

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/338634368>

Job Aggregator – A Final Year Project Report(Information Retrieval)

Technical Report · November 2017

DOI: 10.13140/RG.2.2.21378.22722

CITATIONS

0

READS

5,105

4 authors, including:



Jeevan Chapagain

The University of Memphis

3 PUBLICATIONS 0 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Information Retrieval [View project](#)



CSEdPad: Investigating and Scaffolding Students' Mental Models during Computer Programming Tasks to Improve Learning, Engagement, and Retention [View project](#)

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Aggregator refers to website or computer software that aggregates a specific type of information from multiple online sources [1]. Moreover, it does not produce or warehouse any item but collects information on goods and or services from several competing sources. Review aggregator, Search aggregator, Video aggregator, News aggregator are some of well-known examples of aggregation model. Each Aggregation model has own functionality. For example, News Aggregation is designed in such a way that it can fetch News automatically from various sources such as online newspapers, blogs, podcasts and video blogs and provide over all news in a one single page for easy viewing.

Like other aggregator, a job aggregator has been designed which has capability of comparing the jobs available in Nepali job sites and listing the best match among all the available jobs. This aggregator model is web based platform that aggregates frequently updated content from related job portal sites and consolidates it in one place for viewing A job aggregator is essentially a search engine, like Google or Bing, but specifically designed to pull job postings from a variety of career sites and job boards so they can all be viewed in one location.

This project is aimed at developing an online search portal for the placement details for job seekers. The system is an online application that can be accessed throughout the organization and outside as well with proper login provided. This system is developed as an Online Job Portal that provides the job for job seekers.

This aggregator will be capable of retrieving information from job portal website of Nepal based on different category like job title, company, location and so on in future but till date only company and job title factor is considered. The user is provided with link of the site that drives them to specific page where they find the detail information about the job.

1.2 Problem Definition

Aggregation model is a platform that consists of information about jobs from multiple sources. Till now, we did not find any aggregation site that is based on Nepali job portal that provides information about different job opening in one place. As we know being citizen developing country unemployment is one of the main problems. It is required, to provide a portal that provides the information about different job vacancy for searching the job. That's why before applying to different company and running here and there we visit as many as possible sites to find the job opening that matches our skillset. Some problems are: Jobs are posted in different Job Portal Site so it is difficult to search each job in each site. More time is spent for searching jobs in different sites. Difficult to remember Job Portal Site.

1.3 Objectives

The main objectives of this project are to extract the job posted in different job portal site, organized them in one site and provide the job on the query of the user. The main objectives can be summarized as:

- To develop a dynamic job portal site.
- To allow the job seekers to find the best available job.
- To provide all job in one place.
- To allow the companies to post job vacancy in the Internet.

1.4 Scope

Scope is one of the important things that should be analyzed properly before starting development of any project. We do also see a wide range of scope in context of Nepal. People have to run here and there carrying the curriculum vitae for the jobs on the basis of the newspaper vacancies. Unemployment is growing in each and every corner of the country. And there is still vacancy in many of the companies and offices. The main reason for this is that, there is no proper media of communication between the job seeker and the job giver. The job seeker keeps searching the job and the employer keeps searching for the right employee. And this project brings both the job seeker and the job giver together in a common platform where the needs of both the job seeker and the job giver can be met without any compromises. Not only this, it is time efficient too. This helps to minimize the waste of time

of the job seeker who keeps searching many job boards and also saves time of the employer as they can quickly find the best from many. This helps both the parties to choose one best from many.

1.5 Limitations

The limitations of this project is that it just lists the job posted in other Job Portal Site so there is no option for posting jobs. It's hard for consumers to get the exact information about job of their interests. This online aggregator system is unable to retrieve information that has been updated by the job board sites instantly. Some instantly changed information can be missed by the system. This online aggregator system is based on few factors that are like "job title" and "company". It has a shady reliability since the job seekers cannot be completely relied on the job portal websites.

1.6 Report Organization

In the first task, jobs will be extracted from the web pages. In the second task, the extracted jobs are filtered, parsed and stored in the database.

The rest of the report is organized as follows: Chapter 2 reviews related works, requirement analysis and feasibility analysis of our project. Chapter 3 specify the system design of the project and describe the basic process model. Chapter 4 implementation and testing are done. It also briefly introduces the various tools used to extract and helps to maintain the project. Conclusions and Testing are discussed in Chapter 5.

CHAPTER TWO

REQUIREMENT ANALYSIS AND FEASIBILITY ANALYSIS

2.1 Literature Review

The rapid growth of job aggregator sites has been a responsible factor for the development of the online job aggregation sites. This makes the job recruitment process easier. It helps us to get the job online from among many options of jobs in the internet. The site itself ranks the jobs available according to the requirements of both job seeker and job giver. We also perform comparison of our project to find out the relevance. The most important thing is we have made a system that is ease of use and as simple compared with other existing aggregation system.

Job aggregator first simply scraps the different posted jobs in the different Job Portal site using spider. Spider scraps the job based on the job title, description, location etc. For scraping the job, we use java and python programming language. The scraped job then parsed according to needed information and stored in the database. For database, we use MySQL database server. The stored information in the database are displayed or extracted according to user query.

