A

Major Project

On

BUILDING SEARCH ENGINE USING ML

(Submitted in partial fulfillment of the requirements for the award of Degree)

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

CH.Vani(187R1A0513)

CH.Shravika(187R1A0511)

A.Chakradhar(187R1A0512)

Under the Guidance of

Dr. G.Madhukar

(Associate Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi)Recognized Under Section 2(f) & 12(B) of the UGCAct.1956,
Kandlakoya (V), Medchal Road, Hyderabad-501401.

2018-22

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled "**BUILDING SEARCH ENGINE USING ML**" being submitted by **CH.VANI (187R1A0513), CH.SHRAVIKA (187R1A0511) AND A.CHAKRADHAR (187R1A0512)** in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by him/her under our guidance and supervision during the year 2021-22.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dr. G. Madhukar
Associate Professor
INTERNAL GUIDE

Dr. A. Raji Reddy
DIRECTOR

Dr. K. Srujan Raju
HOD

EXTERNAL EXAMINER

Submitted for viva voice Examination held on _____

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express my profound gratitude and deep regard to my guide

We take this opportunity to express our gratitude and deep regards to our guide **Dr.G.Madukar,**Associate professor for his exemplary guidance monitoring and constant encouragement throughout the project work.The blessing,help and guidance given byhim shall carry us a long way in the journey of life on which we are about to embark. We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) **Mr. A. Uday Kiran, Mr.J.Narsimharao, Dr. T. S. Mastan RaoMr. A. Kiran Kumar, Mrs. G. Latha** for theircordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Srujan Raju,** Head of, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

We are obliged to **Dr. A. Raji Reddy,** Director for being cooperative throughout the course of this project. We also express our sincere gratitude to Sri. **Ch. Gopal Reddy,** Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus**who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion ofthis project.

CH. VANI (187R1A0513)

CH. SHRAVIKA (187R1A0511)

A. CHAKRADHAR (187R1A0512)

ABSTRACT

The web is the huge and most extravagant wellspring of data. To recover the information from the World Wide Web, Search Engines are commonly utilized. Search engines provide a simple interface for searching for user query and displaying results in the form of the web address of the relevant web page, but using traditional search engines has become very challenging to obtain suitable information. This paper proposed a search engine using Machine Learning technique that will give more relevant web pages at top for user queries .

LIST OF FIGURES/TABLES

LIST OF SCREENSHOTS

TABLE OF CONTENTS

1. INTRODUCTION

1.1.INTRODUCTION

World Wide Web is actually a web of individual systems and server which are connected with different technology and methods. Every site comprises the heaps of site pages that are being made and sent on the server. So if a user needs something, then he or she needs to type a keyword. Keyword is a set of words extracted from user search input. Search input given by a user may be syntactically incorrect. Here comes the actual need for search engines. Search engines provide you a simple interface to search user queries and display the results.

1) Web crawler : Web crawlers help in collecting data about a website and the links related to them. We are only using web crawlers for collecting data and information from WWW and storing it in our database.

2) Indexer : Indexer which arranges each term on each web page and stores the subsequent list of terms in a tremendous repository.

3) Query Engine : It is mainly used to reply to the user's keyword and show the effective outcome for their keyword. In the query engine, the Page ranking algorithm ranks the URL by using different algorithms in the query engine.

4)This paper utilizes Machine Learning Techniques to discover the utmost suitable web address for the given keyword. The output of the PageRank algorithm is given as input to the machine learning algorithm.

The first internet search engines predate the debut of the Web in December 1990: Who is user search dates back to 1982, and the Knowbot Information Service multi-network user search was first implemented in 1989.The first well documented search engine that searched content files, namely FTP files, was Archie, which debuted on 10 September 1990. Prior to September 1993, the World Wide Web was entirely indexed by hand. There was a list of webservers edited by Tim Berners-Lee and hosted on the CERN webserver. One snapshot of the list in 1992 remains, but as more and more web servers went online the central list could no longer keep up. On the NCSA site, new servers were announced under

the title "What's New!

The first tool used for searching content (as opposed to users) on the Internet was Archie. The name stands for "archive" without the "v"., It was created by Alan Emtage computer science student at McGill University in Montreal, Quebec, Canada. The program downloaded the directory listings of all the files located on public anonymous FTP (File Transfer Protocol) sites, creating a searchable database of file names; however, Archie Search Engine did not index the contents of these sites since the amount of data was so limited it could be readily searched manually.

In June 1993, Matthew Gray, then at MIT, produced what was probably the first web robot, the Perl-based World Wide Web Wanderer, and used it to generate an index called "Wandex". The purpose of the Wanderer was to measure the size of the World Wide Web, which it did until late 1995. The web's second search engine Aliweb appeared in November 1993. Aliweb did not use a web robot, but instead depended on being notified by website administrators of the existence at each site of an index file in a particular format.

1.2 EXISTING SYSTEM

To create a new Programmable Search Engine, all you have to do is choose which sites to search and give your search engine a name. In the Sites to search box, type one or more sites you want to include in the search results. You can include any sites on the web, even sites you don't own. By our calculations, for mid to large size retailer to build its own highquality, Solar-based site search engine would take 30 to 40 engineers as long as two years.140+ Search Engines and Directories. Search, the holy grail that pushed Google into global Internet domination, is still coveted by many. The fact that most users don't even consider switching Google for anything else doesn't mean that there's no innovation going on in the field of search

Disadvantages

Sometimes the search engine takes too much time to display relevant, valuable, and informative content.

Search engines, especially Google, frequently update their algorithm, and it is very difficult to find the algorithm in which Google runs.

2.LITERATURE SURVEY

2.1 INTRODUCTION

1) Weighted page rank algorithm based on in-out weight of webpages

AUTHORS: Kalyani Desikan, B. Jaganathan.

In its classical formulation, the well known page rank algorithm ranks web pages only based on in-links between web pages. We propose a new in-out weight based page rank algorithm. In this paper, we have introduced a new weight matrix based on both the in-links and out-links between web pages to compute the page ranks. We have illustrated the working of our algorithm using a web graph. We notice that the page rank values of the web pages computed using the original page rank algorithm and our proposed algorithm are comparable. Moreover, our algorithm is found to be efficient with respect to the time taken to compute the page rank values Once the target sentence has been selected, it can be used to generate questions targeting particular linguistic forms contained in the sentence. Heilman (2011) discusses the generation of factual, low-level questions suitable for beginner or intermediate students and gives a comprehensive overview of QG methods. Among the most prominent ones are: replacing the target form with a gap (Agarwal et al., 2011; Becker et al., 2012), applying transformation rules (Mitkov et al., 2006), filling templates (Curto et al., 2012), and generating all possible questions to a sentence and ranking them afterwards using a supervised learning algorithm (Heilman and Smith, 2009). Finally, QG is not an exception to the wave of neural networks, and Du et al. (2017) have recently approached automatic generation of reading comprehension questions on that basis. All of the mentioned QG systems either assess vocabulary or target reading comprehension, which contrasts with the focus of our work on functionally supporting focus on form in language learning.

2)Web Page Ranking Using Machine Learning Approach

AUTHORS: Junaid Khan, Arunima Jaiswal. One of the key components which ensures the

acceptance of web search service is the web page ranker - a component which is said to have been the main contributing factor to the early successes of Google. It is well established that a machine learning method such as the Graph Neural Network (GNN) is able to learn and estimate Google's page ranking algorithm. This paper shows that the GNN can successfully learn many other web page ranking methods e.g. Trust Rank, HITS and OPIC. Experimental results show that GNN may be suitable to learn any arbitrary web page ranking scheme, and hence, may be more flexible than any other existing web page ranking scheme. The significance of this observation lies in the fact that it is possible to learn ranking schemes for which no algorithmic solution exists or is known.

2.2. INPUT AND OUTPUT

i. INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur

OBJECTIVE

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

ii.OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements. 2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the

system.

