Predicting Ranking for Scientific Research Papers using Scalable TensorFlow Library and Learning to Rank

Sarabu Joshna

Abstract: Scientific research papers play a vital role for innovation of new technology. It is the future of the development where a novice person can understand the technology and tries to develop a new idea. In this paper, concentrated on relative order for a group of items applied to scientific research paper. In this process we identify how LTR differs from standard supervised learning in the sense that instead of looking at a precise score or class for each sample, it aims to discover the best relative order for a group of items. Firstly we identified the work of ranking of scientific research papers using traditional method know as supervised learning. Secondly we evaluated and made the comparison between the supervised learning and the scalable Tensor flow library for learning to rank.

Apart from solving information retrieval problems, Learning to Ranking is mostly used in areas like Natural language processing (NLP), Machine translation, Computational biology or Sentiment analysis.

Keywords : Keywords: TensorFlow, PageRank, Binarization, Heatmap.

I. INTRODUCTION

The vast development of information retrieval systems made the universities across the world to use the access for their research activities .Scientific research papers are one of the asset to perform their research work.

Now a day's ranking the scientific research papers has become a interesting area as its providing the priority to most relevant and authors papers. But every ranking algorithm is suffering from one or the other challenges.

Machine learning is a trending application which concentrates the ranking scientific research papers in a crucial research direction which mainly identified the relative order for a group of items.

To be precise a ranking problem is different from classification or regression tasks. It s different in the setting a label is given to each individual paper, to determine the relevance of the paper determine the relevance ordering of the entire research papers. Most of the existing learning to rank algorithms model are constrained to point wise functions. This would lead to many problems due to many reasons.

Revised Manuscript Received on February 28, 2020.
* Correspondence Author
Sarabu Joshna*, Assistant Professor in the Department of Computer Science Engineering ,Bharath Institute of Engineering And Technology, Mangalpally, Ibrahimpatnam, Hyderabad, India.

From the view of machine learning there are three approaches i.e. Point wise , pair wise and line wise. These classifiers allow representing the relation between the different authors and papers. It is used to identify the best relations between two aspects. The main reason to choose this approach is that a researcher tries to identify the relation between the research paper related to their research area and the author who best suit their needs.

To have a clear vision of these approaches and the methodology there have used for the research papers which classified the papers on two axes: the horizontal axis according to the query and second axes .According to the technical analysis link the limitation of supervised learning approach is that it concentrates on the precise score or class but using this approach it has limitation of identifying the score for each class but this could be avoided using learning to rank which identified the score for group of items.

The paper is organized according to the following sections like in the first section gives introduction in the second section related work is described in the third section Learning to rank is explained and it follows by experiment results and conclusion and describe future research directions.

II. RELATED WORK

Now a days ranking has become a important to factor in retrieving the relevant paper and the main aspect to judge the research paper is the impact factor [6], is an index which is scientometric in nature that gives the yearly average number of citations that recent articles published in a given journal received. Among the studies we find:

A. Conceptual Graph model

In this model it identifies the relationship between the items using the concepts that exist between them in a sentence. It constructs a graph based on the pattern [1].

B. N-grams

In this model through N-grams algorithm spelling check is done. The resultant is called as N-grams which is considered as group of characters that are uprooted from the query [2].

C. Okapi BM25

The main inspiration of this model is probabilistic retrieval where the documents are matched according the there probability of matching with the query. WSD In this model ambiguity that exists between the words is eliminated by indexing them semantically to correct sense the indexed words.

For effective retrieval of the relevant document sense based approach is given more important than keyword based [3].

The main inspiration of this model is probabilistic retrieval where the documents are matched according the there probability of matching with the query. WSD In this model ambiguity that exists between the words is eliminated by indexing them semantically to correct sense the indexed words. For effective retrieval of the relevant document sense based approach is given more important than keyword based [3].

In recent years, learning to rank algorithm has came into picture where a class of supervised machine learning techniques is used to overcome the problems of ranking. this is done through grouping the items instead of ranking the single item. The main aim of learning to rank algorithm is to identify from labeled item that maps the vector into scores which is real valued in algorithm. It plays a vital role in learning to rank to sort and rank the research papers. To solve the problem of ranking in machine learning used three approaches.

Among the studies we find According to the technique of link analysis we can add the important ranking model as the second edge.

Among the models we find:

Point wise method:

In this method each sample of ranking is mapped into either continuous or discrete value which converts the problem into regression/classification.