2.2.1 Existing System

There are many aggregation sites available online along with unique purpose. Some of them are designed to perform the policy of the insurance like policy bazaar which compares the insurance policy of various companies. But our project doesn't only compare but it also ranks the best result from among many. It is similar to merojob.com, where the jobs are collected from different sites and are compiled together so that the best matching result to the profile provided by the user as extracted. Some existing system are:

2.2.1.1 Octoparse

Octoparse is a modern visual web data extraction software. Both experienced and inexperienced users would find it easy to use Octoparse to bulk extract information from websites, for most of scraping tasks no coding needed. It makes it easier and faster for you to get data from the web without having you to code. It will automatically extract content from almost any website and allows you to save it as clean structured data in a format of your choice.

2.2.1.2 Winautomation

WinAutomation is the most powerful and intuitive platform for Windows automation that enables users to automate any desktop and web based task with zero effort. WinAutomation intelligent Software Robots can be taught to perform effortlessly any task, empowering organizations to achieve greater efficiencies through automation.

2.2 Requirement Analysis

2.2.1 Functional Requirements

The Functional requirement of the system are as follows:

2.2.1.1 Use Case Diagram

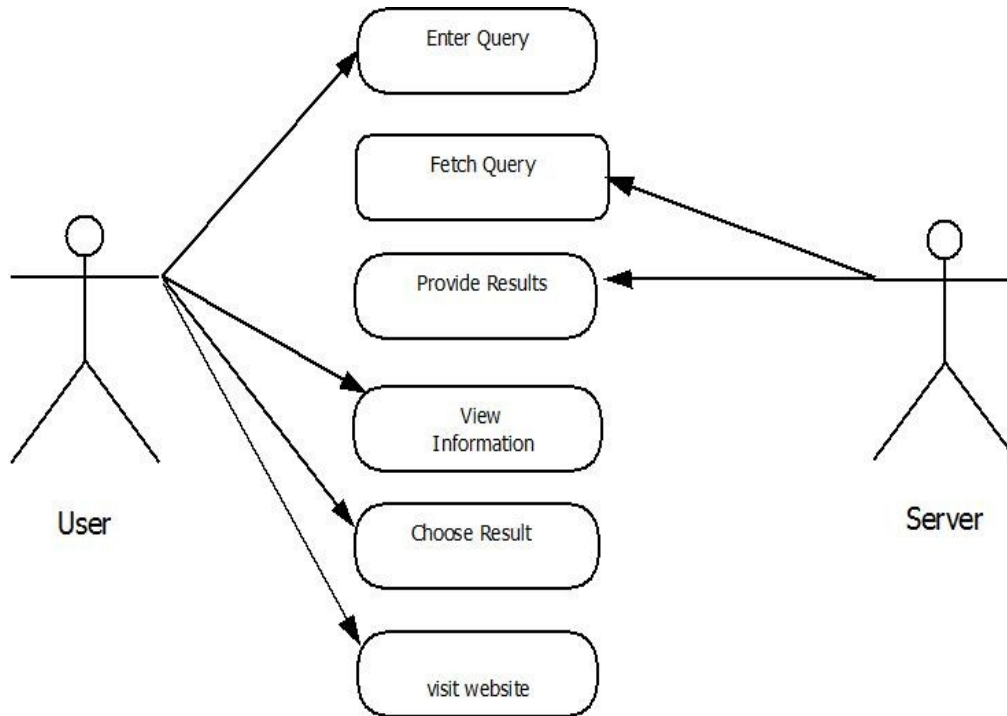


Figure 2.1: Use Case diagram for user and server.

2.2.1.2 Employer (Job Provider) Module:

- Ability to create and edit a company profile.
- Post jobs with detailed job description.
- Employer/Agency Registration.
- Reporting information on how much time left on each job posting and how long left on CV access.
- Information on number of vacancy viewings and applications.
- Quick Employer registration using email and password.
- Update profile at any time.

2.2.1.3 Job Seeker Module:

- Search jobs by keywords, category, and work status.
- Resume posting capabilities. (create, edit, and expire).
- Job Seeker Registration.
- Ability to add and submit photos to profile.
- View how often interested employers viewed the system.
- Ability to apply for a job by clicking the apply button.

2.2.1.4 Admin Module

- Manage Job Seeker and Employer Profiles
- Add New Services
- Ability to create and edit a company profile.
- Set criteria for both users.
- Send newsletters and emails to employers and job seekers.
- Search job based on keyword, categories etc.
- Create / Edit / Save as many icons sets as you like
- Search Jobseekers and Employers on the basis of Keyword, name, Email-address, Country, Industry sector etc.

2.2.1 Non-Functional Requirements

The Non-Functional requirement of the system are as follows:

2.2.1.1 Performance Requirements

- The system provides the features of the online job registration, online apply for job by posting CV and online posting of job vacancies.
- The system shall search record within a short period of time, that is, within few seconds.
- Since the system is web based system, page delays shall be minimized to the possible limit so that the users do not get annoyed.
- The system even displays error message if any error occurred during performing any task in system by the user.

2.2.1.2 Security Requirements

- This system is highly secure since it doesn't allow unauthorized user to login.
- Until and unless the employer or job seekers do not register, they do not have the privilege to post their vacancy for job and resumes respectively.
- The system shall permit only admin or authorized person to add/edit/delete the contents.
- The system shall permit Super administrator to delete/create other administrators.
- The system shall assign different permissions level to different user to perform various functions as per required to fulfill the users request.

2.2.1.3 Software Quality Attributes

- The system shall be available 24hrs to its users over the Internet.
- The system shall be “user friendly”, meaning intuitive and easy to navigate.
- The system shall satisfy specifications, fulfill user’s mission and objectives up to 99%.
- The system is easily maintainable without much effort. It does not require any other additional operation to maintain the system.