2.3 FUNCTIONAL REQUIREMENTS

Outputs from computer systems are required primarily to communicate the results of processing to users. They are also used to provide a permanent copy of the results for later consultation. The various types of outputs in general are:

- External Outputs, whose destination is outside the organization,.
- Internal Outputs whose destination is within organization and they are the
- User's main interface with the computer.
- Operational outputs whose use is purely within the computer department.
- Interface outputs, which involve the user in communicating directly.
- Understanding user's preferences, expertise level and his business requirements through a friendly questionnaire.

2.4 PROPOSED SYSTEM

To build a search engine which gives web address of the most relevant web page at the top of the search result, according to user queries. The main focus of our system is to build a search engine using machine learning technique for increasing accuracy compare to available search engine. Following is the step by step procedure for building the search engine:

1. Collect data from WWW using web crawler.
2. Perform data cleaning using NLP.
3. Study and compare the existing page ranking algorithm.
4. Merge the selected page rank algorithm with current technologies in ML.
5. Implement query engine to display the efficient results for user query.

Advantages

- A search engine saves you time in two ways: by eliminating the need to find information manually, and by performing searches at high speeds.
- Relevance: When a search engine scans a website, it scores the content for relevance to particular search words.

3.SYSTEM ANALYSIS

FEASIBILITY STUDY

Preliminary investigation examine project feasibility, the likelihood the system will be useful to the organization. The main objective of the feasibility study is to test the Technical, Operational and Economical feasibility for adding new modules and debugging old running system. All system is feasible if they are unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigation:

- Technical Feasibility
- Economic Feasibility
- Social Feasibility

3.1. 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.

Here in this application used the technologies like Visual Studio 2012 and SqlServer 2014.

3.2. 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.

3.3.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 efficiently. 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 9 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 performance.

4.IMPLEMENTATION

4.1 SOFTWARE DEVELOPMENT LIFE CYCLE

There is various software development approaches defined and designed which are used/employed during development process of software, these approaches are also referred as "Software Development Process Models". Each process model follows a particular life cycle in order to ensure success in process of software development.

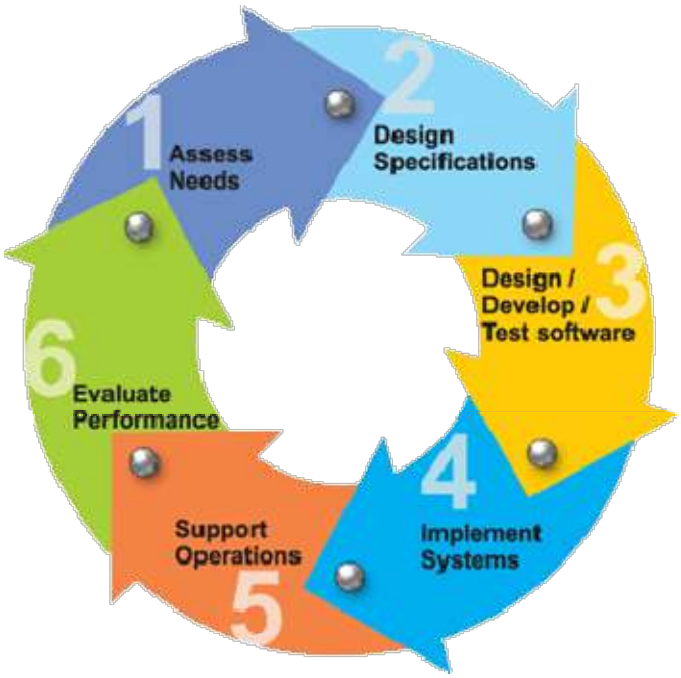


Fig. 4.1 SDLC Life Cycle

Requirements

Business requirements are gathered in this phase. This phase is the main focus of the project managers and stake holders. Meetings with managers, stake holders and users are held in order to determine the requirements. Who is going to use the system? How will they use the system? What data should be input into the system? What data should be output by the system? These are general questions that get answered during a requirements gathering phase.

This produces a nice big list of functionality that the system should provide, which describes functions the system should perform, business logic that processes data, what data is stored and used by the system, and how the user interface should work. The overall result is the system as a whole and how it performs, not how it is actually going to do it.

Design

The software system design is produced from the results of the requirements phase. Architects have the ball in their court during this phase and this is the phase in which their focus lies. This is where the details on how the system will work is produced. Architecture, including hardware and software, communication, software design (UML is produced here) are all part of the deliverables of a design phase.

Implementation

Code is produced from the deliverables of the design phase during implementation, and this is the longest phase of the software development life cycle. For a developer, this is the main focus of the life cycle because this is where the code is produced. Implementation my overlap with both the design and testing phases. Many tools exists (CASE tools) to actually automate the production of code using information gathered and produced during the design phase.

Testing

During testing, the implementation is tested against the requirements to make sure that the product is actually solving the needs addressed and gathered during the requirements phase. Unit tests and system/acceptance tests are done during this phase. Unit tests act on a specific component of the system, while system tests act on the

system as a whole.cycle model. Now let's delve into some of the traditional and widely used variations.

4.2 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 of the project, with an eye toward the end goal of the project.

The following diagram shows how a spiral model acts like:

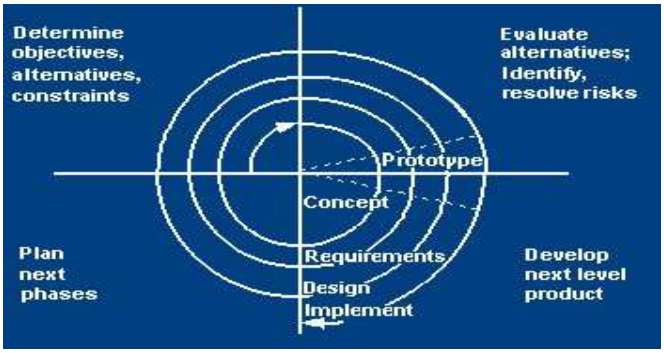


Fig.4.2 Spiral Model

The steps for Spiral Model can be generalized as follows

- 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.
- A preliminary design is created for the new system.
- 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.
- A second prototype is evolved by a fourfold procedure:

1. Evaluating the first prototype in terms of its strengths, weakness, and risks.
2. Defining the requirements of the second prototype.
3. Planning a designing the second prototype.
4. Constructing and testing the second prototype.

- The preceding steps are iterated until the customer is satisfied as the refined prototype.
- The final system is constructed, based on the refined prototype

5.SOFTWARE REQUIREMENT SPECIFICATION

5.1 REQUIREMENTS SPECIFICATION

Requirement Specification provides a high secure storage to the web server efficiently. Software requirements deal with software and hardware resources that need to be installed on a serve which provides optimal functioning for the application. These software and hardware requirements need to be installed before the packages are installed. These are the most common set of requirements defined by any operation system. These software and hardware requirements provide a compatible support to the operation system in developing an application.

i. HARDWARE REQUIREMENTS

The hardware requirement specifies each interface of the software elements and the hardware elements of the system. These hardware requirements include configuration characteristics.

System : Pentium IV 2.4 GHz.

Hard Dis : 100 GB.

Monitor : 15 VGA Color.

Mouse : Logitech.

RAM : 1 GB.

ii. SOFTWARE REQUIREMENTS:

The software requirements specify the use of all required software products like data management system. The required software product specifies the numbers and version. Each interface specifies the purpose of the interfacing software as related to this software product.

Operating system : Windows 10

Coding Language : Html, JavaScript,

Development Kit : Anaconda prompt

Programming language : python

5.2 FUNCTIONAL REQUIREMENTS

The functional requirement refers to the system needs in an exceedingly computer code engineering method.

The key goal of determinant "functional requirements" in an exceedingly product style and implementation is to capture the desired behavior of a software package in terms of practicality and also the technology implementation of the business processes.