Pair wise:

This method uses regression or classification which mainly identifies the relation between two items at a time and it builds the ranking on the group. This process is repeated for all the items in the group.

List Wise:

In this approach it defines the loss function to the whole list of items instead of single item which is done in point wise method.

III. LEARNING TO RANK

In this section we provide how learning to rank technique is used to rank the scientific research papers. Now we begin with the brief description of the learning to rank algorithm.

In order to evaluate the ranking for a specific research paper we need to identify the quality of a ranking which is a direct optimization than using direct ranking metric. In order to improve the quality of the ranking for a given scientific research paper we need to calculate Discounted cumulative gain. The main intention of using the the DCG is to get the relationship that exist between the items in the list using the logarithm.

$$DCG\ p = \sum_{i=1}^{p} \frac{2^{relevance}\ i-1}{\log_2(i+1)} \qquad (1)$$

The above logarithmic is used vastly than the point wise or pair wise approaches because this gives the efficient relationship between the items.

The main aim of the learning to rank is to learn a ranking function from a given training data .So the above logarithm is used on the given data set in order to get the relation between the items. Further we will see how this logarithm is applied for data set.

A. Tensor Flow Library for Training the Dataset

Tensor Flow Library is used for ranking the paper where it is providing scalable on large amount of data. The problem that a raised using the Lambda Loss is overcome by TF-Ranking using the library of Tensor Flow.

Firstly we will look at the problem which arises from the traditional methods. consider the query which is used to retrieve a research paper based on the relevance so in order to provide the efficient output to the user we are using ranking system. Without concept of ranking which is done automatically we need the advice of human expert where they will give the suggestion of which paper should be taken in order to get the knowledge of concerned domain .This is having a drawback of people should depend on the expert in order to get the efficient output.

One of the technique to train the dataset is whenever the user click on the research paper the click logs is counted either (x=0) or (x=1). In both the Cases x results in a partial ranking documents.

B. Ranking Scientific Research Paper using Learn To Rank Framework

Many ranking model have emerged to rank the data items among them Machine learning technologies has become so popular to rank data item among them learning to rank is one of the model to rank the data set.

Generally all the methods uses the feedback which is related and turns the features automatically using Information Retrieval model. But most of the ranking algorithms uses the combination of features extracted from query document through the training data set.

The following two properties are necessary for ranking methods to call them as learning to rank algorithms.

Feature Based:

The documents used to rank are represented in the form of feature vectors which reflects how the documents are related to query. To represent them as vectors we take Φ is a feature extractor in which for a given query u and its relevant document r can be represented as vector y= Φ(r,u) .one of the features or terms need to consider while implementing the ranking the scientific research paper is query keywords, page rank and the relationship that exist between the paper and other research papers. One of the assumption done while ranking the research paper using learning to rank is that parameter in the model is fixed even if the feature is the output of an existing retrieval model. The main advantage of learning to ranking is that even if we use different model to rank the scientific research paper is that for previously automated parameter can be incorporated in to the learning to rank so that progress can be visible by including the output of one of the feature .This advantage made the learning to rank to rank the documents not only for scientific research paper but also the areas where the complex information is to be ranked.

Discriminative training:

In discriminative methods feature of different kinds are widely used without necessity of representing the framework to the objects and prediction.

Learning to rank method has its particular input space, output space ,hypothesis space and loss function. so as it is having its specific functions previously done training data is not considered as learning to rank. This training is used to rank the scientific research paper as it will receive a lot of researcher feedback and usage logs which identifies the poor ranking of research papers or some queries. So they is emergence to learn from feedback and improve the ranking methodology.

For a given research paper we have a given to search the relevant paper so the relevant papers are taken as training set for given n queries qj=(j=1,2,….n), their relevant documents are represented as feature vectors[j]={y(j)i)n(j)i=1(where n(j) is the no of research papers with query qj a typical training set consists of n training queries qi(i = 1,...,n), their associated documents represented by feature vectors x (i) = {x (i) j } m(i) j=1 (where m(i) is the number of documents associated with query q(i), and the corresponding relevance judgments. After training the research papers a rank is allocated when the query is given the testing phase makes the relevant documents to be the highest priority.

