CAREER RECOMMENDER SYSTEM

Author

TAKREEM SAEED

FALL 2016-MS-16(CE) 00000170606

Supervisor

DR.FARHAN HUSSAIN

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

AUGUST 2020

CAREER RECOMMENDER SYSTEM

Author

TAKREEM SAEED

FALL 2016-MS-16(CE) 00000170606

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Engineering

Thesis Supervisor:

DR.FARHAN HUSSAIN

Thesis Supervisor's Signature: _____

DEPARTMENT OF COMPUTER &SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

AUGUST 2020

Declaration

I certify that this research work titled *"Career recommender system"* is my work under the supervision of Dr. Farhan Hussain. The work has not been presented elsewhere for assessment. The material that has been used from other sources is properly acknowledged/referred to.

_____

Signature of Student

Takreem Saeed

00000170606

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical, and spelling mistakes. The thesis is also according to the format given by the university.

_____

Signature of Student

Takreem Saeed

00000170606

_____

Signature of Supervisor

Dr. Farhan Hussain

Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

_____

Signature of Student

Takreem Saeed

00000170606

_____

Signature of Supervisor

Dr. Farhan Hussain

Copyright Statement

Acknowledgments

All Praise and thanks be to Allah! We praise and thank Him, ask Him for His help and forgiveness, and we seek refuge in Allah from the evils of our souls and the mischiefs of our deeds. He whom Allah guides will not be misled, and he whom Allah misleads will never have a guide. I testify that there is no deity but Allah alone, without any partners, and that Muhammad (SAW) is His 'Abd (worshiper) and Messenger.

I would like to express special thanks to my supervisor Dr.Farhan Hussain for his exemplary supervision, unparalleled insight, and constant encouragement during this thesis. His constant curiosity with the challenges that I faced and his helpful notions gave me the direction to complete my research.

I owe the deepest gratitude to my honorable co-supervisor Dr. Arslan Shoukat without whom it was almost impossible to complete my research work. I am highly obliged for his support and guidance throughout my research work. I wouldn’t be here today without his encouragement throughout the final stages of this research.

I would also like to thank Dr. Wasi Haider and Dr. Saad Rehman for being on my thesis guidance and examination committee and for guidance and cordial support, which helped me in completing this task through various stages.

Finally, I would like to thank my husband M. Sufian and my sister Swabia Saeed that has been a constant source of love, support, and inspiration for me throughout my thesis.

To my father who has given me the life I love today

Abstract

In recent years, Recommender systems are utilized in a variety of areas. One reason behind why we want a recommender system in current society is that an individual has a large number of alternatives to use because of the pervasiveness of the Internet. A recommender system seeks to estimate and predict user content preference. Old recommender systems used State-of-the-art recommender algorithms like content based filtering to predict ratings. Career Recommender system provides Engineering candidates the best possible available jobs relevant to their skills, qualification, etc. Four to six major engineering disciplines are covered in this recommender system. The proposed approach is tested using a career recommendation dataset which is collected from many students of different disciplines of various universities. A deep NLP based CNN model is used to predict the best jobs with maximum precision.512 hidden layers are used to increase the performance of this system. Career recommendation takes care of the users and saves their cost and time spending on traditional job searching methods. Comparative study demonstrations that the proposed methodology of prediction of the best jobs achieves better results with an accuracy of 84% when matched with content based filtering technique where 81% accuracy is gained for content based career recommender system.

Key Words: *Content based Filtering, Deep NLP(Natural Language Processing), Convolutional neural network (CNN Model)*

Table of Contents

List of Figures

List of Tables

List of Acronyms

CF	Collaborative Filtering
CNN	Convolutional Neural Network
NLP	Natural Language Processing
FoDRA	Four Dimension Reommenation Algorithm
SJH	Students Job Hunting
SVM	Support Vector Machine
WWW	World Wide Web
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
IDE	Integrated Development Environment
ANN	Artificial Neural Network
EDA	Exploratory Data Analysis
NLTK	Natural Language ToolKit
ID	One Dimensional
CRS	Career Recommender System

CHAPTER 1: INTRODUCTION

This chapter provides a detailed introduction to the research and research concepts. This section is organized in multiple sub-sections. The background study is discussed in **Section 1.1**, types of recommender systems are discussed in **Section 1.2**, the problem statement of this research is presented in **Section 1.2**, in **Section 1.3** the proposed methodology is discussed, the details about research contribution is given in **Section 1.4**, and thesis organization is presented in **Section 1.5**.

1.1 Background Study

Due to the invention of the internet, everyone now has access to online data. This huge amount of data poses the question: "From all this data how one can find useful information about his interest". To overcome this issue recommender systems are used, so that they can help users to find data of their own interests. Nowadays recommender systems are used in many applications like recommending movies, music, and books, etc.

To facilitate engineering graduates about their career according to their area of interest is the need of this modern time. The career recommender system will suggest different career options to engineering graduates regarding their skills. This system will save their time by recommending them relevant jobs of their fields.

Recommender systems mainly focus on predictions, i.e. to predict that item or information which seems to be interesting or useful for the user. A recommender system usually consists of three elements.

- Recommendation Content
- User's preferences
- Recommendation Technique

Prediction of a recommender system depends upon user's profile. The aim of profile making is to get an idea about user's interests or skills. For this purpose, information from the user is extracted [10].

1.1.1 User Profile

The user profile is usually constructed by two things:

- Knowledge of the user
- Behavior of the user

Knowledge of the user can be gathered by arranging an interview or by doing a survey of a questionnaire.

In the second approach, the user's behavior acts as a model. Different machine learning techniques are used to find out useful patterns in the behavior.

1.1.1.1 Explicit Profile

Ask the candidate to fill out the questionnaire by a specific form. Advantage of this explicit profiling is that recommender system can determine user’s needs directly through this technique.

1.1.1.2 Implicit Profile

In this technique, users' behavior is kept under observation. This method is usually obvious to the users. To track the user's browsing activity a log file is used. This file keeps some specific user’s identification and behavior information. For example amazon.com has a log file which has each customer’s buying history and based on that history amazon.com recommends specific products to the customers.