5.3 NON -FUNCTIONAL REQUIREMENTS

All the other requirements which do not form a part of the above specification are categorized as Non-Functional needs. A system perhaps needed to gift the user with a show of the quantity of records during info. If the quantity must be updated in real time, the system architects should make sure that the system is capable of change the displayed record count at intervals associate tolerably short interval of the quantity of records dynamic. Comfortable network information measure may additionally be a non-functional requirement of a system.

The following are the features:

- Availability
- Backup
- Certification
- Compliance
- Configuration Management
- Documentation
- Disaster Recovery
- Efficiency
- Interoperability

5.4 PERFORMANCE REQUIREMENTS

Performance is measured in terms of the output provided by the application. Requirement specification plays an important part in the analysis of a system. Only when the requirement specifications are properly given, it is possible to design a system, which will fit into required environment. It rests largely with the users of the existing system to give the requirement specifications because they are the people who finally use the system. This is because the requirements have to be known during the initial stages so that the system can be designed according to those requirements. It is very difficult to change the system once it has been designed and on the other hand designing a system, which does not cater to the requirements of the user, is of no use.

The requirement specification for any system can be broadly stated as given below:

- The system should be able to interface with the existing system
- The system should be accurate.

6.SYSTEM DESIGN

6.1 SYSTEM ARCHITECTURE

- Design various blocks for overall system processes.
- Design compact,smaller and workable modules in each block.
- Design various database structures.
- Specify details of programs to achieve desired functionality.
- Design the form of inputs,and outputs of the system,
- Perform documentation of the design.
- System reviews.

The purpose of the design phase is to arrange an answer of the matter such as by the necessity document. This part is that the opening moves in moving the matter domain to the answer domain. The design phase satisfies the requirements of the system. The design of a system is probably the foremost crucial issue warm heartedness the standard of the software package. It's a serious impact on the later part, notably testing and maintenance.

The output of this part is that the style of the document. This document is analogous to a blueprint of answer and is employed later throughout implementation, testing and maintenance. The design activity is commonly divided into 2 separate phases System Design and Detailed Design.

System Design conjointly referred to as top-ranking style aims to spot the modules that ought to be within the system, the specifications of those modules, and the way them move with one another to supply the specified results.

At the top of the system style all the main knowledge structures, file formats, output formats, and also the major modules within the system and their specifications square measure set. System design is that the method or art of process the design, components, modules, interfaces, and knowledge for a system to satisfy such as needs,detailed design, the inner logic of every of the modules laid out in system design is determined. Throughout this part, the small print of the info of a module square measure sometimes laid out in a high-level style description language that is freelance of the target language within which the software package can eventually be enforced.

In system design the main target is on distinguishing the modules, whereas throughout careful style the main target is on planning the logic for every of the modules.

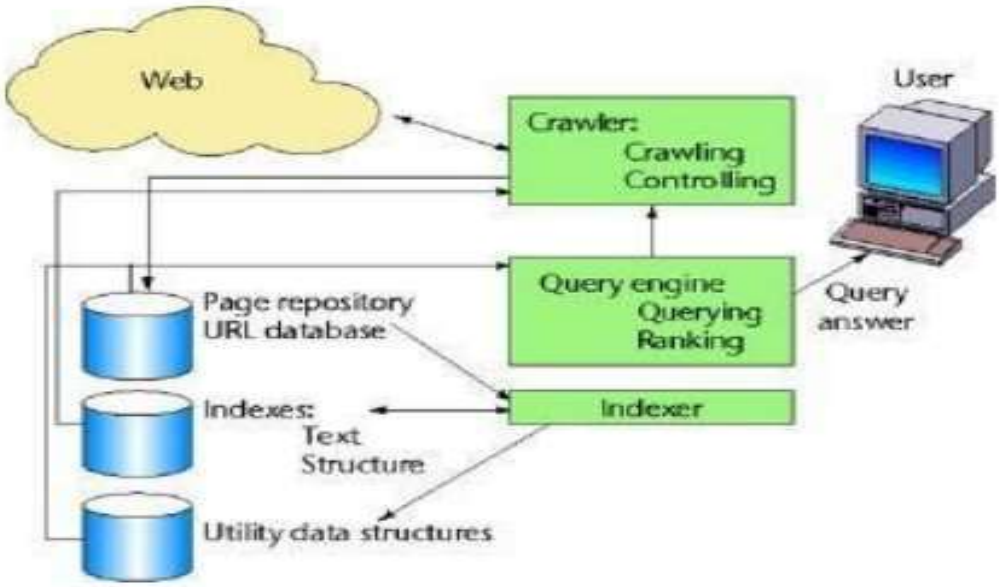


Figure 6.1: Architecture diagram

6.2 UML DIAGRAMS

The Unified Modeling Language allows the software engineer to express an analysis model using the modeling notation that is governed by a set of syntactic semantic and pragmatic rules.

A UML system is represented using five different views that describe the system from distinctly different perspective. Each view is defined by a set of diagram, which is as follows.

User Model View

This view represents the system from the users perspective. The analysis representation describes a usage scenario from the end-users perspective.

Structural Model view

In this model the data and functionality are arrived from inside the system. This model view models the static structures.

Behavioral Model View

It represents the dynamic of behavioral as parts of the system, depicting the interactions of collection between various structural elements described in the user model and structural model view.

Implementation Model View

In this the structural and behavioral as parts of the system are represented as they are to be built.

i. USE CASE DIAGRAM

A use case diagram at its simplest is a representation of a user's interaction with the system and depicting the specifications of a use case. A use case diagram can portray the different types of users of a system and the various ways that they interact with the system. This type of diagram is typically used in conjunction with the textual use case and will often be accompanied by other types of diagrams as well.

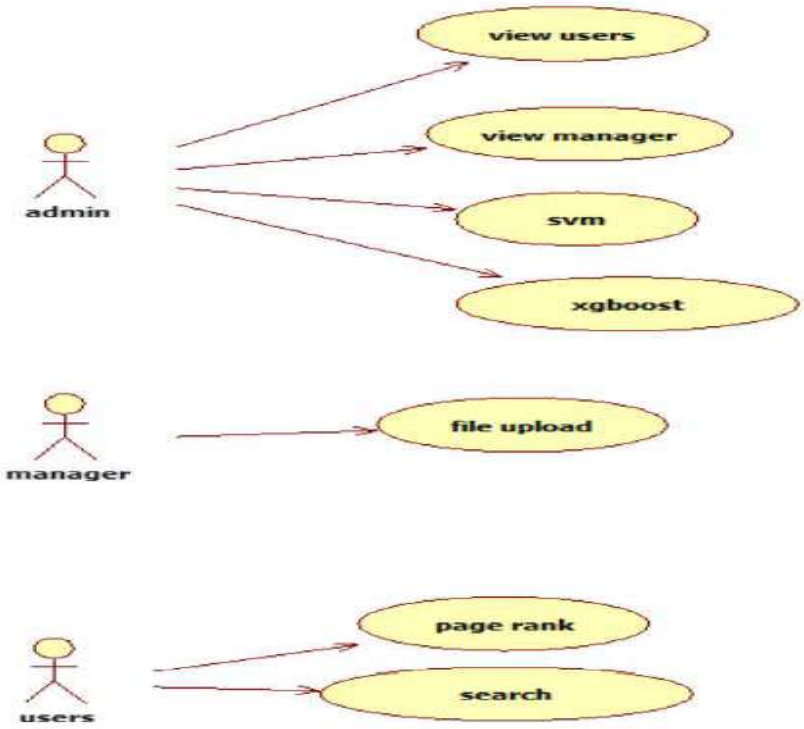


Figure 6.2 : Use Case Diagram

ii. CLASS DIAGRAM

The class diagram is the main building block of object oriented modeling. It is used both for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. Class diagrams can also be used for data modeling. The classes in a class diagram represent both the main objects, interactions in the application and the classes to be programmed. A class with three sections, in the diagram, classes is represented with boxes which contain three parts

The upper part holds the name of the class.