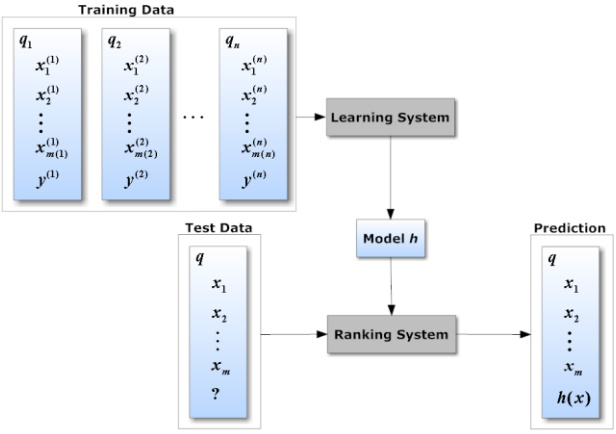


Fig 1. Learning-to-rank framework.

C. Approaches to Rank Scientific Research paper

The above explained framework can be used to any learning to rank algorithms. So to have a clear idea on this organization of these algorithms we can perform categorization. So already discussed in the above section the approaches like pointwise, pair wise approach, and the listwise approach. These approaches have their own technique to rank a research paper such as they may have their own input space, output space, loss functions and hypothesis each have they own advantage and drawback.

For the given approach the output space which is generated from learning to ranking algorithm using different approach may have they own format of determination or decision The pointwise approach

In this approach the input space for pointwise is the feature vector for one research paper. After applying the point wise approach output space consists of relevance degree of each scientific research paper. The conversion of input space to output space is done with help of hypothesis space that contains the functions from the feature vector and using the scoring functions it predict the relevance degree of the scientific research paper. Using this score final ranked list of scientific research papers are generated according to the ranking generated.

Pairwise classification or regression

RankNet

In ranking the scientific research papers there is a chance to mismark the paper so in order to avoid such problem the RankNet has introduced where its main aim is to avoid the number of inversions in ranking.

LambdaMART

In this algorithm it makes use of decision trees for determining the rank of documents .on experiments of documents [7] it is revealed that LambdaMart is better than other ranking algorithms[4].

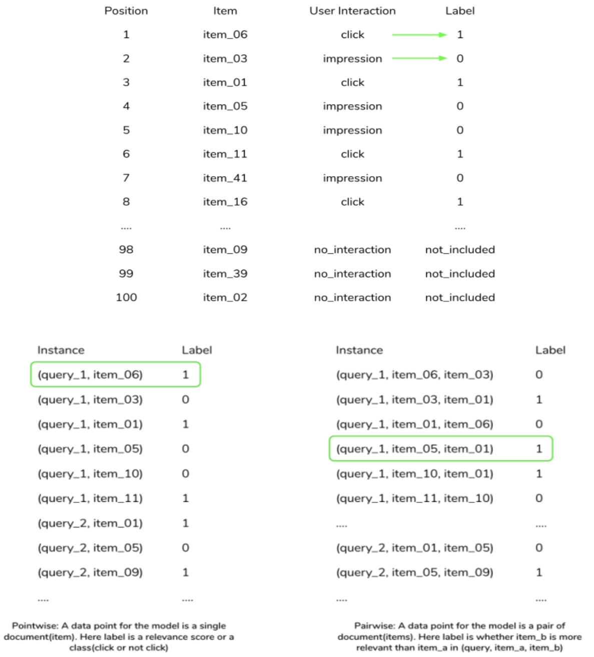
The listwise approach

In this approach all the research papers available online is grouped and it is associated with query q e.g y= {yj}

There are two types of output spaces used in the listwise approach.

Here the output space for the research paper can bedegree of all the papers which is attached with query [5].

The above approaches can be used for ranking but to be practical pair wise methods have been found better than pointwise methods because in pairwise method it mainly concentrates on determining the relative order which is similar to the ranking whereas determining the class label or relevance score independently [6].



Position	Item	User Interaction	Label
1	item_06	click	1
2	item_03	impression	0
3	item_01	click	1
4	item_05	impression	0
5	item_10	impression	0
6	item_11	click	1
7	item_41	impression	0
8	item_16	click	1
....	....		....
98	item_09	no_interaction	not_included
99	item_39	no_interaction	not_included
100	item_02	no_interaction	not_included

Instance	Label
(query_1, item_06)	1
(query_1, item_03)	0
(query_1, item_01)	1
(query_1, item_05)	0
(query_1, item_10)	0
(query_1, item_11)	1
(query_2, item_01)	1
(query_2, item_05)	0
(query_2, item_09)	1
....	....