1.2 Types of Recommender System

Important types of recommender systems are:

- Knowledge based
- Content based
- Collaborative filtering
- Hybrid filtering

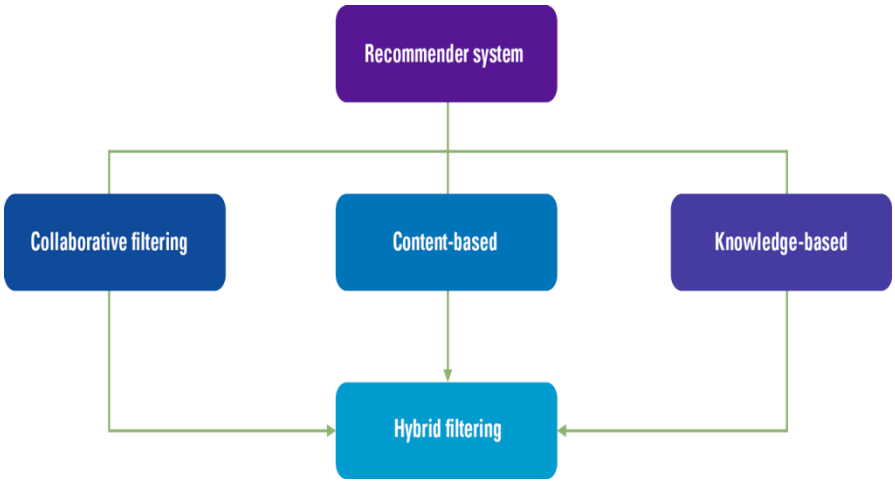

Figure 1.1: Types of Recommendation Systems

1.2.1 Knowledge based

Knowledge based recommender systems recommend items to the user by knowing the knowledge of the item or the user. Knowledge based decision rules are used for recommendation of items to the users who fulfill the rules. Explicit user profiles are extracted in this technique.

Knowledge based filtering recommendation technique proceeds in these steps:

Set of all items is denoted by I, set of all users is represented as U.

I. Features of all items in I are identified.
II. A description of User's interests in U's are identified.
III. Matching between I and U's interests are generated.
IV. Based on this matching, Top N Items are recommended.

1.2.2 Content Based Filtering

Content based filtering completely depends on information retrieval and filtering. Each item has a set of attributes which is called as content profile. It measures the similarity between the new item and with that item which is preferred by the user with the help of content profile .The content profile is developed by means of extracting a set of features from an item. In domains such as text files and electronic items, key phrases of a report or physical characteristics of an item are used to construct such item profiles [11].

Content Based filtering recommendation techniques involves following steps:

I. Features of all items in I or personal features of all users in U are identified.
II. Features of items in I (which are selected by active users U) are identified.
III. Generate a classifier that fits U's rating behavior
IV. Use U's rating on i.
V. Based on this classifier top N recommendations are made.

1.2.3 Collaborative Filtering

The basic idea in collaborative filtering is that if two users have exactly same likeliness for a number of items then they must have same likeliness for other items as well [12].

Collaborative filtering based recommender systems follow these steps:

I. For a target user, the set of his ratings are identified.
II. The users who are more similar to target users are identified.
III. The items purchased by these similar users are identified.
IV. For each one of these items, a prediction is generated (of the rating that is given by the target user to the item).
V. Based on this predicted rating, a set of top N items are recommended.

1.2.4 Hybrid Filtering

By combining content based and collaborative filtering, hybrid recommender systems are developed. These systems have qualities of both content and collaborative filtering recommender systems.

Clustering based hybrid filtering recommender systems have following these steps:

I. For all users, a description of his interests are identified.
II. Clusters are made with related interests by using clustering algorithm.
III. For a new user, a set of his interests are identified.
IV. A prediction is generated to find the best cluster, in which new user's interests best fits in.
V. Top N items of this cluster are recommended.

1.3 Recommendation System Generations

Recommender systems are growing day by day. Recommender systems can be categorized as first generation, second generation and third generation. First generation recommender systems

consists of E-commerce domain. For second generation, social network and social contextual information is required to give accurate recommendations. Third generation recommender systems deals with location based information and internet of things to give recommendations [13].

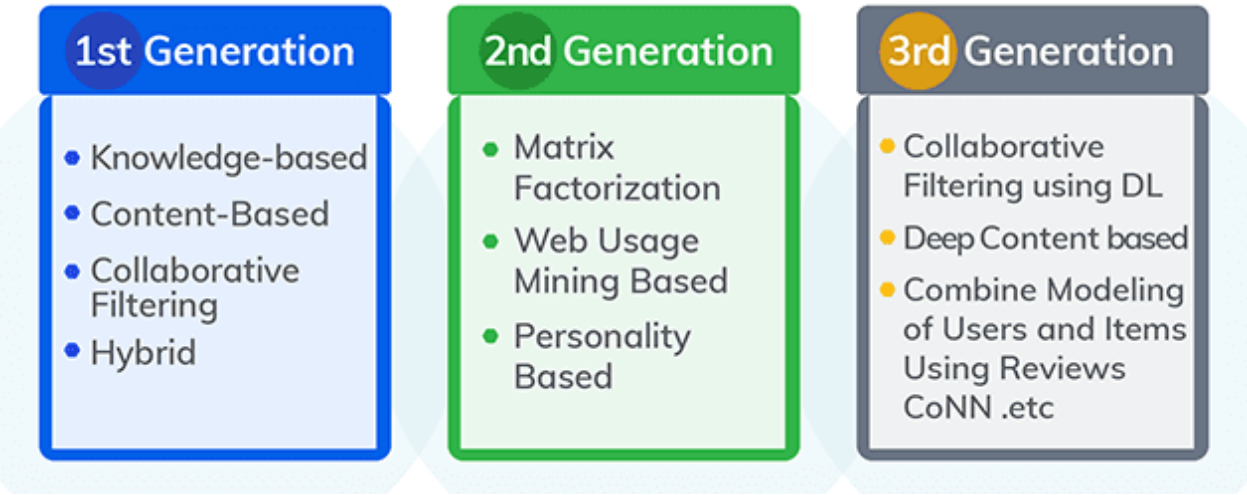

Figure.1.2: Recommendation System Generations

1.4 Problem Statement

The problem discovered in this thesis is to suggest a recommender system for engineering students carrying out their studies to give them ease for their career selection using machine learning techniques. With the advent of the internet, recommender systems are becoming more popular nowadays in every area, Netflix and amazon give us the best examples of the importance of recommender systems.

Building a career recommender system for engineering students has following challenges,

- Non-availability of dataset
- Validation of dataset
- Choosing the right machine learning technique

This thesis will try solving the above challenges by reviewing different ways of making a dataset, validating them and choosing the right machine learning technique which gives us the optimum results.

1.5 Proposed Methodology

We try to perform this research in a systematic way. Figure 1.3 represents the flow of research step by step. In the first step we identify the problem. Then we proposed a solution for the problem identified in the first step. We carried out a detailed and comprehensive literature review which helps us to identify the potential solution for the problem. We reviewed the researches carried out related to our proposed solution, analyze, and compared it.

This thesis suggests another kind of recommender system called Career recommender system which will recommend different career options to engineering graduates by using Convolutional Neural Networks (CNN) for Natural Language Processing technique. Dataset for this recommender system is divided into two parts, the first part is termed as a user profile and is collected by doing a survey of a questionnaire from engineering graduates, and the second part which is job profile is collected from job holder engineers who are working in different companies or universities of the country.

