some preprocesses are implemented to the click-through logs including enriching URLs and term processing. In our approach, we have two parameters to be fixed: K in K-means clustering and _ in (10). When clustering feedback sessions of a query, we try five differentKð1; 2; . . . ; 5Þ in K-means clustering.

Then, we restructure the search results according to the inferred user search goals and evaluate the performance by CAP, respectively. At last, we select K with the highest CAP. Before computing CAP, we need to determine _ in (10). We select 20 queries and empirically decide the number of user search goals of these queries.

Then, we cluster the feedback sessions and restructure the search results with inferred user search goals. We tune the parameter _ to make CAP the highest when K in K-means accord with what we expected for most queries. Based on the above process, the optimal _ is from 0.6 to 0.8 for the 20 queries. The mean and the variance of the optimal _ are 0.697 and 0.005, respectively. Thus, we set _ to be 0.7. Moreover, we use another 20 queries to compute CAP with the optimal _ (0.7) and the result shows that it is proper to set _ to be 0.7. In the following, we will first give intuitive results of discovering user goals to show that our approach can depict user search goals properly with some meaningful words. Then, we will give the comparison between our method and the other two methods in restructuring web search results.

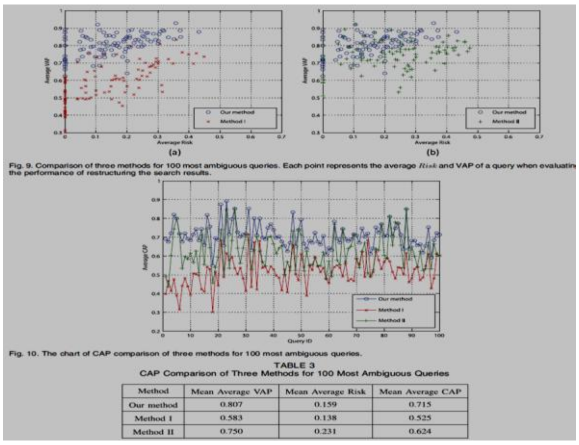


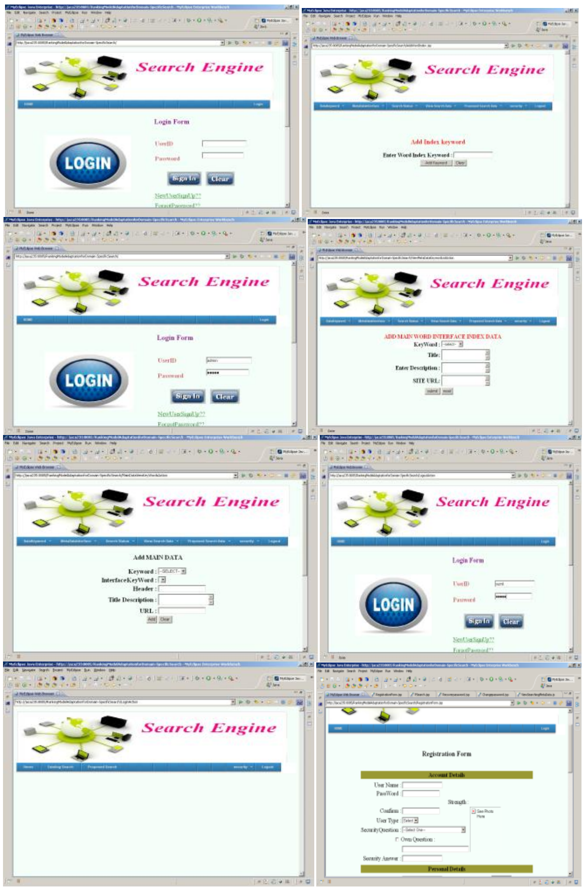
Fig. 9. Comparison of three methods for 100 most ambiguous queries. Each point represents the average *Risk* and VAP of a query when evaluatin the performance of restructuring the search results.

Fig. 10. The chart of CAP comparison of three methods for 100 most ambiguous queries.

TABLE 3
CAP Comparison of Three Methods for 100 Most Ambiguous Queries

Method	Mean Average VAP	Mean Average Risk	Mean Average CAP
Our method	0.807	0.159	0.715
Method I	0.583	0.138	0.525
Method II	0.750	0.231	0.624

X. SREEN SHOTS



XI. CONCLUSION

In this paper, a novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents.

First, we introduce feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Both the clicked URLs and the unclicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions. Therefore, feedback sessions can reflect user information needs more efficiently.

Second, we map feedback sessions to pseudo documents to approximate goal texts in user minds. The pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets. Based on these pseudo-documents, user search goals can then be discovered and depicted with some keywords.

Finally, a new criterion CAP is formulated to evaluate the performance of user search goal inference. Experimental results on user click-through logs from a commercial search engine demonstrate the effectiveness of our proposed methods. The complexity of our approach is low and our approach can be used in reality easily.