The middle part contains the attributes of the class.

The bottom part gives the methods or operations the class can take or undertake.

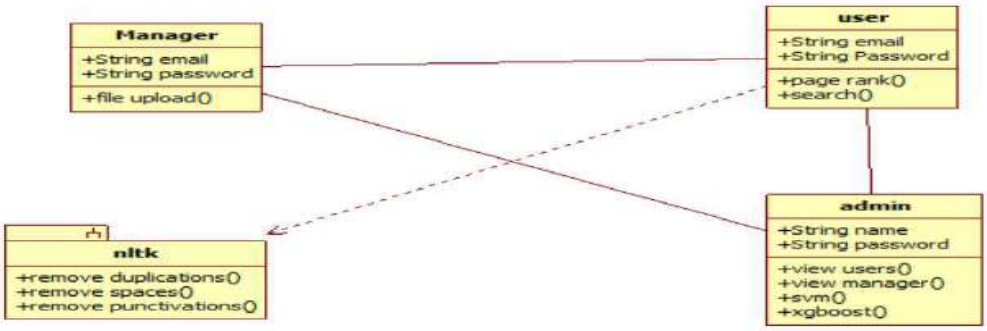


Figure 6.3: Class Diagram.

iii. SEQUENCE DIAGRAM

A sequence diagram is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams and timing diagrams.

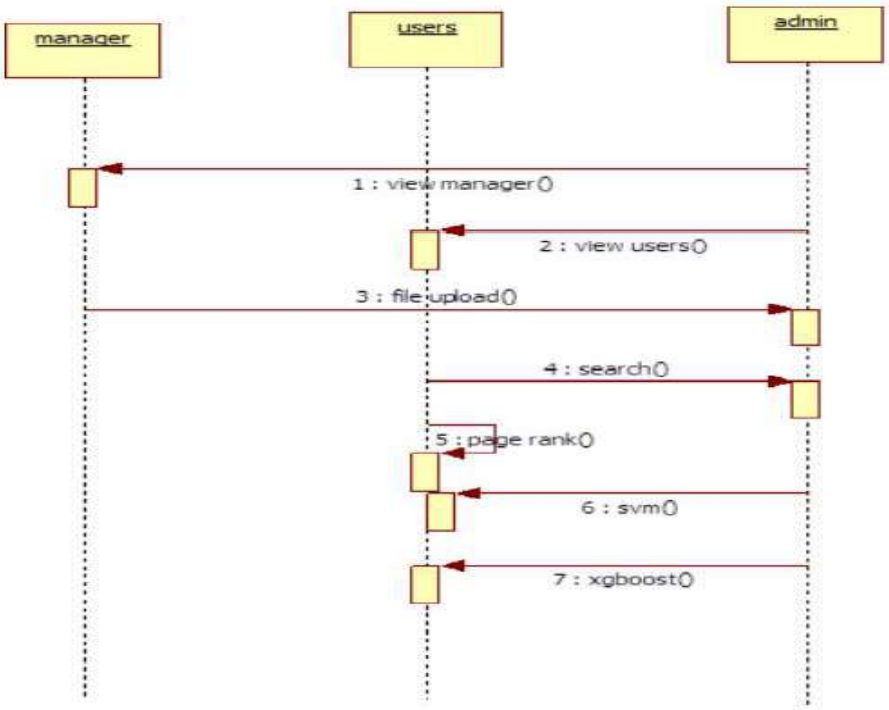


FIGURE 6.4: Sequence diagram

iv. ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

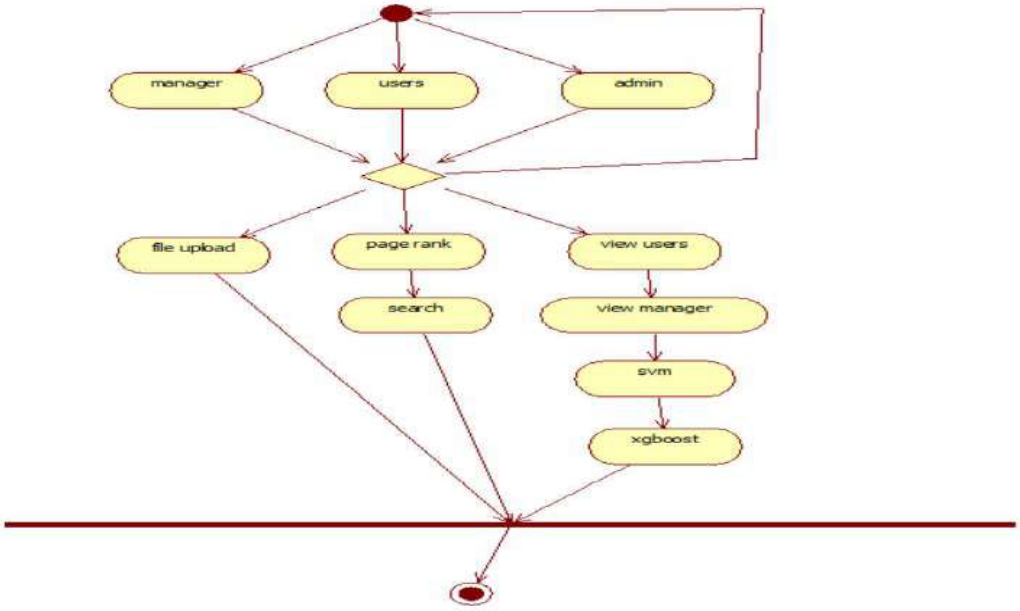


FIGURE 6.5:Activity Diagram

7.TESTING

7.1 SYSTEM TESTING

Testing has become an integral part of any system or project especially in the field of information technology. The importance of testing is a method of justifying, if one is ready to move further, be it to be check if one is capable to with stand the rigors of a particular situation cannot be underplayed and that is why testing before development is so critical. When the software is developed before it is given to user to user the software must be tested whether it is solving the purpose for which it is developed. This testing involves various types through which one can ensure the software is reliable. The program was tested logically and pattern of execution of the program for a set of data are repeated. Thus the code was exhaustively checked for all possible correct data and the outcomes were also checked.

7.2 MODULE TESTING

To locate errors, each module is tested individually. This enables us to detect error and correct it without affecting any other modules. Whenever the program is not satisfying the required function, it must be corrected to get the required result. Thus all the modules are individually tested from bottom up starting with the smallest and lowest modules and proceeding to the next level. Each module in the system is tested separately. For example the job classification module is tested separately. This module is tested with different job and its approximate execution time and the result of the test is compared with the results that are prepared manually. Each module in the system is tested separately. In this system the resource classification and job scheduling modules are tested separately and their corresponding results are obtained which reduces the process waiting time.

7.3 INTEGRATION TESTING

After the module testing, the integration testing is applied. When linking the modules there may be chance for errors to occur, these errors are corrected by using this testing. In this system all modules are connected and tested. The testing results are very correct. Thus the mapping of jobs with resources is done correctly by the system .

7.4 ACCEPTANCE TESTING

When that user fined no major problems with its accuracy, the system passers through a final acceptance test. This test confirms that the system needs the original goals, objectives and requirements established during analysis without actual execution which elimination wastage of time and money acceptance tests on the shoulders of users and management, it is finally acceptable and ready for the operation.

7.5 TEST CASES

S.no	Test Case	Expected Result	Result	Remark (IFFails)
1.	User Register	If User registration success fully.	Pass	If already user email exists then it fails.
2.	User Login	If the Username and password is correct then it willbe a valid page.	Pass	Unknown Register Users willnot be logged in.
3.	Manager login	If the Manager name and password is correct then it willbe a valid pag.	Pass	.Unknown Register Managerwill not log in.
4.	Admin can activate the register managers	Admin can activate the register managerid.	Pass	If the manager did not find it then it won't login
5.	Admin login	Admin can loginwith his login credential. If success he get ishome page	Pass	Invalid login details will not allowed here
6.	Admin can activate the register users	Admin can activate the register user id .	Pass	.If the user did notfind it then it won't login.