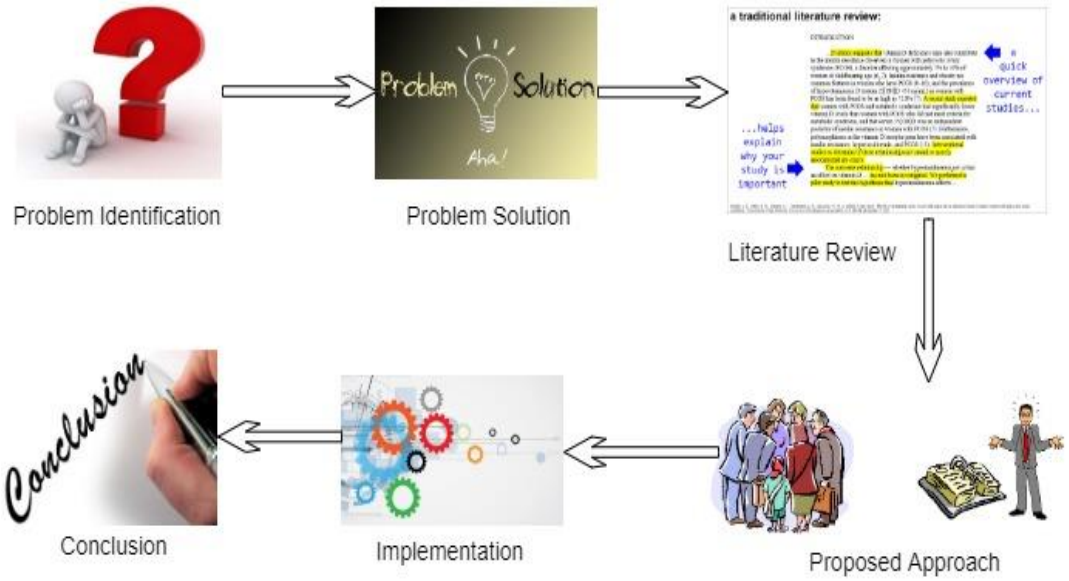

Figure 1.3: Research Flow

1.6 Research Contribution

We conducted a survey and collected the dataset, extracted the relevant features and proposed a career recommender system. A detailed overview of the proposed approach is as follows:

- We have presented a model that will take the dataset as input and give the best job according to the features as output.
- Convolutional neural network for natural language processing based model is used.
- This model first learns from the training data
- The best job prediction is done by using the model, in which hidden layers of the Neural Network will do all the predictions.
- We have provided validation of our work by applying the same dataset on content based filtering and then by comparing the results of CNN based model with traditional content based recommender system.

1.7 Thesis Organization

Figure 1.4 represents the organization of the thesis**. Error! Reference source not found.** eals with the introduction having detailed background study about the concepts used in the research, problem statement, research contribution, and thesis organization. **Error! Reference ource not found.** contains the literature review which provides a description of work done in the area of recommendation systems using machine learning. In the Literature review we also highlight the research gaps that we encountered. **Error! Reference source not found.** covers the etails of the proposed methodology used for the identification of problems. **Error! Reference source not found.** presents the detailed implementation regarding the proposed model, dataset formation, feature selection, and implementations. **Error! Reference source not found.** rovides the results for our proposed methodology by testing on the dataset and a comparison table by comparing it with 4 other recommendation models. **Error! Reference source not ound.** contains a summary of our work and extension of our thesis which can be done in the future.

Figure 1.4: Thesis Outline

CHAPTER 2: LITERATURE REVIEW

This chapter presents a brief assessment of job recommendation systems and the existing work of techniques used for job recommendation systems. After a brief literature review of work conducted in this area we enlightened the research gaps that we found in previous works.

2.1 Literature Review

Karbhari, Deshmukh, Shinde (2018) proposed a job recommender system for college students on the behalf of their marks in the exams from 1st year to final year. The dataset including student details, mark sheet and add-ons were collected from database. The technique used to build job recommender system was content based filtering by finding out correlation between the content of jobs specifically words in a document and user's marks in each subject, add-ons like certificates, short courses etc. Proposed model included multiple phases like preprocessing of the data, filtering out relevant information from the raw data, clustered users into various categories, identifies users who were most interested to the specific jobs and finally jobs recommended through content based filtering. The results obtained from correlation in content based and jobs were suggested to the college campus students. The Proposed model of content based recommendation system is shown in figure 2.1 [1].

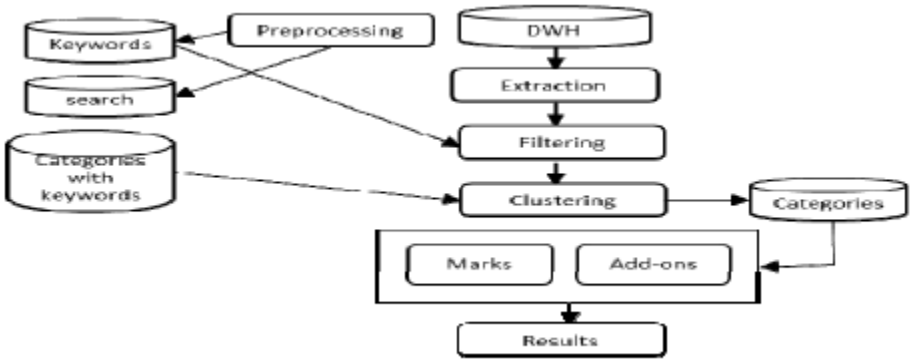

Figure 2.1: Proposed model of content based recommendation system

Almalis, Tsihrintzis, Karagiannis, Strati (2016) emphasized on updated minkowski distance to build a content based job recommender system. The dataset used in the recommendation were taken from an online data repository www.kaggle.com [2]. This paper

proposed a Four Dimension Recommendation Algorithm (FoDRA) by using Content based filtering technique. FoDRA used 8- dimensional vector for both the job seeker and
Job. FoDRA used structured data of Job and candidates profile which was extracted from unstructured form by doing content analysis on candidate's CV and job description.
Results of the content analysis for the job seeker is shown is figure 2.1.

Attribute names (required skill)	Attribute values (required skill level)
S_1	25
S_2	20
S_3	11
S_4	20
S_5	29
S_6	19
S_7	21
S_8	21

Figure 2.2: Content Analysis Results for the Job Seeker

By doing a comparison between Job seeker's profile and Job's profile, FoDRA suggested different top 5 jobs to the job seekers for different values of 'p' like Manhattan distance and Euclidean distances. The top 5 jobs are shown in the figure 2.3.