2.3 Feasibility Analysis

Preliminary investigation examines 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 of the 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:

2.3.1 Technical feasibility

The technical issue usually raised during the feasibility stage of the investigation includes the following:

- Does the necessary technology exist to do what is suggested?
- Do the proposed equipment’s have the technical capacity to hold the data required to use the new system?
- Will the proposed system provide adequate response to inquiries, regardless of the number or location of users?
- Can the system be upgraded if developed?
- Are there technical guarantees of accuracy, reliability, ease of access and data security?

The current system developed is technically feasible. It is a web based system for searching and posting jobs. Thus, it provides an easy access to the users. The database’s purpose is to create, establish and maintain a workflow among various entities in order to facilitate all

concerned users in their various capacities or roles. Permission to the users would be granted based on the roles specified. Therefore, it provides the technical guarantee of accuracy, reliability and security. The work for the project is done with the current equipment and existing software technology.

2.3.2 Economic Feasibility

A system can be developed technically and that will be used if installed must still be a good investment for the organization. In the economic feasibility, the development cost in creating the system is evaluated against the ultimate benefit derived from the new systems. Financial benefits must equal or exceed the costs. The system is economically feasible. It does not require any addition hardware or software as the interface for this system is developed using the existing resources and technologies available.

2.3.3 Operational Feasibility

Proposed projects are beneficial only if they can be turned out into information system. That will meet the organization's operating requirements. Operational feasibility aspects of the project are to be taken as an important part of the project implementation. Some of the important issues raised are to test the operational feasibility of a project includes the following:

- Is there sufficient support for the management from the users?
- Will the system be used and work properly if it is being developed and implemented?
- Will there be any resistance from the user that will undermine the possible application benefits?

This system is targeted to be in accordance with the above-mentioned issues. Beforehand, the management issues and user requirements have been taken into consideration. So there is no question of resistance from the users that can undermine the possible application benefits. The well-planned design would ensure the optimal utilization of the computer resources and would help in the improvement of performance status.

2.4 System Requirements

2.4.1 Data Modeling

2.4.1.1 Activity Diagram

Activity diagram are used for business process modeling to model the detailed logic of a business rule. Although UML activity diagrams could potentially model the internal logic of a complex operation it would be far better to simply rewrite the operation such that it is simple enough that an activity diagram is not required. The activity diagram for the project is shown in Figure 2.1.

2.4.2 Process Modeling

2.4.2.1 DFD

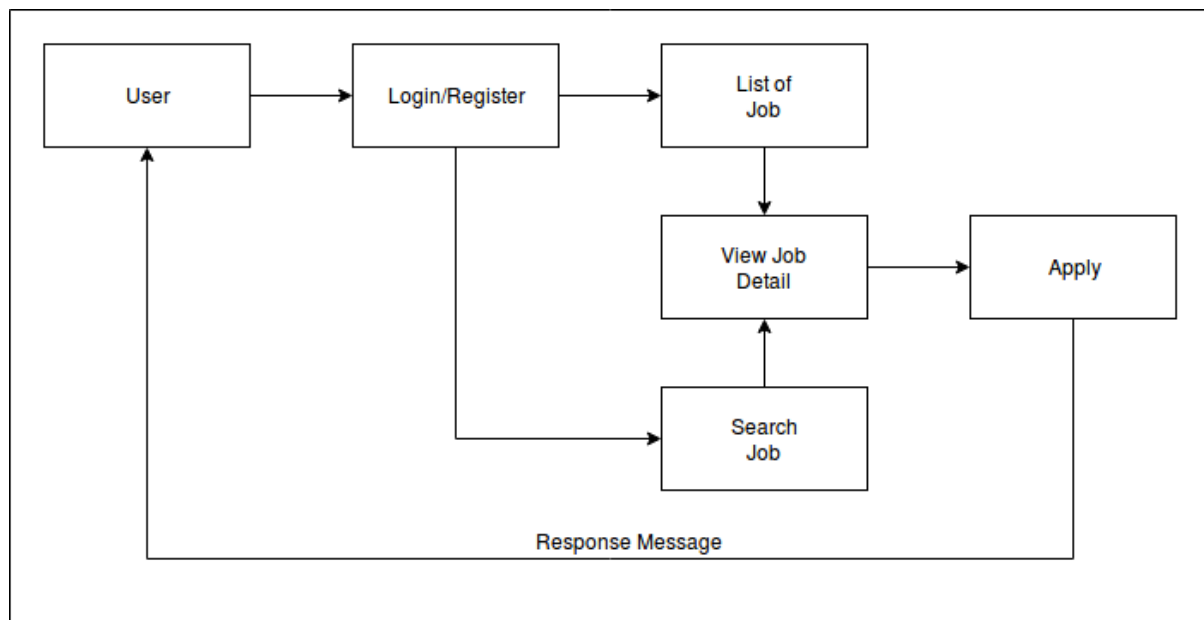


Figure 2.2: DFD for purposed system.

CHAPTER THREE

SYSTEM DESIGN

This document contains the overall design of the system. The system has been designed to enhance effectiveness, incorporate free and open source, platform independent and local language support as well as user friendly solution to the terms related with jobs. The modules have been developed for the online job posting, resumes posting and online processing of the various activities to foster the advanced job placement to the deserving people as per their desire in the prospective firm within the short time span.

The design process includes modular decomposition of the whole system, functional partitioning of the system, ER diagram, DFDs, user interfaces, information flow diagrams etc. The design document acts as a guideline for the system implementation.