7.	admin can get thesvm results	by clicking svm itwill display svm prediction	Pass	prediction of svm won't get..
8.	admin can get the xgboost results	by clicking xgboost it will display xgboost prediction.	Pass	predictio n of xgboost won't get..
9.	user login page	user can search the weig ht of parti cular docu ment	Pass	we won't get theweight of document.
10.	Query search	Can search using search button	Pass	Won't get the weight

Table 7.1: Test Cases

8.SOURCE CODE

SOURCE CODE

```
 path('userlogin/',user.userlogin,name='userlogin'),
path('userregister/',user.userregister,name='userregister'),
path('userlogincheck/',user.userlogincheck,name='userlogincheck'),
path('pagerank',user.pagerank,name='pagerank'),
path('search/',user.search, name="search"),
path('search1/',user.search1, name="search1"),
path('usersearchresult/',user.usersearchresult, name="usersearchresult"),
path('usersearchresult1/',user.usersearchresult1, name="usersearchresult1"),
path('weight/', user.weight, name="weight"),
path('logout/',user.logout,name='logout'),

path('managerlogin/',manager.managerlogin,name='managerlogin'),
path('managerregister/',manager.managerregister,name='managerregister'),
path('managerlogincheck/',manager.managerlogincheck,name='managerlogincheck'),
path('fileupload/', manager.fileupload, name='fileupload')

path('admin1/',search.adminlogin,name='admin1'),
path('adminloginentered/',search.adminloginentered,name='adminloginentered'),
path('userdetails/',search.userdetails,name='userdetails'),
path('Managerdetails/',search.managerdetails,name='Managerdetails'),
path('activateuser/',search.activateuser,name='activateuser'),
path('activatemanager/',search.activatemanager,name='activatemanager'),
```

views.py:

```
def managerlogin(request):
return render(request,'manager/managerlogin.html')
def          managerregister(request):
   if request.method=='POST':
      form1=managerForm(request.POS
      T)if form1.is_valid():
```

```
            form1.save()
            print("succesfully saved the data")
            return render(request, 'manager/managerlogin.html')
            #return HttpResponse("registreration succesfully completed")
        else:
            print("form not valied")
            return HttpResponse("form not valied")
    Else

form=managerForm()
        return render(request,"manager/managerregister.html",{"form":form})
def managerlogincheck(request):
    if request.method == 'POST':
        sname =
        request.POST.get('email')
        print(sname)
        spasswd =
        request.POST.get('upasswd')
        print(spasswd)
        try:
            check = managerModel.objects.get(email=sname,passwd=spasswd)
            # print('usid',usid,'pswd',pswd)
            print(check)
            # request.session['name'] =
            check.name#
            print("name",check.name)
            status =
            check.status
            print('status',stat
            us)
            if status == "Activated":
                request.session['email'] = check.email
                return render(request,
```

```
            'manager/managerpage.html')else:
                messages.success(request, 'manager is not activated')
                return render(request, 'manager/managerlogin.html')
        except Exception as e:
            print('Exception is
            ',str(e))pass
        messages.success(request,'Invalid name and password')
        return render(request,'manager/managerlogin.html')
```

models.py:

```
from django.db import models
class userModel(models.Model):
    name =
    models.CharField(max_length=50)email
    = models.EmailField()
    passwd = models.CharField(max_length=40)
    cwpasswd = models.CharField(max_length=40)
    mobileno = models.CharField(max_length=50, default="", editable=True)
    status = models.CharField(max_length=40, default="", editable=True)

    def str (self):
        return
        self.email

    class Meta:
        db_table='userregi
        ster'
class weightmodel(models.Model):
    filename =
    models.CharField(max_length=100)file =
    models.FileField(upload_to='files/pdfs/')
    weight=models.CharField(max_length=100)
    rank=models.CharField(max_length=100,default="", editable=False)
```

```
    label=models.CharField(max_length=100,default="", editable=False)


    def__str__(self):
        return
        self.filename
    class Meta:
        db_table='we
        ight'
from django.db import models
class userModel(models.Model):
    name =
    models.CharField(max_length=50)email
    = models.EmailField()
    passwd = models.CharField(max_length=40)
    cwpasswd = models.CharField(max_length=40)
    mobileno = models.CharField(max_length=50, default="", editable=True)
    status = models.CharField(max_length=40, default="", editable=True)


    def str (self):
        return
        self.email


    class Meta:
        db_table='userregi
        ster'
class weightmodel(models.Model):
    filename =
    models.CharField(max_length=100)file =
    models.FileField(upload_to='files/pdfs/')
    weight=models.CharField(max_length=100)
    rank=models.CharField(max_length=100,default="", editable=False)
    label=models.CharField(max_length=100,default="", editable=False)
```

```
    def__str__(self):
        return
        self.filename
    class Meta:
        db_table='we
        ight'
```

forms.py:

```
from django import
forms from
user.models import *
from django.core import validators


class userForm(forms.ModelForm):
    name = forms.CharField(widget=forms.TextInput(), required=True, max_length=100,)
    passwd = forms.CharField(widget=forms.PasswordInput(), required=True,
max_length=100)
    cwpasswd = forms.CharField(widget=forms.PasswordInput(), required=True,
max_length=100)
    email = forms.CharField(widget=forms.TextInput(),required=True)
    mobileno= forms.CharField(widget=forms.TextInput(), required=True,
max_length=10,validators=[validators.MaxLengthValidator(10),validators.MinLengthValidat
or(10)])
    status = forms.CharField(widget=forms.HiddenInput(), initial='waiting', max_length=100)


    def str (self):
        return
        self.email


    class Meta:
        model=userM
        odel
```

```
        fields=['name','passwd','cwpasswd','email','mobileno','status']
```

userdetails.html:

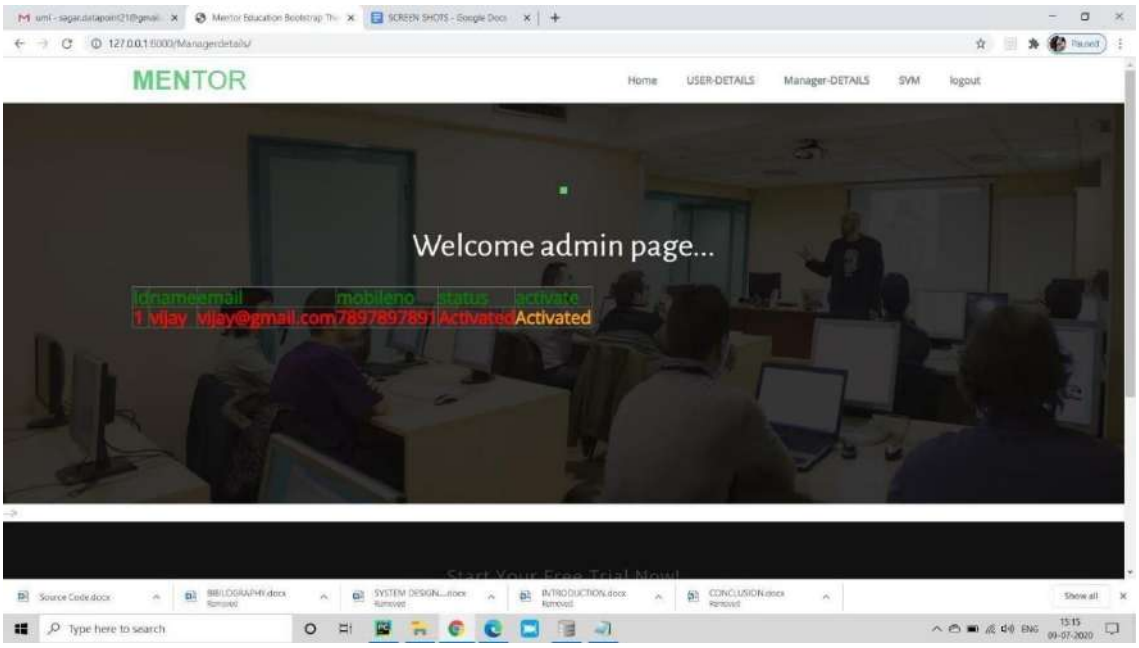
```
{% extends 'adminbase.html' %}

{% load static %}
{% block contents^}

<div class="modal fade" id="login" role="dialog">

   <div class="modal-dialog modal-sm">

     <!-- Modal content no 1-->

     <div class="modal-content">

       <div class="modal-header">

         <button type="button" class="close" data-dismiss="modal">&times;</button>

         <h4 class="modal-title text-center form-title">Login</h4>

       </div>

       <div class="modal-body padtrbl">
         <div class="login-box-body">

           <p class="login-box-msg">Sign in to start your session</p>

           <div class="form-group">

             <form name="" id="loginForm">

               <div class="form-group has-feedback">

                 <!____username_____________>

                 <input class="form-control" placeholder="Username" id="loginid" type="text"
autocomplete="off" />

                 <span style="display:none;font-weight:bold; position:absolute;color: red;position:
absolute;padding:4px;font-size: 11px;background-color:rgba(128, 128, 128, 0.26);z-index: 17;
right: 27px; top: 5px;" id="span_loginid"></span>

                 <!---Alredy exists ! -->

                 <span class="glyphicon glyphicon-user form-control-feedback"></span>

               </div>

               <div class="form-group has-feedback">

                 <!____password_____________>

                 <input class="form-control" placeholder="Password" id="loginpsw"
type="password" autocomplete="off" />
```

```
                <span style="display:none;font-weight:bold; position:absolute;color:
grey;position: absolute;padding:4px;font-size: 11px;background-color:rgba(128, 128, 128,
0.26);z-index: 17; right: 27px; top: 5px;" id="span_loginpsw"></span>
                <!---Alredy exists ! -->

                <span class="glyphicon glyphicon-lock form-control-feedback"></span>
              </div>
              <div class="row">
                <div class="col-xs-12">
                  <div class="checkbox icheck">
                    <label>
                          <input type="checkbox" id="loginrem" > Remember Me
                        </label>
                  </div>
                </div>
                <div class="col-xs-12">
                  <button type="button" class="btn btn-green btn-block btn-flat"
onclick="userlogin()">Sign In</button>
                </div>
              </div>
            </form>
          </div>
        </div>
      </div>
    </div>


  </div>
</div>
<!--/ Modal box-->
<!--Banner-->
<div class="banner">
  <div class="bg-color">
    <div class="container">
      <div class="row">
```

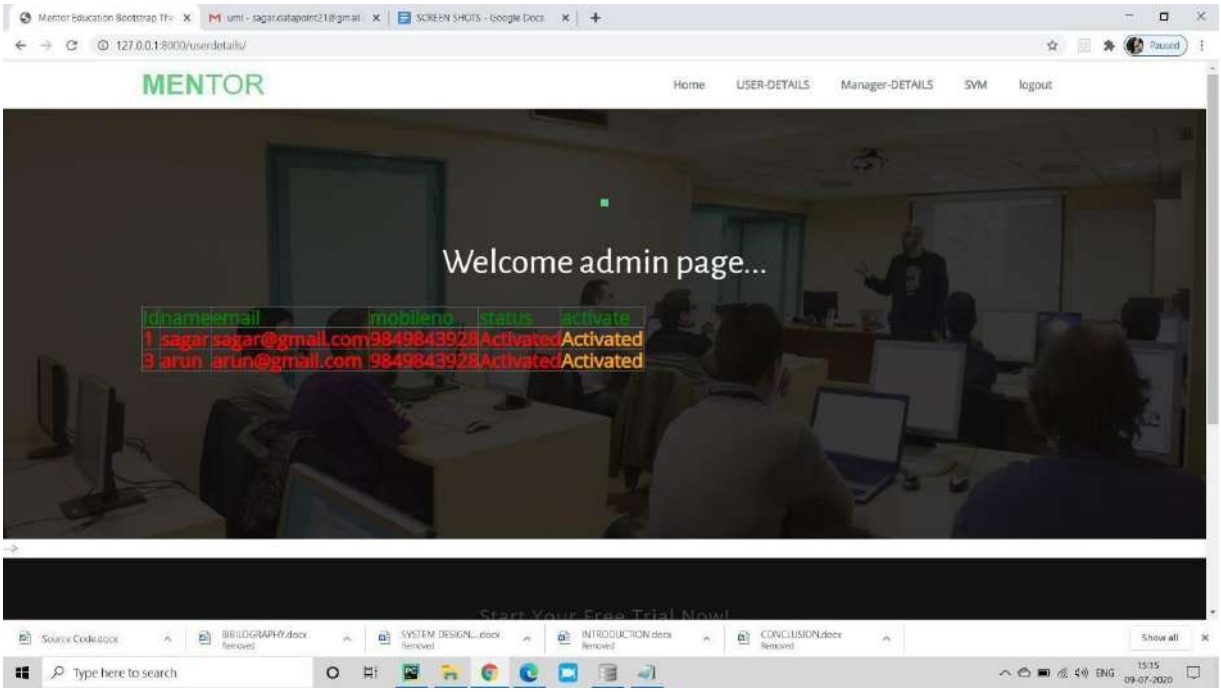
```
<div class="banner-text text-center">
<div class="text-border">
<!--<h2 class="text-dec"></h2>-->
</div>
<div class="intro-para text-center quote">
<p>
Welcome admin page...
</p>
<center><h3>
<table border="2px solid red" align="left">
<tr><th style="color:green">Id</th>
<th style="color:green">name</th>
<th style="color:green">email</th>
<th style="color:green">mobileno</th>
<th style="color:green">status</th>
<th style="color:green">activate</th>
</tr>
{% for x in qs %}
<tr>
<td style="color:red">{{x.id}}</td>
<td style="color:red">{{x.name}}</td>
<td style="color:red">{{x.email}}</td>
<td style="color:red">{{x.mobileno}}</td>
<td style="color:red">{{x.status}}</td>
>Activate</a></td>
{% if x.status == 'waiting' %}
<td style="color:orange"> <a href="/activateuser/?pid={{ x.id }}"
```

```
{% else %}
        <td style="color:orange"> Activated</td>

                                    {% endif %}

                                </tr>
                                {% endfor %}

                            </table>
                    </h3>
                  </center>
                </div>

                </a>
              </div>
            </div>
          </div>
        </div>
      </div>-->
    {% endblock %}
```