Sorted list of jobs for different values of p			
$p = 1$[e]		$p = 2$[f]	
Job	Suitability Value	Job	Suitability Value
Job_5	10	Job_5	10,585786
Job_{25}	7	Job_{25}	7,763932
Job_{11}	6	Job_{11}	6,763932
Job_{14}	5	Job_{14}	6,171573
Job_{22}	3	Job_{22}	4,171573

[e] p=1 is Manhattan distance

[f] p=2 is the Euclidian distance

Figure 2.3: The Top 5 Jobs after the FoDRA Execution for different values o

Paparrizos, Cambazoglu, Gionis (2011) provided supervised machine learning based job recommender system. This system predicted the next best job company/ institute for the employee, as job seeker wanted to switch his job from one company to the next [3]. Millions of job transitions and other linked details were extracted from openly available employee profiles in

web. Directed graphs (I, T) were used to represent job transitions. Institutions were represented by the set of nodes I, and set of edges T shows the transitions between institutions. Dataset statistics are shown in the figure 2.4.

Description	Value
Number of profiles	5,298,912
Number of unique company affiliations	1,278,240
Number of unique university affiliations	195,849
Average number of company affiliations	3.04
Average number of university affiliations	1.27

Figure 2.4: Dataset Statistics

Prediction model trained through a lot of job transition extracted from the web. Supervised machine learning naïve bayes hybrid classifier gave the highest prediction accuracy. Prediction accuracy is shown in figure 2.5.

	Accuracy (%)		
Setup	Baseline	DTNB	Difference
I	15.21	66.78	51.57
II	15.40	78.26	62.86
III	15.97	86.09	70.12

Figure 2.5: Prediction Accuracy

Liu, Ouyang, Rong, Song, Tang, Xiong (2016) suggested a job recommender system to college students to help them in finding out best job according to their potential. Basic challenge in this study was zero working experience, as students had no history. This paper provides jobs to the students in a very short span of time. The dataset used in this research was real dataset obtained from Beihang University in 2016, which consists of 658 Master Graduates and 3830 ratings given by them to 856 employers. Experimental data included 80% ratings as the training data and 20% as the test data. Cold start ratings experimental data partition dataset is shown in figure 2.6 [4].

	#Student	#Employers	#Rating	#Acceptance
Training dataset	526	748	2858	1091
Test dataset	132	400	972	292

Figure 2.6: Cold start ratings experimental data partition dataset

Similarity calculated on the basis of major subjects, Home town and GPA. Similarity then used to rate prediction. This paper proposed a rating based prediction algorithm to recommend jobs to the college students by obtaining the feedback from the graduates. For better performance, student's interest is also considered as an important attribute. Rating based job recommendation framework is shown in figure 2.7.

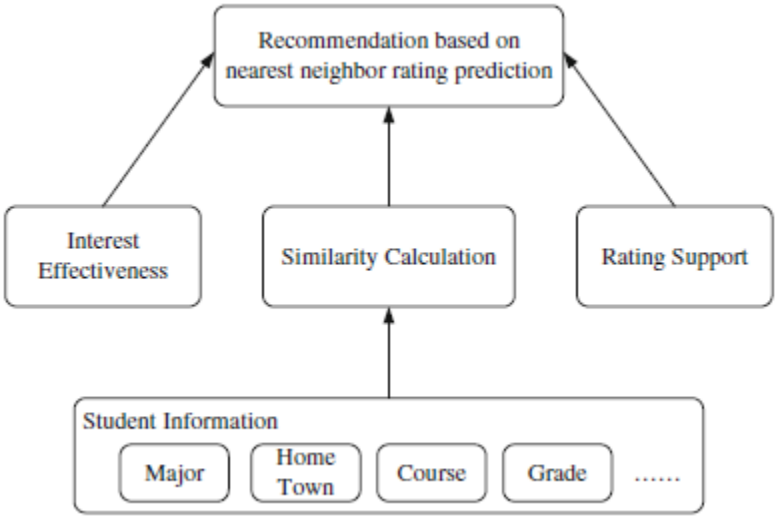

Figure 2.7: Rating based job recommendation framework

Al-Otaibi, Ykhlef (2012) discussed the e-recruiting process and different efficient ways of constructing customized job recommender system. This paper addressed the most common problem of the internet era, most of the companies now a days use internet-based recruiting system to decrease the advertisement cost and recruitment time and as a result a lot of irrelevant information is obtained. It decreases the chances of most eligible applicant to be hired for a suitable position. To improve the E-recruiting, different recommender system techniques have been proposed. The recommender system helped users to find out jobs related to their interests. The advantages and disadvantages of some job recommender approaches discussed in fig 2.8.

Many approaches combined to provide the best fit results to the candidates for their job recommendations [5].

Recommendation approach	Techniques	Advantages	Disadvantages
Hybrid job recommender systems	Probabilistic hybrid approach.	Bidirectional recommendation. Relational aspects are included.	Binary representation only. Less attributes used. No perfect measures.
	Proactive job recommender system.	Adaptive system. Use many attributes. Use ontology to categorize jobs and as a knowledge base to define features (attenuate cold-start problem).	Key words search method. One way recommendation. Knowledge acquisition and knowledge engineering problems. No relational aspects are included.
	Semantic matchmaking for job recruitment	Bidirectional recommendation. Effective matching methods. Includes many attributes. Relational aspects are included. Qualitative and quantity representation (proficiency level for skills is included). Use two levels in skills matching (constrains and preferences).	Knowledge acquisition and Knowledge engineering problems. Tools and technologies skills excluded.
	Fuzzy multiple criteria method for recruitment.	Use many attributes. Relational aspects are included. Effective matching methods. Use linguistic variables to determine skill levels.	One way recommendation.
Content-based job recommender systems	Machine learned recommender system	Use many attributes. Transition history is included.	One way recommendation. No relational aspects are included. Scalability, ramp-up, and data sparsity problems.
	System for screening candidates	Use many attributes. Various information retrieval techniques are used. Constrains used to eliminate candidates before ranking.	Inefficient measures. One way recommendation. No relational aspects are included. Ramp-up and data sparsity.
	Reciprocal recommendation for recruitment	Bidirectional recommendation. Effective matching methods. Use integration-based similarity in skills matching (explicit and implicit preferences).	No relational aspects are included. Ramp-up and data sparsity.

Figure 2.8: Advantages and Disadvantages of Job recommendation Approaches

Zhang, Yang, Niu (2014) proposed a job recommender system based on collaborative filtering. Due to advent of the internet, users spent a lot of time in the search of accurate job. To overcome this challenge, to reduce laborious work of finding a suitable job, a recommendation system was implemented in this study. Model proposed was verified by using the real dataset in implementation of the algorithm. Student's resume and details of recruitment information were extracted to suggest jobs to the users. Four similarity calculation methods used to get job preferences like Cosine similarity, Tanimoto coefficient, Log likelihood, the city block distance. Collaborative filtering technique is of two basic types, User-based and item based. This paper evaluated both of these types and choose item-based collaborative filtering algorithm for its better performance. Procedure of student job hunting recommendation model are shown in figure 2.9 [6].