3.1 Overall Architecture of System

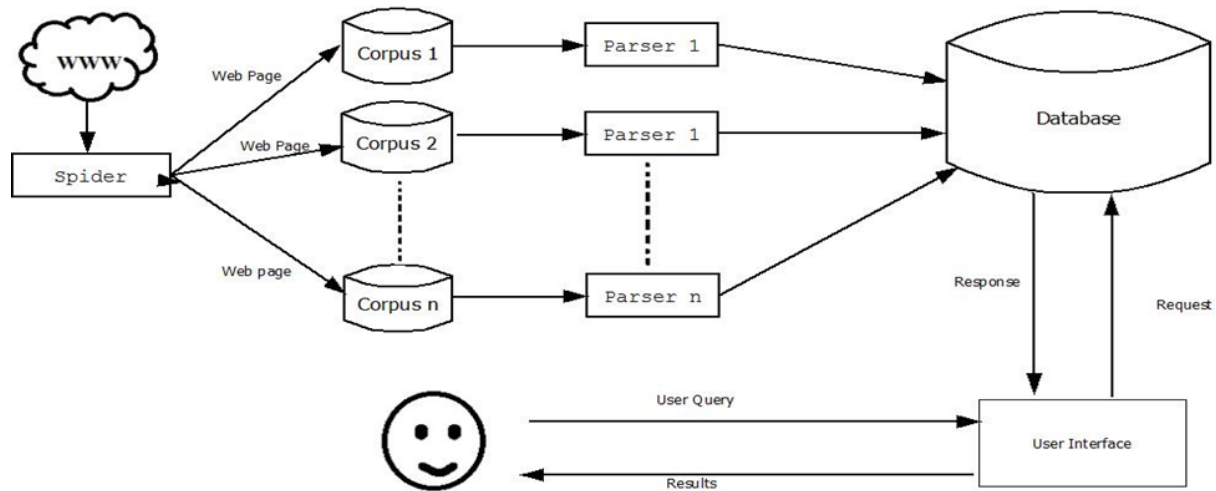


Figure 3.1: Architecture of System.

a. World Wide Web(WWW)

World Wide Web (WWW) contains the different web pages where different categories of the job are posted. From WWW, we extract the needed or required information.

- **Spider:** A program that automatically fetches web pages is called Spider.
- **Corpus:** A corpus is a collection of pieces of languages that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.
- **Parser:** A parser is a compiler or interpreter component that breaks data into smaller element for easy translation into another language.

b. Web Crawler:

Web crawlers are an essential component to search engines; running a web crawler is a challenging task. There are tricky performance and reliability issues and even more importantly, there are social issues. Crawling is the most fragile application since it involves interacting with hundreds of thousands of webs servers and various name servers, which are all beyond the control of the system. Web crawling speed is governed not only by the speed of one's own Internet connection, but also by the speed of the sites that are to be crawled. Especially if one is a crawling site from multiple servers, the total crawling time can be significantly reduced, if many downloads are done in parallel. Despite the numerous applications for Web crawlers, at the core they are all fundamentally the same.

Following is the process by which Web crawlers work:

- Download the web page.
- Parse through the downloaded page and retrieve all the links.
- Parse the html content to find the job vacancy post.
- To parse the organizational information, follow the above two processes and retrieve the "Contact Us" that contains the organizational information.

c. Database:

Database stores the parsed data which we extract using spider. It stores data like as Title, Heading, and Job Description, CV, Company Descriptions etc.

d. Job Seeker (User):

Job Seekers are those who search for the available jobs. Job seekers may be the registered users or visitors. For the registered users, they have to follow the process of registration to facilitate job feature. After registration process, the user's profile is created where he/she has to fill up the profile information which is then used as the contents for the resume. Based on that resume, the system displays the best available job.