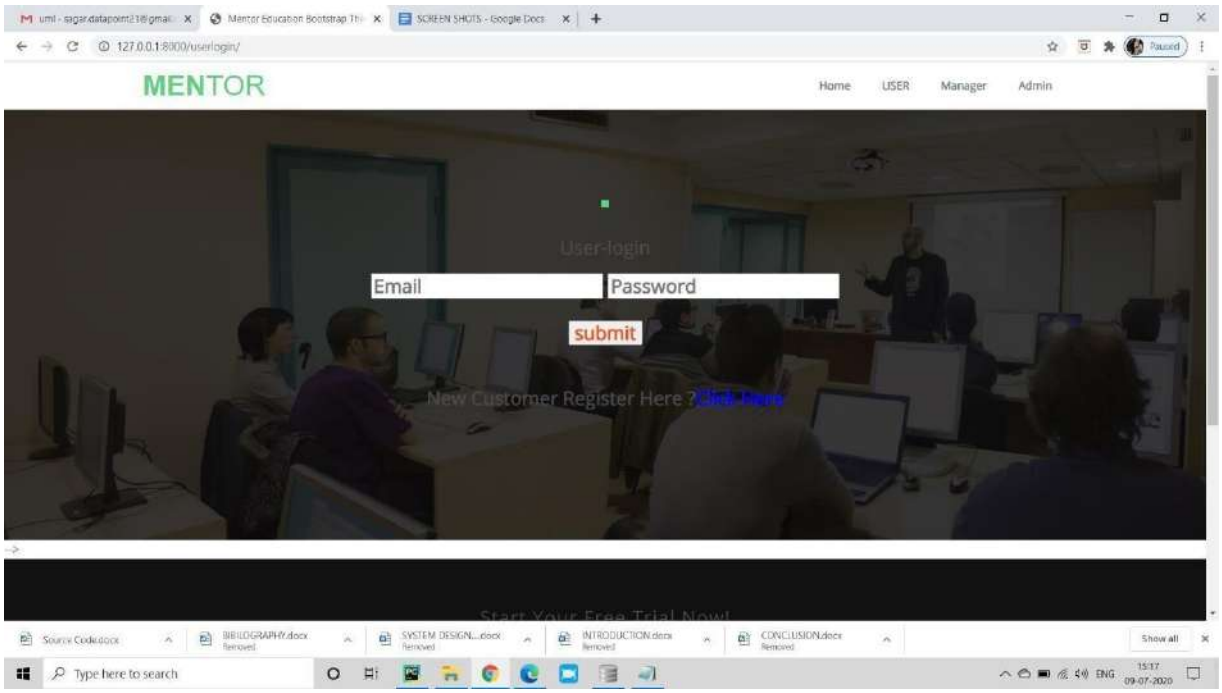
9.SCREENSHOTS



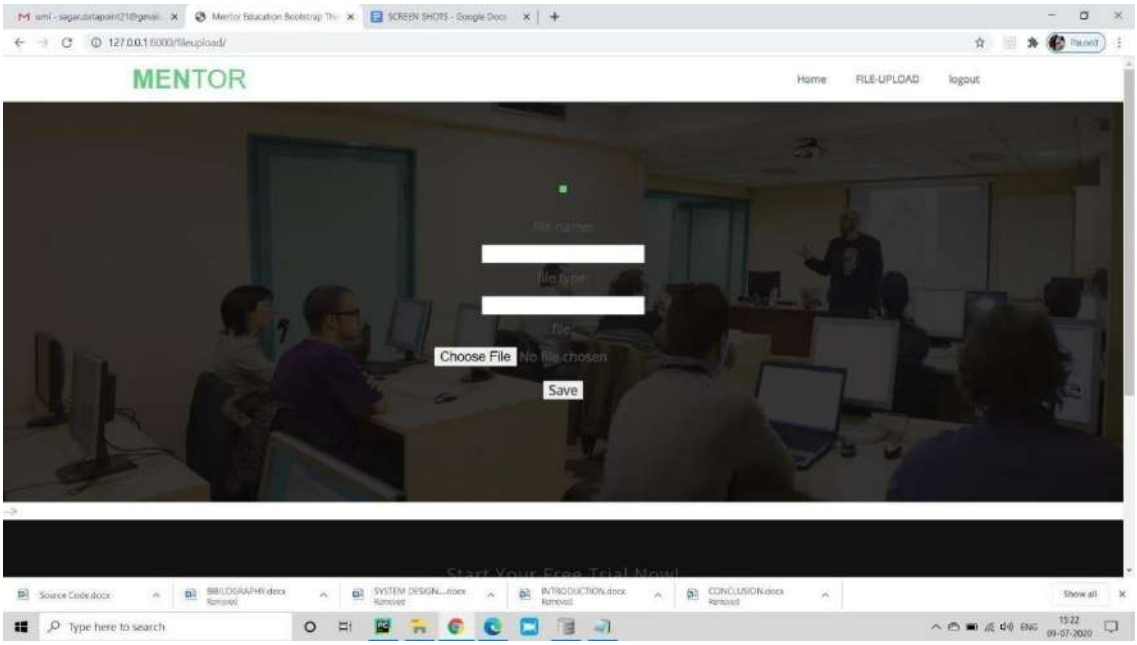
SCREENSHOT 1 MANAGER DETAILS



SCREENSHOT 2: USER DETAILS



SCREENSHOT 3: USER LOGIN



SCREENSHOT 4: FILE UPLOAD

10.CONCLUSION

CONCLUSION

Search engines are very useful for finding out more relevant URLs for given keywords. Dueto this, user time is reduced for searching the relevant web page. For this, Accuracy is a very important factor. From the above observation, it can be concluded that XGBoost is better in terms of accuracy than SVM and ANN. Thus, Search engines built using XGBoost and PageRank algorithms will give better accuracy.

11.REFERENCES

REFERENCES

[1] Manika Dutta, K. L. Bansal, "A Review Paper on Various Search Engines (Google, Yahoo, Altavista, Ask and Bing)", International Journal on Recent and Innovation Trends in Computing and Communication, 2016.

[2] Gunjan H. Agre, Nikita V.Mahajan, "Keyword Focused Web Crawler", International Conference on Electronic and Communication Systems, IEEE, 2015.

[3] Tuhena Sen, Dev Kumar Chaudhary, "Contrastive Study of Simple PageRank, HITS and Weighted PageRank Algorithms: Review", International Conference on Cloud Computing, Data Science & Engineering, IEEE,2017.

[4] Michael Chau, Hsinchun Chen, "A machine learning approach to web page filtering using content and structure analysis", Decision Support Systems 44 (2008) 482–494,scienceDirect,2008.

[5] Taruna Kumari, Ashlesha Gupta, Ashutosh Dixit, "Comparative Study of Page Rank and Weighted Page Rank Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, February 2014.

[6] K. R. Srinath, "Page Ranking Algorithms – A Comparison", International Research Journal of Engineering and Technology (IRJET), Dec2017.

[7] S. Prabha, K. Duraiswamy, J. Indhumathi, "Comparative Analysis of Different Page Ranking Algorithms", International Journal of Computer and Information Engineering, 2014.

[8] Dilip Kumar Sharma, A. K. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", International Journal on Computer Science and Engineering, 2010.

[9] Vijay Chauhan, Arunima Jaiswal, Junaid Khalid Khan, “Web Page Ranking Using Machine Learning Approach”, International Conference on Advanced Computing Communication Technologies, 2015.

[10] Amanjot Kaur Sandhu, Tiewei s. Liu., “Wikipedia Search Engine: Interactive Information Retrieval Interface Design”, International Conference on Industrial and Information Systems, 2014.

[11] Neha Sharma, Rashi Agarwal, Narendra Kohli, “Review of features and machine learning techniques for web searching”, International Conference on Advanced Computing Communication Technologies, 2016.

[12] Sweah Liang Yong, Markus Hagenbuchner, Ah Chung Tsoi, “Ranking Web Pages using Machine Learning Approaches”, International Conference on Web Intelligence and IntelligentAgent Technology, 2008.

[13] B. Jaganathan, Kalyani Desikan,“Weighted Page Rank Algorithm based on In-OutWeight of Webpages”, Indian Journal of Science and Technology, Dec

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Building Search Engine using Machine Learning

Vani Chinni, Shravika Challa, Chakradhar Annandi

CH.VANI (CSE,CMR Technical Campus)
CH.SHRAVIKA(CSE,CMR Technical Campus)
A.CHAKRADHAR(CSE,CMR Technical Campus)

Abstract –

The web is the huge and most extravagant wellspring of data. To recover the information from the World Wide Web, Search Engines are commonly utilized. Search engines provide a simple interface for searching for user query and displaying results in the form of the web address of the relevant web page, but using traditional search engines has become very challenging to obtain suitable information. This paper proposed a search engine using Machine Learning technique that will give more relevant web pages at top for user queries.