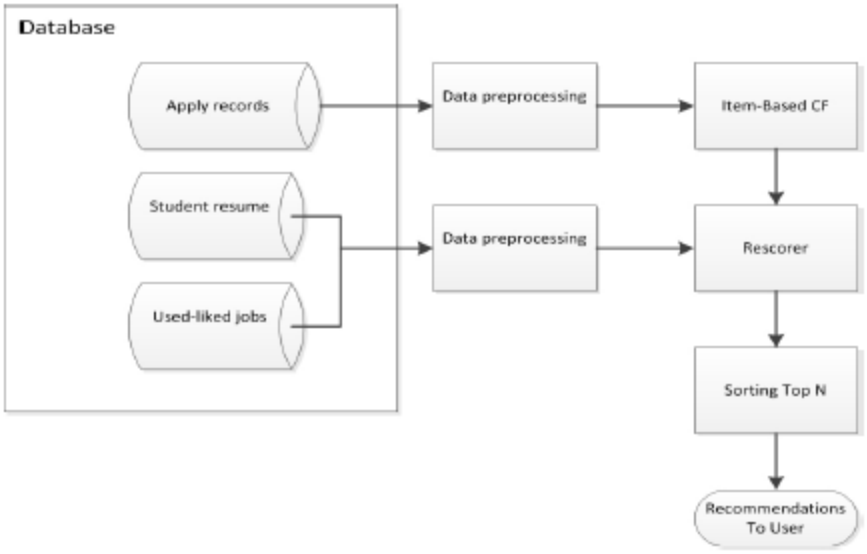

Figure 2.9: Procedure of Students Job Hunting (SJH) recommendation

Siting, Wnxing, Ning and Fan (2012) discussed in their paper that customized recommender system proposed to resolve the problem of information overload in multiple domains. Because of the rapid development in technology, organizations post the jobs and job seekers disclose their personal data on the internet. Due to increase in web technology there is a lot of personal and recruiting information available. In this way information overloaded which cause the low utilization. The recommender system handle this issue for enterprises and job seekers. The recommender system can retrieve the list of jobs which satisfy a job seeker need or meet the requirements of an organization. Authors addressed some issues that how to extract information about people and jobs and contributed this information for matching jobs. Also, they studied that which technique is suitable based on user profile in job recommender system and build a recommender system using real data with some background applications. First, they target the users with different backgrounds. For job seekers, recommender system allows to upload their resume and receive preferred job. Similarly, for recruiter's, recommender system provides a list of candidates who meet their requirements with uploaded information. Second, user profile : Procedure of Students Job Hunting (SJH) recommendation discussed in which initial profile and feedback included. Initial profile covered the user's basic characteristics and feedback relevance was about the user actions or behavior. Furthermore, hybrid model extracted general information from resume and prepare block selection and then extracted detailed information as shown in Figure 2.10. Collaborative filtering also known as user-to user correlation method used. This

technique find the same users who had same background and recommend the items which match with the target user. Multiple architectures were surveyed in this paper in which researchers' emphasized on the recommendation technology and user profile at the aim of improving performance and the accuracy [7].

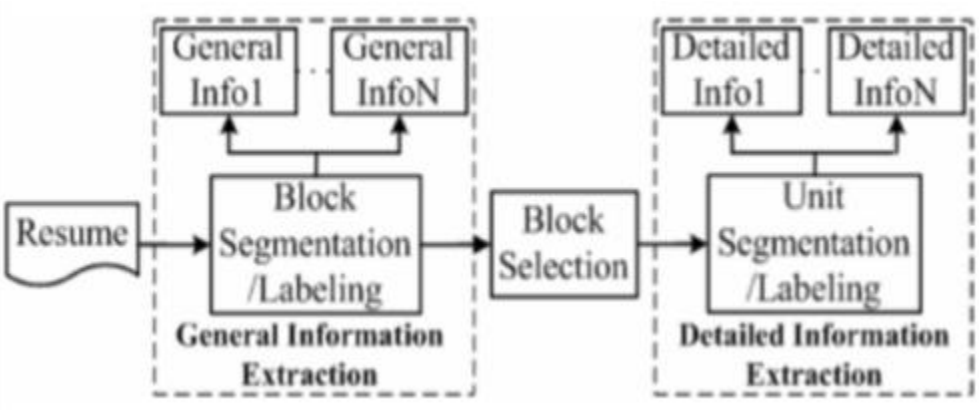

Figure.2.10: Structure of cascaded Hybrid model

Diaby, Viennet and Launay (2013) discussed a recommender system which used a content-based technique and hit the LinkedIn and Facebook users to recommend a job. This technique used by some world lead software companies that offer recruitment solution of Facebook. LinkedIn and Facebook users explicitly accessed to parts of their data. The technique match the job description with the user profile and recommend a job. The profile of a user divided into two types of data: users' personal information and users' friend's information. Furthermore, user profile divided into multiple parts called fields. The proposed study estimated fields importance of each user and jobs in the suggestion section. Experiment predict the jobs similar with the user interest. Linear SVM also trained to improve the results gained with similarity measures. This SVM yield improved results. But to make accurate suggestions it's essential to be trained on a lot of job links. There was also a problem with social network-based recommender system that fields of most Facebook users were empty. So thereby recommendations predicted to users were not accurate using their recommender system [8].

Hong, Zheng and Wang (2013) focused on those users who do not updated their user information in a specified time. For this purpose, a dynamic user profile-based recommender system presented. In this system based on the historical behavior of jobs and applied jobs they extend and update the user profile dynamically. Basic features which show users preferences were

updated after regular intervals. Extracted features were used to extend the feature numbers. The basic features of a job applicant and post are shown in Figure 2.11 [9].

Job Applicant	Job Post
Sex: Male Age: 26	Job_name: Programmer
Degree: Master	Salary: 7000
Education: Xiamen University	Location: Xiamen
Work_length: 1 year	Job_type: Full time
Need_job_type: Full time	Need_sex: Unlimited
Need_salary: 8000	Need_age: 20-35
Need_location: Xiamen	Need_degree: Master

Figure 2.11: The Basic features of a Job applicant and a Job post

Dynamic user profile contain behavior and all the information on the recruiting website that belong to the job applicant. Dynamic user profile generated by extending the extracted feature and updating the basic features. The workflow of dynamic extracted feature is shown in Figure 2.12.

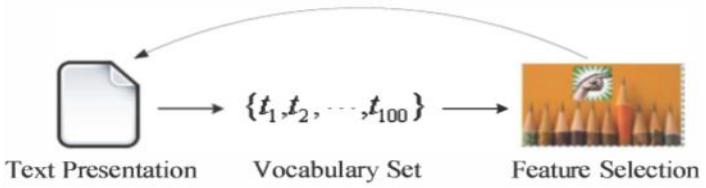

Figure 2.12: The workflow of dynamic extracted feature

Based on the dynamic user profile, dynamic recommendation achieved to improve the recommendation results. Initial recommendation was generated by using the user-based collaborative filtering. Workflow of initial recommendation is shown in Figure 2.13.