3.2 Process Design

3.2.1 Overall Flowchart of the System

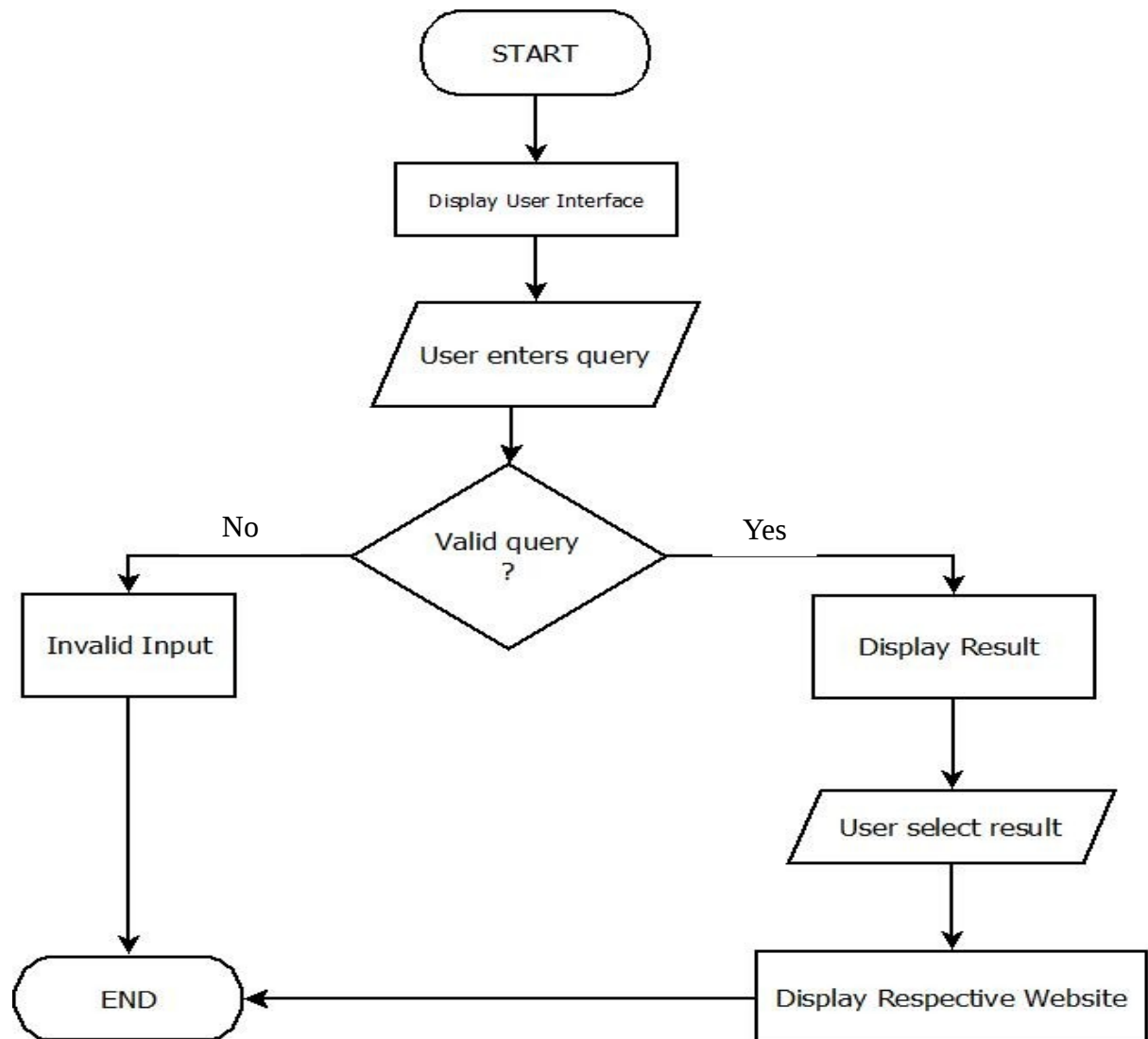


Figure 3.2: Flowchart of System.

3.3 Schema Design

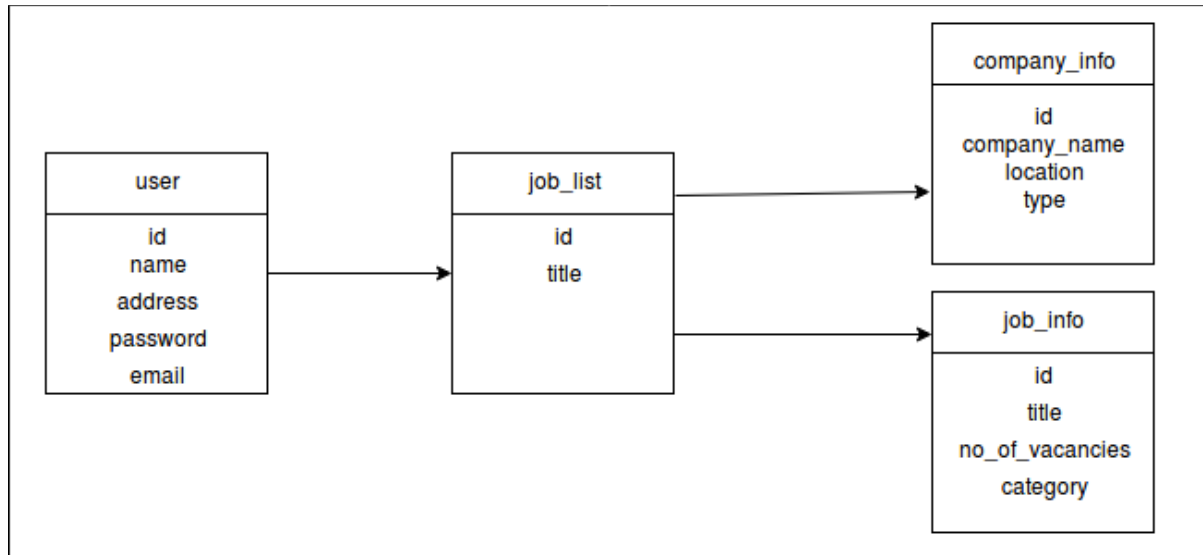


Figure 3.3: Schema diagram of System.