Pointwise: A data point for the model is a single document(item). Here label is a relevance score or a class(click or not click)

Instance	Label
(query_1, item_06, item_03)	0
(query_1, item_03, item_01)	1
(query_1, item_01, item_06)	0
(query_1, item_05, item_01)	1
(query_1, item_10, item_01)	1
(query_1, item_11, item_10)	0
....	....
(query_2, item_01, item_05)	0
(query_2, item_05, item_09)	1
....	....

Pairwise: A data point for the model is a pair of document(items). Here label is whether item_b is more relevant than item_a in (query, item_a, item_b)

Fig 2: pairwise vs pointwise

IV. EXPERIMENTS AND RESULTS

Dataset

To create a training data set there are two ways .firstly we can use manual judgement in which a person is asked to choose a relevant document which meets the requirement of the user in getting the related research paper.

This can be done using the query which is given by the user taking the query context in to consideration dataset is generated and the human will judge the research papers. The main drawback of this approach is the generation of huge data set is very costly.

Due to the drawback of the first method we use second approach ,in this approach past search log data is used to determine the related labels. Now during searching for the relevant papers a user query in the search engine then the search engine the user can get many research papers out of which the user go through with few papers and interact with them. the next step after interactions is to convert them into labels called as relevance labels. These labels acts as a basis for the classification problem .later during prediction time this relevant labels are used for the ranking score. Here there is a possibility of getting multiple relevance labels in that case we can use the methods like Mcrank ,Prank to convert these multiple labels into ranking score.

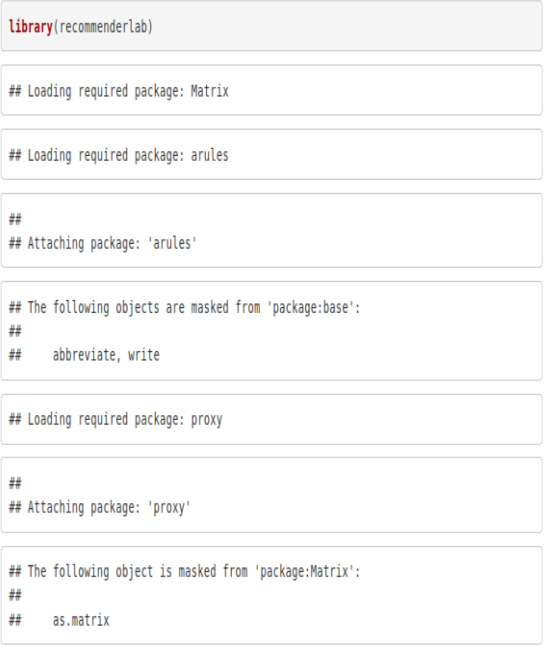
Importing Essential Libraries

In order to rank the research paper we have used four packages – *'rankpaper', 'gplot2', 'data1.researchtable' and 'reshape21'.*

Code:

library(recommenderlab)

Output Screenshot

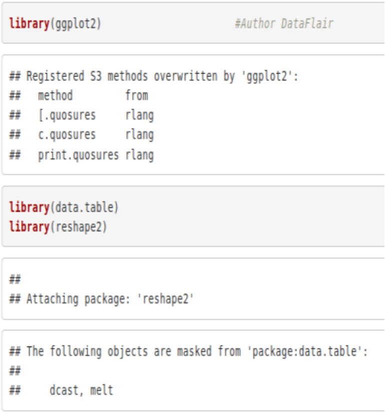


```
library(recommenderlab)

## Loading required package: Matrix

## Loading required package: arules

##
## Attaching package: 'arules'

## The following objects are masked from 'package:base':
##
##     abbreviate, write

## Loading required package: proxy

##
## Attaching package: 'proxy'

## The following object is masked from 'package:Matrix':
##
##     as.matrix
```

Code:

1. **library**(gplot2)
2. **library**(data1.table)
3. **library**(reshape21)