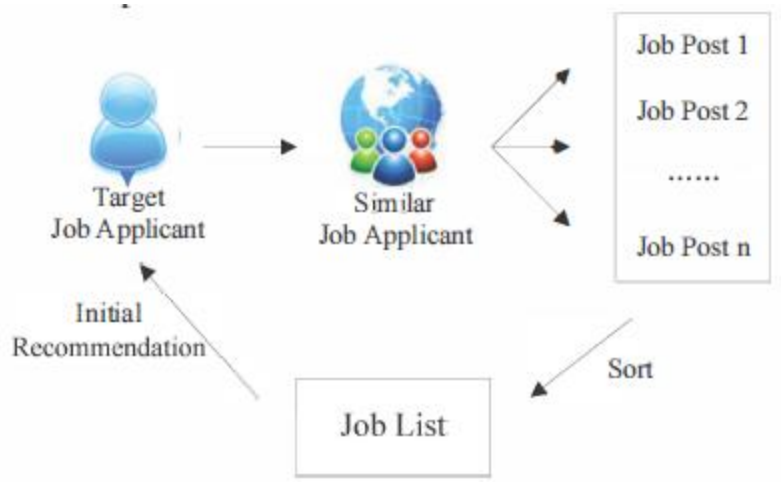

Figure 2.13: The workflow of user-based collaborative initial recommendation

In general, by combining two feature which were extracted and basic features, dynamic user profile was created, and then dynamic recommendation achieved by using the hybrid recommendation algorithm.

Recommender systems are becoming popular in academia as well as in enterprise domain. Due to rapid advancement of technology, huge amount of data slows down the utilization of relevant information for the recommendation in the academic domain.

To overcome this challenge, researchers find a way of asking the users to upload their resume initially and on the basis of user's data and behavior, jobs are recommended to the users.

Uploading the resume to recommend a job arises a problem of unstructured data, it contains unnecessary information too, which is not anymore useful for recommendation. To solve this issue FoDRA algorithm was proposed, which converted the unstructured form of data to the structured form. Candidate profile and job data recommends jobs to the users.

In traditional techniques, unstructured data used to recommend jobs and then later on it's converted into structured form through different algorithms. Now data is in the structured form but background of the user is unavailable. For recommendation to these users a rating based algorithms is presented. This algorithm worked by extracting rating from users and then calculating the similarity. Jobs recommended to those users whose ratings similarity is high.

In case of dynamic users, social network based recommender systems used to recommend jobs. It targets fb and LinkedIn users. Personal data of user and their friend's data used to suggest jobs.

Similarity measures improved through Linear SVM model. Limitation of this study showed most of the fields were empty.

To solve this issue, Dynamic user profile based recommender system discussed.

It targeted those users who didn't updated their profiles within time. Features extracted from applied jobs and used for empty fields. Jobs recommended on the basis of applied jobs.

CHAPTER 3: PROBLEM STATEMENT AND APPROACH

3.1 Problem Statement

With the advent of the internet era, everyone has access to the overwhelming information available on the World Wide Web (www). Internet users have seldom idea of relevant information about their needs. To overcome this problem of information overload, recommender systems are becoming more popular nowadays for almost all fields. Netflix and amazon give us the best examples of recommender systems.

The most challenging task is to suggest a recommender system for engineering students in the education domain to give them ease for career selection by using machine learning techniques. This recommender system will help optimizing the efforts, which they have to do for searching and applying for a suitable job according to their skills.

3.2 Approach

3.2.1 Machine learning

Machine learning is the subset of AI (artificial intelligence). It provides us statistical tools to explore and understand the data and focuses on prediction models through the use of computers. Machine learning has three different approaches:

- Supervised learning
- Unsupervised learning
- Reinforcement or semi supervised learning

Supervised machine learning technique deals with labeled data and based on learning from this labeled data, they would be able to do some prediction for the future.

In unsupervised learning, model works with unlabeled data and unaware of the output of this data initially. Usually unsupervised learning solves clustering type of problems like k-means clustering, hierarchical clustering. Clustering means, based on the similarity through mathematical concepts like Euclidean distance etc. it will grouped the data together. Reinforcement or semi supervised learning is the combination of supervised and unsupervised learning. Reinforcement learning trains models on reward and punishment basis.

3.2.2 Deep Learning

Deep learning is the subset of machine learning, which uses algorithms that are inspired by the working of brain. Usually we have a lot of neurons in our brain that is responsible for passing the signals from one part of the brain to the other and finally getting the output which is basically understanding the information. Basic idea behind deep learning is to mimic human brain.

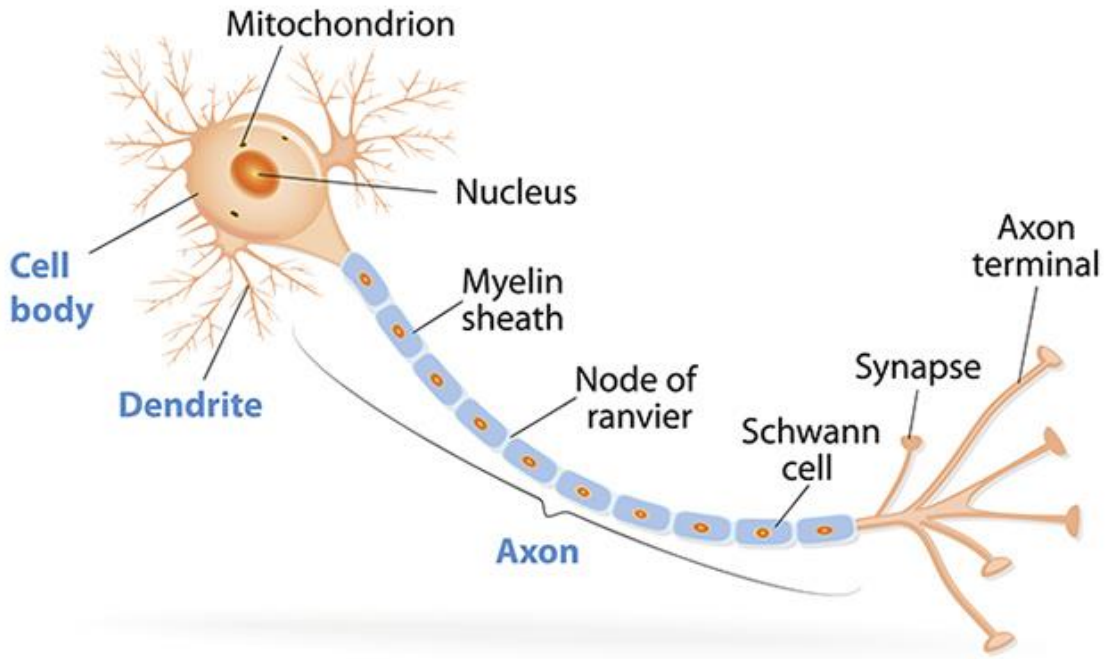

Figure 3.1: Structure of the neuron

There are various deep learning architectures. We have basically deployed the following two:

- Artificial neural network (ANN)
- Convolutional neural network (CNN)

3.2.2.1 Artificial neural network (ANN):

Artificial neural network consists of an input layer, a hidden layer and an output layer. Input layer is basically the layer where we provide the input features for example images of a cat and dog. And model will classify the input data whether it is a cat or dog image at the output. Similarly there is a hidden layer which is splitted into various neurons. Information from the input layer is passed to the hidden layer and from the hidden layer it is transferred to the output layer.