For each query, the running time depends on the number of feedback sessions. However, the dimension of Ffs in (3) and (5) is not very high. Therefore, the running time is usually short. In reality, our approach can discover user search goals for some popular queries offline at first.

Then, when users submit one of the queries, the search engine can return the results that are categorized into different groups according to user search goals online. Thus, users can find what they want conveniently.

XII. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 1999.

[2] R. Baeza-Yates, C. Hurtado, and M. Mendoza, “Query Recommendation Using Query Logs in Search Engines,” Proc. Int’l Conf. Current Trends in Database Technology (EDBT ’04), pp. 588-596, 2004.

[3] D. Beeferman and A. Berger, “Agglomerative Clustering of a Search Engine Query Log,” Proc. Sixth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’00), pp. 407-416, 2000.

[4] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, “Varying Approaches to Topical Web Query Classification,” Proc. 30th Ann. Int’l ACM SIGIR Conf. Research and Development (SIGIR ’07), pp. 783-784, 2007.

[5] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, “Context-Aware Query Suggestion by Mining Click-Through,” Proc. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’08), pp. 875-883, 2008.

[6] H. Chen and S. Dumais, “Bringing Order to the Web: Automatically Categorizing Search Results,” Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI ’00), pp. 145-152, 2000.

[7] C.-K Huang, L.-F Chien, and Y.-J Oyang, “Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs,” J. Am. Soc. for Information
Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.

[8] T. Joachims, “Evaluating Retrieval Performance Using Clickthrough Data,” Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.

[9] T. Joachims, “Optimizing Search Engines Using Clickthrough Data,” Proc. Eighth ACM SIGKDD

Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’02), pp. 133-142, 2002.

[10] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, “Accurately Interpreting Clickthrough Data as Implicit Feedback,” Proc. 28th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’05), pp. 154-161, 2005.

[11] R. Jones and K.L. Klinkner, “Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs,” Proc. 17th ACM Conf. Information and Knowledge Management (CIKM ’08), pp. 699-708, 2008.

[12] R. Jones, B. Rey, O. Madani, and W. Greiner, “Generating Query Substitutions,” Proc. 15th Int’l Conf. World Wide Web (WWW ’06), pp. 387-396, 2006.

[13] U. Lee, Z. Liu, and J. Cho, “Automatic Identification of User Goals in Web Search,” Proc. 14th Int’l Conf. World Wide Web (WWW ’05), pp. 391-400, 2005.

[14] X. Li, Y.-Y Wang, and A. Acero, “Learning Query Intent from Regularized Click Graphs,” Proc. 31st Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’08), pp. 339-346, 2008.

[15] M. Pasca and B.-V Durme, “What You Seek Is what You Get: Extraction of Class Attributes from Query Logs,” Proc. 20th Int’l Joint Conf. Artificial Intelligence (IJCAI ’07), pp. 2832-2837, 2007.

XIII. AUTHORS PROFILE

Ratna Raju Mukiri M.Tech(CSE), S.E.T.,(P.hD)., is having 10+ years of experience in the field of teaching in various Engineering Colleges and PG colleges. At present he is working as Assoc. Professor in Eswar College of Engineering, Narasaraopet, Guntur, India. He published 7 international journals and attended 2 national conference and 1 international conference and qualified state eligibility test twice in 2012 & 2013. He has given many guest lecturers to M.C.A. students in the subject areas of Micro processors, artificial intelligence, data structures etc., He also guided many **B.Tech**, **MCA** and **M.Tech** projects. He attended two weeks **ISTE workshop** on “**Data Base Management Systems**” conducted by **IIT Bombay**. His interested areas are Data Mining, Mobile Computing, Software Project Management, E - Commerce, C Programming, Computer Networks, UNIX Programming, Data Structures, Data Base Management Systems, etc. He can be contacted with email mukiriratnaraju001@gmail.com, and Phone No. 9700763540, 9848439256.

Dr. G. Manoj Someswar, B.Tech., M.S.(USA), M.C.A., Ph.D. is having 20+ years of relevant work experience in Academics, Teaching, Industry, Research and Software Development. At present, he is working as PRINCIPAL and Professor CSE Department Anwarul-uloom College of Engineering & Technology, Yennepally, Vikarabad - 501101, RR Dist., A.P., and utilizing his teaching skills, knowledge, experience and talent to achieve the goals and objectives of the Engineering College in the fullest perspective. He has attended more than 100 national and international conferences, seminars and workshops. He has more than 10 publications to his credit both in national and international journals. He is also having to his credit more than 50 research articles and paper presentations which are accepted in national and international conference proceedings both in India and Abroad. He received National Awards like Rajiv Gandhi Vidya Gold Medal Award for Excellence in the field of Education and Rashtriya Vidya Gaurav Gold Medal Award for Remarkable Achievements in the field of Education. He can be contacted at: manojgelli@yahoo.co.in.