3.4 Sequence Diagram for each use case in use case diagram

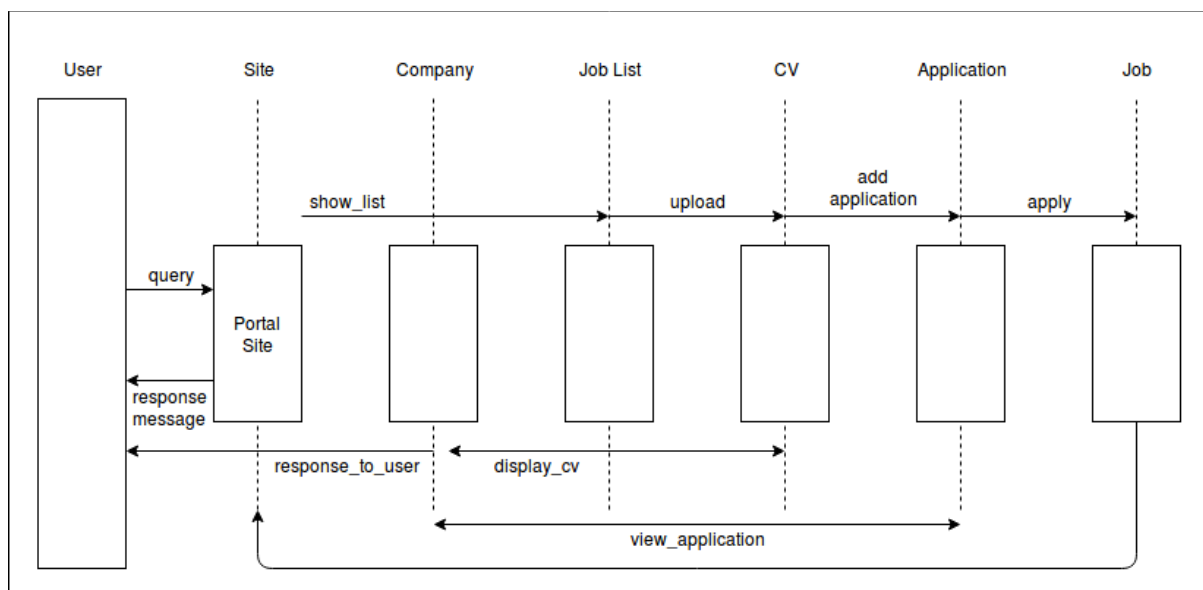


Figure 3.4: Sequence diagram of System.

3.5 Class Diagram

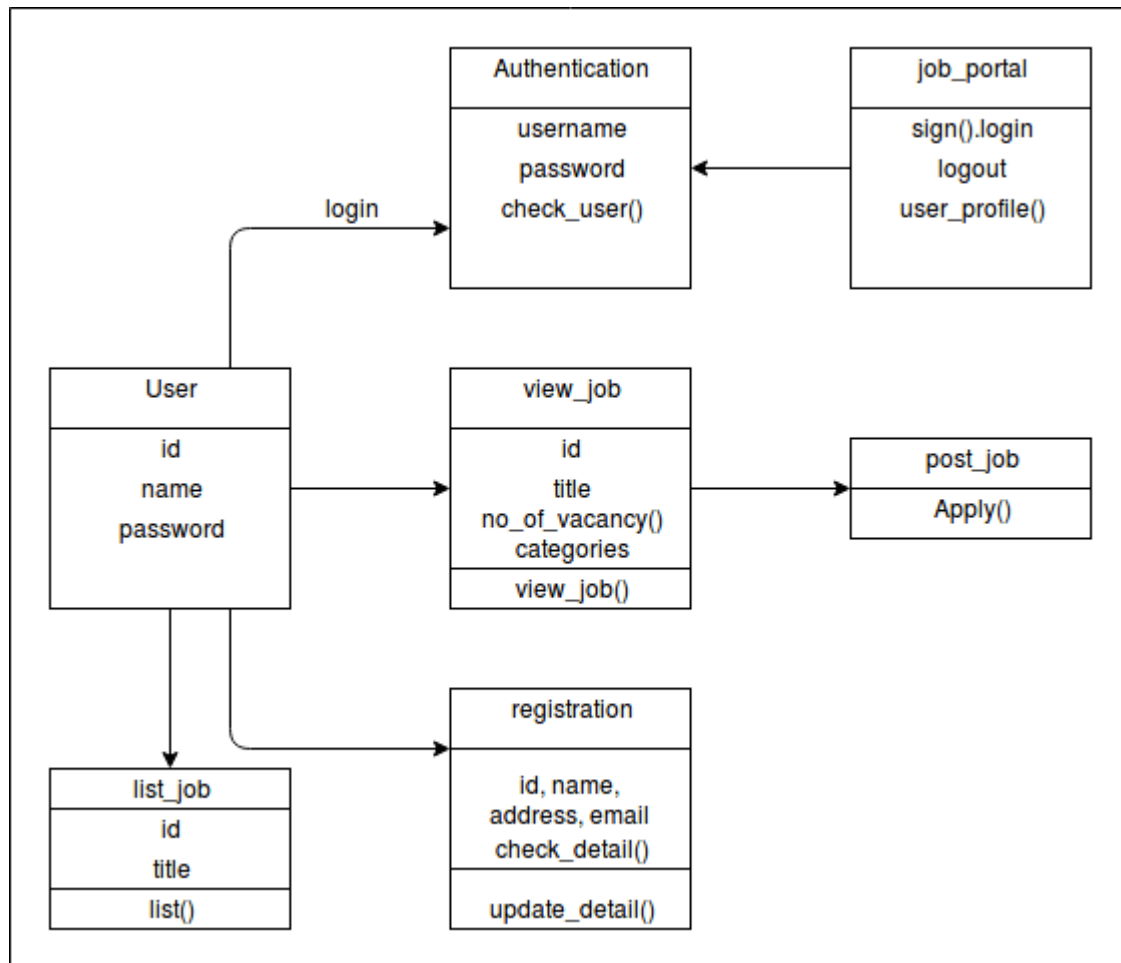


Figure 3.5: class diagram.

CHAPTER FOUR

IMPLEMENTATION AND TESTING

4.1 Implementation

4.1.1 Implementation Model

A system must be implemented for its use. The system is made with two-tier architecture. The system consists of the data-access layer for the data repository via the database server and the service layer for processing the business logic, and an interface. Therefore, we need to install a database server application at the database server and database server's client tools at the individual machines to access the database server. However, the database server can be physical servers or the local system itself. The next step is to migrate the database to the database server that is intended to serve the application. Then, we can host the site; the system is ready to be used.