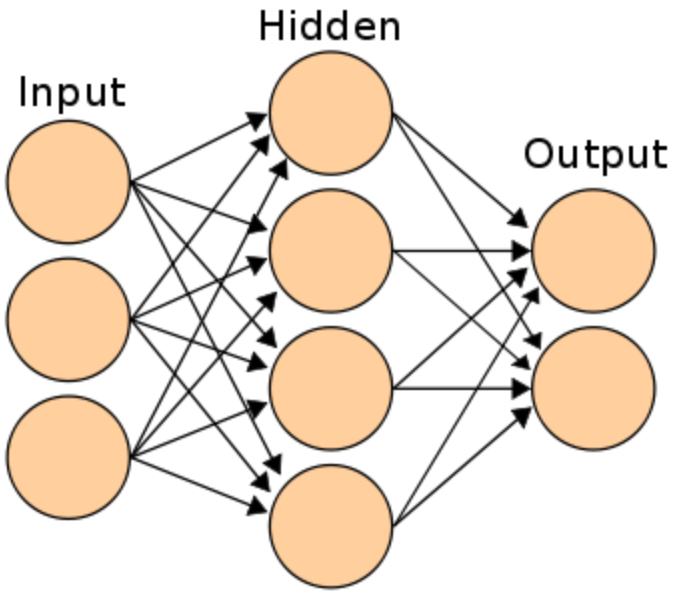

Figure 3.2: an artificial neural network

3.2.2.2 Convolutional neural network (CNN):

Convolutional neural network usually works with image inputs to make them more simplified so it can be better processed and understood. A convolutional neural network made up of multiple layers like convolutional layer and pooling layer etc. CNN learns by the back propagation algorithm.

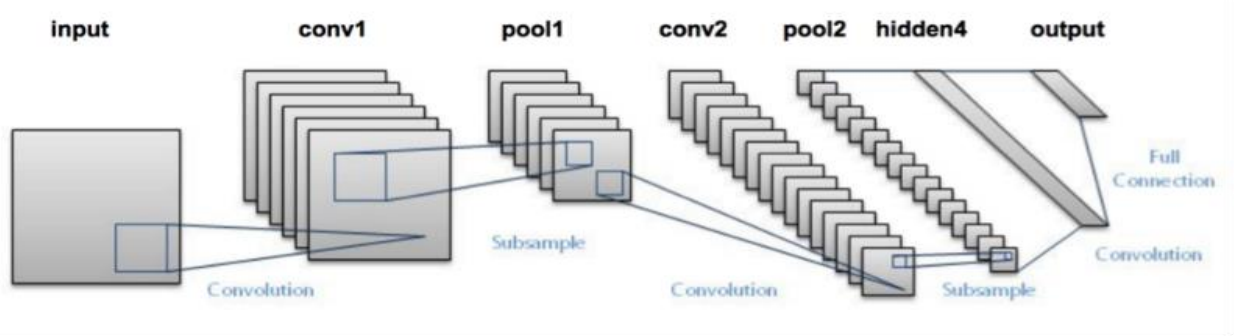

Figure 3.3: a convolutional neural network

3.2.3 Python

Python is the currently the most vastly used programming language for research and development in machine learning. Recently the google trend from the past five years has shown that the interest in python programming language has gone to a whole new level when compared

to other programming language like Java, Julia, R etc. and also google trends showa python is the most popular machine learning language. It was created by Guido van Rossum during 1985-1990.

Python has following features:

Python is easy to use: The syntax of python is very simple. If you have basic knowledge of any programming language, you will be able to learn it very quickly within two to three months. And it uses very less lines of code for implementation.

It supports many libraries and frameworks: As python is an open source programming language. It has hundreds of libraries along with different types of frameworks which the developers can use to basically complete their work. These libraries are NumPy, Pandas, scikits learn. NumPy is basically used for creating multidimensional array. Pandas library help you to read from different kind of data resources so that you can retrieve the data and then with the help of data frames you can do different kind of operations. Scikits library has all the types of algorithms both supervised and unsupervised. Python is also being used for web services.

Community and corporate support: This basically means many developers are using Python skills to basically increase the machine learning knowledge apart from that they are using Python for implementing various deep learning projects, this is basically leading to the increase in the popularity of python even though if you go and search in the google you'll be finding various GitHub links many online repositories and much more online learning resource apart from youtube

3.2.4 Anaconda Tool

Anaconda is basically both a package manager and it allows you to also manage your virtual environment so to create and manage these virtual environments on your machine. The normal python pip is virtual and approach is basically made for any purpose, you can create environments user packages but anaconda come with a lot of packages required for data analysis and data science or at least these packages make working in these fields a lot easier and they also make it easier for beginners to get started with python. Anaconda comes with Python 2.7 or Python 3.4.

3.2.4.1 Python usage in anaconda

For python coding, many people use Emacs or Vim as text editors. And for visual studio rest of the users prefer to use an IDE spyder or python tools. Jupyter notebook is a free IDE having Anaconda embedded.

3.2.4.2 Natural language processing in Python

Natural language processing deals with processing and inferring from text data. NLP falls under the broader domain of artificial intelligence. Artificial intelligence (AI) is about a computer performing tasks that a human can do so that includes how humans interpret language. NLP deals with how can we make the machine interpret language. To start NLP in python the most popular library natural language toolkit (NLTK) is used. NLTK is written in python. Its an efficient Library for machine learning.

3.3 System Requirements

3.3.1 Hardware Requirements

- 8 GB RAM

3.3.1 Software Requirements

- Dataset in CSV format
- Windows OS
- Python 3.7

CHAPTER 4: DATASET AND IMPLEMENTATION

In this chapter we describe the dataset, the specific implementation details of the proposed model.

4.1 Dataset Description

The data collection phase consisted of two main steps, one is collection of job data having details of the jobs from different engineers already working in different sectors of the country like industries, companies, universities, research areas etc. and second step is user data collection containing user information and description of those students who are almost graduating from different engineering universities.

4.2 Data Set regarding Job Profiles

Data regarding job was collected by conducting a survey from different job holders, doing jobs in different companies, industries and academia like universities etc. Job data was extracted mainly from these Engineering categories:

- Electrical / Electronics
- Mechanical
- Computer / Software
- Mechatronics

The raw data which was obtained from the survey was converted into a CSV format to perform modifications on it. The list of attributes in the job dataset are follows:

Company name: Name of the company / industry / institute in which the employee is working.