Output Screenshot:



```
library(ggplot2)                    #Author DataFlair

## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang

library(data.table)
library(reshape2)

##
## Attaching package: 'reshape2'

## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

Extracting the Data

The next step is to extract the data from the research papers datasets into ranking papers.we can use the str() function to display information about the research paper dataframes [8]

Code:

setwd("/home/ arXiv ")

arXiv _data <- read.csv("arXiv ",stringsAsFactors=FALSE)

rating_data <- read.csv("ratings.pdf")

str(rating_data)

Output Screenshot:

head(rating_data)

Data Pre-processing

Data preprocessing aims to extract only the related information which is used in the algorithms. In this phase the useful information for ranking the research papers is titles, year of publication and citations, author(s), conference of publication. In this section of the Learning to Rank , we will explore the most viewed Research papers in our dataset. We will first count the number of views for a given research paper and then organize them in a table that would group them in descending order.

Code:

1. library(gplot2)
2. rating_views <- Countofcol(ratingMatrix)
3. table_views <- data.frame(movie = names(rating_views),
4. views =rating_views) # create dataframe of views
5. table_views <- table_views[order(table_views$views,
6. decreasing = TRUE),] # sort by number of views
7. table_views$title <- NA
8. for (index in 1:10325){
9. table_views[index,3] <- as.character(subset(arXiv_data,
10. movie_data$paperId == table_views[index,1])$title)

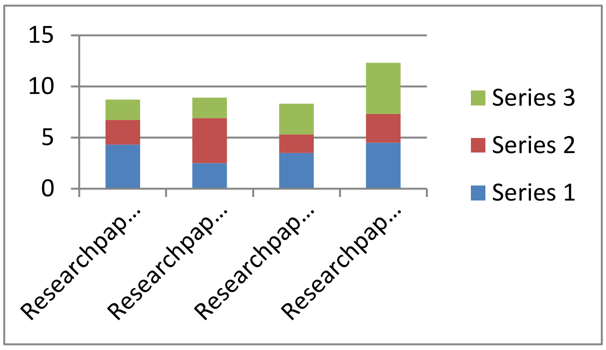


Fig. 3.RESEARCHPAPERS WITH VIEW COUNT

From the above bar-plot, we observe that researchpaper4 is viewed many times when compared to the other papers.

Data Preparation

We will conduct data preparation in the following three steps –

- Selecting useful data.
- Normalizing data.
- Binarizing the data.

For extracting the useful research paper in the dataset arXiv we have taken the maximum limit of the rating as 50.this is same as minimum number of views for a given research paper ratings per research paper. In this way the most relevant and popular research papers are rated than the least views papers.

For extracting the useful research paper in the dataset arXiv we have taken the maximum limit of the rating as 50.this is same as minimum number of views for a given research paper ratings per research paper. In this way the most relevant and popular research papers are rated than the least views papers.

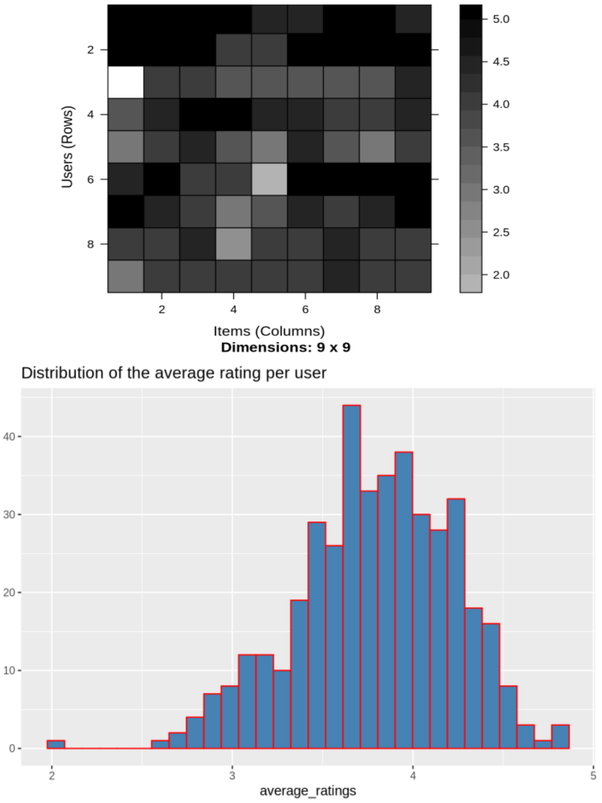


Fig 3:Heat map of the top research papers

Data Normalization

In the Data Normalization some scientific papers will have high ratings and some may have low ratings. In order to avoid this bias we follow the normalization which gives the consistency to the research papers. In this section we can standardize the values to some standard format to a common values. This can be helpful to avoid distortion to the ranges [9].