4.1.2 Experimental Setup

In chapter 3, we presented the framework for aggregation of the online jobs. The confidence in the overall system performance is yet to be justified by experimental results. In this chapter, we discuss the experimental setup to evaluate the whole system.

Job aggregation experiments rely on a large corpus of documents. Many large corpora are available for aggregation research. However, our interest here is in job aggregation. Since, the goal of our project work was to create a model for aggregating online jobs. So, we have to rely on web documents in our experiments.

4.1.2.1 Corpus

The document corpus for carrying out our project study consisted of the jobs collected from five different job portals site.

4.1.2.2 Dataset

The dataset was constructed by retrieving the structured text from xml file which consisted of the heading, location, description and the content of the jobs.

4.1.3 Preprocessing

4.1.3.1 Crawling

A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Other terms for Web crawlers are ants, automatic indexers, bots, [1] Web spiders, [2] Web robots [2], etc.

The process followed by the web crawler is called Web crawling or spidering. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for sending spam).

A Web crawler is one type of bot, or software agent. In general, it starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies.

The large volume implies that the crawler can only download a limited number of the Web pages within a given time, so it needs to prioritize its' downloads. The high rate of change implies that the pages might have already been updated or even deleted.

The number of possible crawlable URLs being generated by server-side software has also made it difficult for web crawlers to avoid retrieving duplicate content. Endless combinations of HTTP GET (URL-based) parameters exist, of which only a small selection will actually return unique content. For example, a simple online photo gallery may offer three options to users, as specified through HTTP GET parameters in the URL. If there exist four ways to sort images, three choices of thumbnail size, two file formats, and an option to disable user-

provided content, then the same set of content can be accessed with 48 different URLs, all of which may be linked on the site. This mathematical combination creates a problem for crawlers, as they must sort through endless combinations of relatively minor scripted changes in order to retrieve unique content.

Crawling Process

- The crawler begins with a seed set of URLs to fetch
- The crawler fetches and parses the corresponding web pages, and extracts both text and links
- The text is fed to a text indexer, the links (URL) are added to a URL frontier (\approx crawling agenda)
- (continuous crawling) Already fetched URLs are appended to the URL frontier for later re-processing
- traversal of the web graph
- Reference point: fetching a billion pages in a month-long crawl requires fetching several hundred pages each second
- Some potential issues:
- Links encountered during parsing may be relative paths
- normalization needed
- Pages of a given web site may contain several duplicated links
- Some links may point to robot-free areas

4.1.3.2 HTML Parsing

An HTML parser analyzes HTML markup tags. The parser is the front-end for a markup analyzer, the entity designed for the identification of document structure. The markup analyzer contains the core logic for identifying different document parts and assigning levels of significance for them according to the scheme described earlier.

The parser also takes care of decoding encoded character entities, removing whitespaces, detecting block-level structures, inline structures, and stripping tags.

4.1.5 Tools Used

The tools used for the design and developments of the system are as follows:

a. Development Environment

PHP, Python: For business logic and interface design.

MySQL: For database design.

HTML, CSS, JavaScript: For validation, styling and presenting document.

b. Development Tool

Sublime Text for text editing.

c. Database

MySQL is finally a client/server database produce that can deliver world-class performance at a price that most enterprises can afford not only to purchase, but also to support. MySQL Server has the capabilities to create a solid production data server.

d. Bit bucket:

Bit bucket is another tool that we have used in our project management for source code management. We team members worked from different place and use “git” for our source code management. Bit bucket is an application to code, test, and deploy code together. It provides Git repository management with fine grained access controls, code reviews, issue tracking, activity feeds, wikis, and continuous integration.

4.2 Testing

4.2.1 Unit Testing

Unit Testing concentrates on each unit of software as implemented in the source code. It only tests the functionality of the units themselves.

We have used different test cases in this project for unit testing. We used unit testing to test:

- Existence of classes
- Existence of methods
- Exception thrown properly or not
- Actual implementation works or not.

Table 4.1: Unit Testing.

Test Case	Description	Expected result	Actual result	Status
Job Search	Users puts the desired job keyword in the search box.	If the keyword matches the keyword in database, the result is shown.	Users get their search result.	Successful
Job Search	Users puts the desired job keyword in the search box	If the keyword doesn't match the result is not shown to the user	Users doesn't get their search result	Successful
Job Search	Users puts the desired job keyword in the search box	If numerous job found of that keyword, search result is shown in pagination	Users get more search result, result shown in pagination	Successful

4.2.2 Integration Testing

After all the modules were developed and tested separately, they were merged systematically to form a complete system. The system was tested again to know whether it is functioning properly or not after integration.

Table 4.2: Integration Testing.

Test Case	Description	Expected result	Actual result	Status
Job Search	User searches a job or company on the system.	User should be able to see the result matching their search queries.	The search result was displayed on the window.	Successful

4.2.3 System Testing

The system was then tested as a whole. Different varieties of testing and training datasets were used to check whether the system is giving accurate result or not.

Table 4.3: System Testing.