Key Words: web crawling, Indexer, Search Engine,Webpages.

1.INTRODUCTION

World Wide Web is actually a web of individual systems and servers which are connected with different technology and methods. Every site comprises the heaps of site pages that are being made and sent on the server. So if a user needs something, then he or she needs to type a keyword. Keyword is a set of words extracted from user search input. Search input given by a user may be syntactically incorrect. Here comes the actual need for search engines. Search engines provide you a simple interface to search user queries and display the results.

1) Web crawler : Web crawlers help in collecting data about a website and the links related to them. We are only using web crawlers for collecting data and information from WWW and storing it in our database.

2) Indexer : Indexer which arranges each term on each web page and stores the subsequent list of terms in a tremendous repository.

3) Query Engine : It is mainly used to reply to the user's keyword and show the effective outcome for their keyword. In the query engine, the Page ranking algorithm ranks the URL by using different algorithms in the query engine.

4)This paper utilizes Machine Learning Techniques to discover the utmost suitable web address for the given keyword. The output of the PageRank algorithm is given as input to the machine learning algorithm.

2. Body of Paper

The Crawler based Search Engines Crawler based search engines[3] such as Google create their listings automatically. They crawl or spider the web, then people search through what they have found. If you change your webpages, crawler based search engine will find these changes and that can affect how you are listed. Three elements in crawler based search engines are:

• Crawler or spider

• index or catalog

• search engine

software Crawler or spider visits webpages and reads it and index or catalog is like a giant book containing a copy of every webpage that crawler or spider finds. If a webpage changes, then this book is updated with a new one. Human Powered Directories A human powered directory such as the open directory depends on humans for its listings. In this type of search engine, site owner submits a short description of the site to the directory along with category it is to be listed. Submitted site is then manually reviewed and added in the appropriate category or rejected for listing. Keywords entered in a search box will be matched with the description of the sites. This means the changes made to the content of web pages are not taken into consideration as it is only the description that matters. A good site with good content is more likely to be reviewed for free compared to a site with poor content.

Meta Search Engines Meta search engines[4] gives results based on a combination of results from other search engine databases. It uses complex algorithms and virtual databases. A search engine that queries other search engines and then combines the results that are received from all. In effect, the user is not using just one search engine but a combination of many search engines at once to optimize web searching. For example, Dog pile is a meta search engine directories as secondary mechanism. For example, google may take the description of a webpage from human powered directories and show in the search results. As human powered directories are disappearing, hybrid types are becoming more and more crawler based search engines.

But still there are manual filtering of search result happens to remove the copied and spammy sites. When a site is being identified for spammy activities, the website owner needs to take corrective action and resubmit the site to search engines. The experts do manual review of the submitted site before including it again in the search results. In this manner though the crawlers control the processes, the control is manual to monitor and show the search results naturally. B.5 Speciality Search Engines Speciality search engines search a specially created database which is limited to a particular subject. A

speciality search engine, sometimes called a topican or vertical search engine, searches a specially-created database limited to a particular subject. Speciality search engines fall into two main categories: • service • subject-specific Speciality service search engines provide services that are often not available from larger general search engines. Subject-specific search engines search a database tailored to a particular subject. Depending on your area of interest and the type of information you are seeking, speciality search engines can provide more relevant results more quickly than a general purpose search engine such as Google or Yahoo.

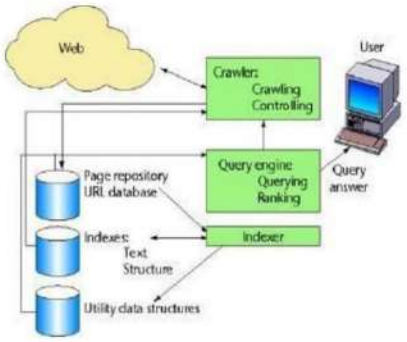


Fig-1: System Architecture

Because of this it would be wise to also submit your blog or website to some of the speciality search engines that cater for your niche. C Search Engine Working While you should always create website content geared to your customers rather than search engines, it is important to understand how a search engine works. Most search engines build an index based on crawling, which is the process through which engines like Google, Yahoo and others find new pages to index. Mechanisms known as bots or spiders crawl the web looking for new pages. The bots typically start with a list of website. URL's determined from previous crawls. When they detects new links on these pages, through tags like HREF and SRC, they add theses to the list of sites to index. Then, search engine use their algorithms to provide you with a ranked list from their index of what pages you should be most interested in based on the search terms you used.

Then, the engine will return a list of web results ranked using its specific algorithm. On Google, other elements like personalized and universal results may also change your page ranking. In personalized results, the search engine utilizes additional information it knows about the user to return results that are directly catered to their interests. Universal search results combine video, images and Google news to create a bigger picture result, which can mean greater competition from other websites for the same keywords. Search engine optimization is a set of rules that can be followed by website owners to optimize their websites for search engines and thus improve their search engine ranking. In addition, it is a great way to increase the quality of your website by making it user friendly, faster and easier to navigate. Steps in search engine optimization are as follows: • Website analysis • Client requirements

• Keyword research

• Content writing

• Website optimization

3. CONCLUSION

Web page ranking is a global ranking of all web pages, regardless of their content, based solely on their location in the web's graph structure. Using these web page ranking techniques, we are able to order search results so that more important and central web pages are given preference. All things considered, search engine optimization will become more resourceful in the upcoming years, but also more complex, forcing marketers to develop more elaborate strategies that bring more types of content, devices and tools into the equation. But no matter which combination of elements you see, the focus should stay on the user and their needs, as machine learning and artificial intelligence technologies will transform ranking factors that can better reflect the needs and expectations of searchers.

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express my profound gratitude and deep regard to my guide.

Dr. G. Madhukar, Associate Professor for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark. We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) **Dr. M. Varaprasad Rao**, **Dr. G.Madhukar**, **Dr. T. S. Mastan Rao**, **Dr. Suwarna Gothane**, **Mr. A. Uday Kiran**, **Mr. A. Kiran Kumar**, **Mrs. G. Latha** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Srujan Raju**, Head, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

We are obliged **to Dr. A. Raji Reddy**, Director for being cooperative throughout the course of this project

We also express our sincere gratitude to **Sri. Ch. Gopal Reddy**, Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help. Finally, we would like to take this

Oppurunity to thank our family for their constant encouragement without which this project would not be completed we sincierly acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

REFERENCES

1. Vishwas Ravall and Padam Kumar, "SEReleC (Search Engine Result Refinement and Classification) – A meta search engine based on combinatorial search and search keyword based link classification," in IEEE-International Conference on advances in Engineering, science and management(ICAESM-2012), March 30,31,2012.Saad ALBAWI, Tareq Abed MOHAMMED, Saad AL-ZAWI, "Understanding of a convolutional neural network," ICET 2017.

2. Vijay Chauhan, Arunima Jaiswal, Junaid Khalid khan, "Web page ranking using machine learning approach," in Fifth International Conference on Advanced Computing and Communication Technologies2015.T. Yamunarani, G.Kanimozhi, "Hand gesture recognition system for disabled people using arduino," vol. 4, 2018

3. Farzaneh Shoeleh, Mohammad Sadegh Zahedi, MoiganFarhoodi, "Search Engine Pictures: Empirical analysis of a web search engine query log," in Third International Conference on Web Research(ICWR), 19,20 April 2017.

4. Donghong Liu, Xan Xu, Yu Long, "On member search engine selection using artificial neural network(ann) in meta search engine," in 2017.

5. P. Kshirsagar, S. Akojwar, Nidhi D. Bajaj , "A hybridised neural network and optimisation algorithms for prediction and classification of neurological disorders" International Journal of Biomedical Engineering and Technology ,vol. 28,Issue 4,Pp. 307-321,2018.