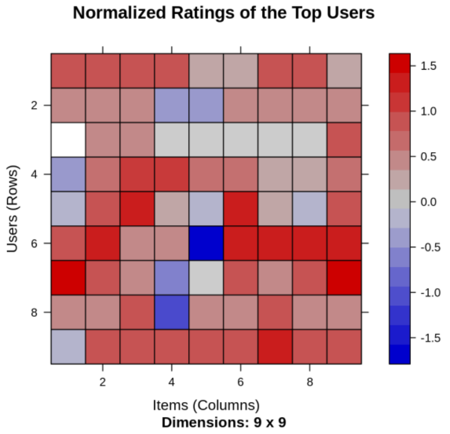


Fig 4: Data Normalization

Data Binarization

In the final step of data preparation we can binarize the data. This binarization will help the user to get the research paper very efficiently as it is allows only two values 0 or 1 so it is easy to decide the values easily [9].

Results

We have taken the arXiv dataset for applying the ranking using the algorithms discussed in the previous sections. arXiv contains 24,000 research papers from various areas of research . The main aim of this results is that how Discounted cumulative gain algorithm generate the better results in ranking the research paper .The arXiv dataset uses click through log data of a search algorithm .In the dataset for the experiment results collected 50 queries treated as training data set .In each phase there are 1000 web documents with clicks.

In order to have clear idea on labeling the click , a number is assigned to the clicks as 0 to 20 with 1 and 20 to 40 with click number 2 and so on.

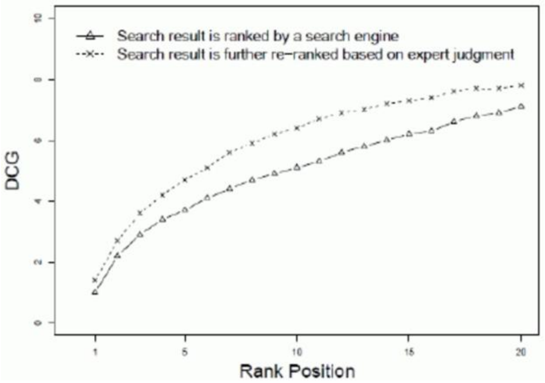


Fig 5: Creating labels from search log data

Table- I:Table of Ranking the Research Papers using ImpactFactor.

Ranking(Number published)	Journal	Impact factor
1(9)	International orthopaedics	2.387
2(7)	Journal of the Medical Association of Thailand	0.314
3(5)	Journal of Foot and Ankle Surgery	0.845
4(5)	Indian Journal of orthopaedics	0.64

V. CONCLUSION AND FUTUREWORK

Ranking plays a vital role in searching the relevant research paper as its helping the researcher in getting the valid data without time consuming .Machine learning is a emerging technology now a days where searching made easy for every area. so this research can be extended further using different methodologies using machine learning and Deep learning techniques.

REFERENCES

1. http://www.dirf.org/jdim/v4i1a12.pdf
2. http://www.iraj.in/journal/journal_file/journal_pdf /12-387-150487359549-56.pdf
3. http://www.iraj.in/journal/journal_file/journal_pdf/12-387-150487359549-56.pdf
4. https://github.com/arifqodari/ExpediaLearningToRank/tree/master/src
5. http://arogozhnikov.github.io/2015/06/26/learning-to-rank-software-datasets.html
6. https://tech.olx.com/ranking-ads-with-machine-learning-ee03d7734bf4
7. https://medium.com/@nikhilbd/intuitive-explanation-of-learning-to-rank-and-ranknet-lambdarank-and-lambdamart-fe1e17fac418
8. https://mail.google.com/mail/u/0/?tab=rm&ogbl#label/Research+papers/KtbxLrjCKpJRdbZLrDPptKcHpBKzBTnFkg?projector=1&messagePartId=0.1
9. https://mail.google.com/mail/u/0/?tab=rm&ogbl#label/Research+papers/KtbxLrjCKpJRdbZLrDPptKcHpBKzBTnFkg?projector=1&messagePartId=0.1

AUTHORS PROFILE

Sarabu Joshna , working as a Assistant Professor in the Department of Computer Science Engineering ,Bharath Institute of Engineering And Technology, Mangalpally, Ibrahimpatnam, Hyderabad. she published a paper in International Journal of Scientific Research and Review .Her area of research is machine learning .