Test Case	Description	Expected result	Actual result	Status
Job Aggregator	User inputs and keyword in the system to search the job.	The system should take a keyword as input and display the job that matches the keyword.	The system was able to display the job that matches the search keyword.	Successful

CHAPTER FIVE

CONCLUSION AND LIMITATION

5.1 Conclusion

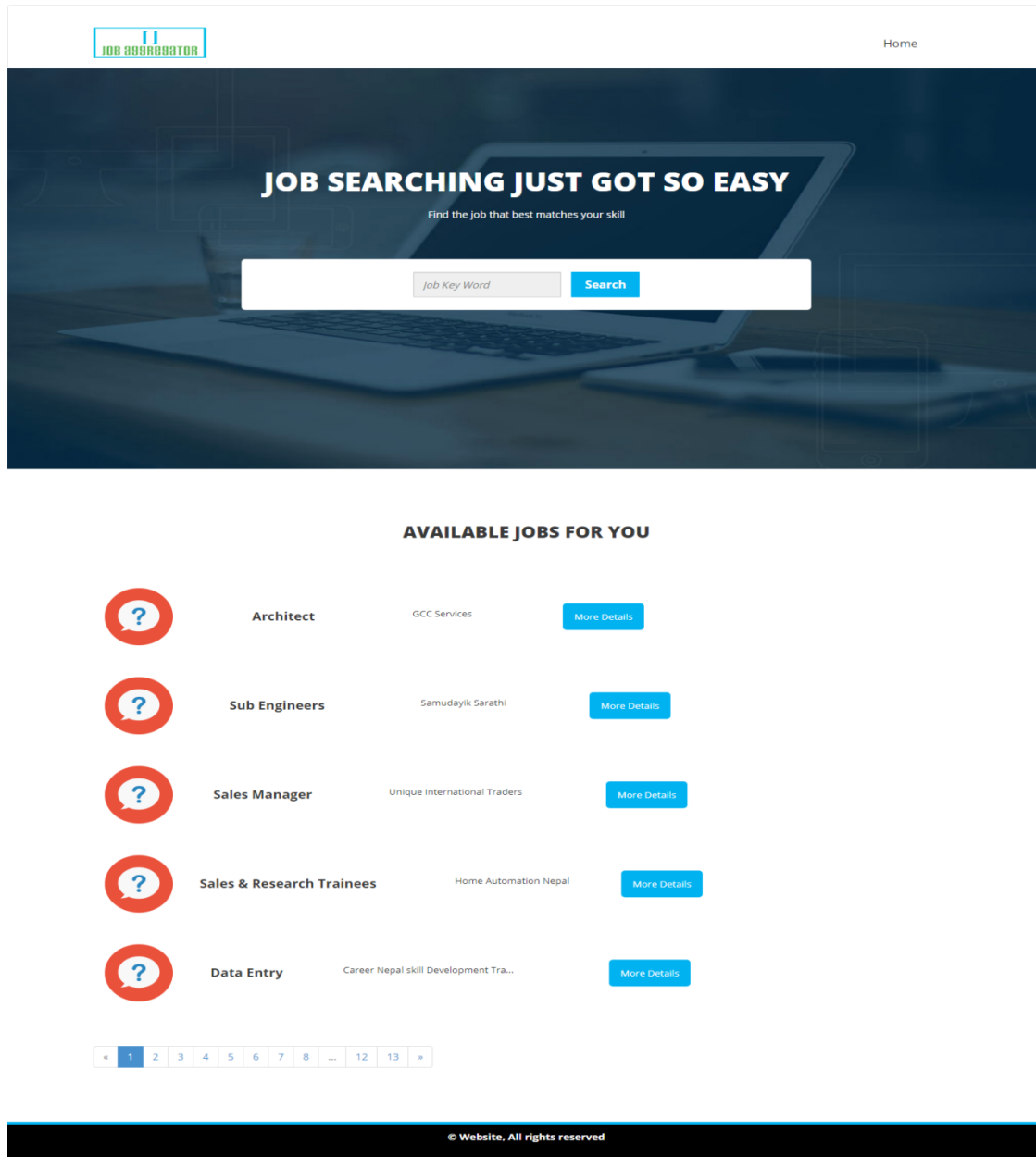
Online Aggregation System has been developed as a prototype to search a job through the job portal sites. The main thing that we want is that our website must be user friendly, any user who is capable to use web can visit to our website and perform their shopping. As we know about the condition of Nepal's internet connection problem, we also put some effort towards making our website as light as possible so that it could load easily in slow internet connection. As our website is domain specific to jobs that the user is searching for. It makes Nepali online job seekers easier to perform their job searching easily, efficiently and as quick as possible.

5.2 Recommendations


The usage of this web application can be increased if the mobile version of this application can be releases which will be able to give some facility even in offline mode. Additional features like instant chat room, video conferencing, can be added to upgrade the web application and make it more useful. If internet infrastructure is made available this web application can be effectively used in the areas where skilled teaching manpower's are lacking. In such places, the absence of skilled teachers can be filled through this application.

APPENDICES

Screenshot of Home Page



Screenshot of Result Page




Home

JOB SEARCHING JUST GOT SO EASY

Find the job that best matches your skill


Search result for:**accountant**



Accountant

Hotel Holiday Home

More Details



Receptionist Cum Accountant

Real Land Pvt. Ltd

More Details

© Website. All rights reserved

References

- [1] Kobayashi, M. and Takeda, K [2000]. "Information retrieval on the web".
- [2] Karloss, Artto. [2007, 4-5] "What is project Strategy?", International Journal of Project Management
- [3] Spetka, Scott. "The TkWWW Robot: Beyond Browsing"
- [4] <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue1/art2/node4.html>
- [5] <http://www.businessdictionary.com/definition/aggregator-model.html>
- [6] <http://business.everestcollege.edu/articles/7-essential-strategies-of-successful-project-management>
- [7] <https://www.quora.com/How-does-the-aggregate-business-model-work>
- [8] https://www.sciencedaily.com/terms/web_crawler.htm