Employment status: Status of the job either it is part-time, full-time, contract basis or un-employed

Unique-job-id: An id is assigned to each job uniquely to make all jobs different from each other.

Job-description: Explanation for each job in text form.

Location: Place of the company / institute etc.

Skills: Qualities of the employees

Job Experience: Number of years the job holder has been working in the company

Satisfaction: user's level of satisfaction in the job

4.3 User Dataset

User dataset show the personal and academic details of the users. It was collected by doing a survey from different engineering candidates of various universities. The four major engineering disciplines were mainly involved in this survey. The raw data extracted from the survey was later on converted into a CVS format for processing in Python. It contains almost 10,000 entries. Some of the important features of the user dataset are as follows:

Qualification: formal education of the candidate either it is Bachelor's degree, Master's Degree or PhD.

Discipline: Engineering field either it is Electrical/ Electronics, Mechanical/ Computer/Software, Mechatronics

Major Subjects: subjects of undergraduate degree

GPA: Grade point average of the candidate

Languages: List of languages in which the candidate has worked

Frameworks: Frameworks in which the candidate has worked

Platforms: Platforms in which the candidate has worked

4.4 Data preprocessing

Whenever the data is collected from different sources it has some abnormalities, contains inconsistency and has missing values. To make it feasible before feeding into an algorithm preprocessing is done .Preprocessing the data is an important task for the accuracy.

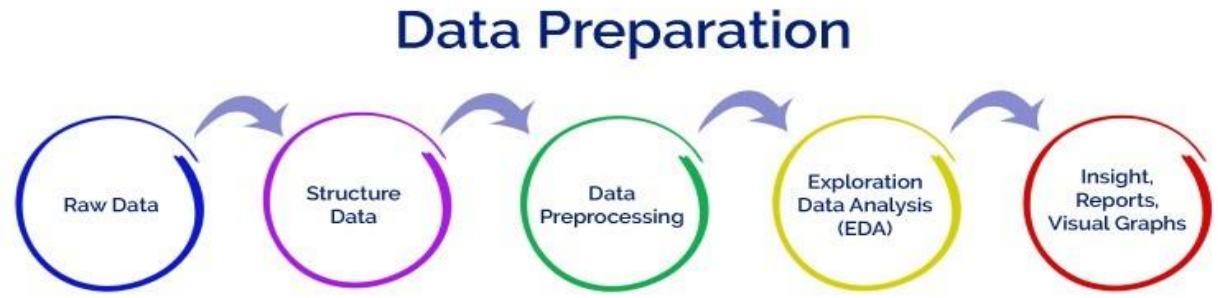

Figure 4.1: Data preprocessing steps

4.4.1 User data cleaning

Data cleaning was performed for user data. In User data some columns having attributes like name, gender, Contact details, hobbies, age etc. were removed as they are not useful in the further process.

4.4.2 Job data cleaning

In order to clean job data, NLTK packages were used in the python. The job attributes like job description and skills are mostly in the form of text therefore, numeric characters and stop words form the entire text were removed by using NLTK packages.

4.5 Exploration Data Analysis

Exploratory data analysis or EDA is a data exploration technique to understand the various aspects of data. While exploring the data it make sure that data is clean and does not have any redundancies or missing values or even null values in the dataset, and we identify the important attributes in the dataset and remove all unnecessary attributes that may actually hinder the accuracy of our conclusions when we work on model building, we must understand the relationship between the attributes through EDA. The main objectives of EDA are to identify false attributes, removing them and the next objective is that EDA helps us to understand the relationship between attributes which gives us a wider perspective on data. To perform EDA, python provide us a lot of libraries. We will use Pandas and Matplotlib for performing EDA, Pandas will work on data manipulation and by using Matplotlib graphs will be plotted.

Many patterns have been obtained when exploratory analysis executed on the job dataset. Following points extracted from the analysis:

- Satisfaction level of each job
- Count of each job in the company
- Interests of the Job holder

Exploratory analysis is also performed on the user dataset and it provides the following points:

- Qualification of the users
- Employment status of the user
- Top languages of the users

4.6 Feature Extraction

Feature extraction builds a set of derived values from a set of measured data. Derived values intended to be non-redundant and informative. Feature extraction enhance the speed of supervised learning leading to better human evaluation. Feature extraction usually transform the attributes. It also reduces the dimensionality of the raw data and converts it into more managed featured data for further processing.

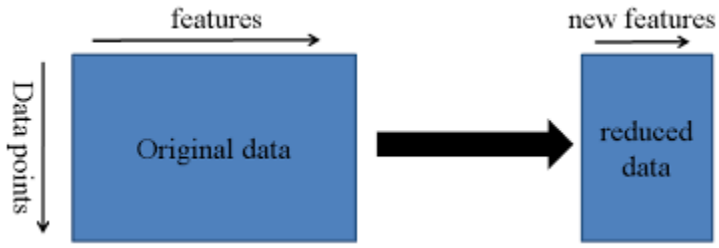

Figure 4.2: Feature Extraction

4.6.1 Feature Extraction of User Data

User dataset has many attributes like Domain, frameworks, languages, Databases, development type etc. In order to get unique values from above mentioned attributes feature extraction is performed on the user dataset mentioned above in section 4.3. All the unique values of each attribute are represented in a list. Categorical attributes like employment status (employed/unemployed) are converted into numerical values with the help of LabelEncoder python package. We created a CSV file and store extracted features for each attribute separately.

4.6.2 Feature Extraction of Job Data

After preprocessing of the job dataset. Two steps were performed for feature extraction from the job data.

- Categorize the jobs using some predefined categorize. An advanced NLP package named as spacy provides word embedding. Word embedding mapped vocabulary based words or phrases into real number vectors
- Extract the features from the job description based on user data extracted features. Now by using these features matching jobs were found.

4.7 Implementation

4.7.1 Content based Career Recommender System

In the content based career recommender system the engineering students or graduates are considered to be the term user and career/job is considered to be an item in the filtering technique. The content of the item i.e. description of the job and skills of the job holders are compared with the skills mentioned in the profile of engineering candidates/graduates. Now the matching is performed between engineering students and career options/job by using cosine similarity.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

In the result of cosine similarity, those career options/jobs (of the job holders already working in different sectors) which have high score against user's skills will be recommended to the engineering candidates.

Table 4.1: Job Feature Matrix

	Domain	Language	Framework	Platform	Database
Network Engineer	Network System	Python, Perl	Flask, Flex	SharePoint	Greenplum
Mobile Developer	Mobi	Java, C#	Ionic, Corona	Microsoft Azure	MySQL
Electrical Engineer	Design, Development	C++, Arduino	Power eSim	Orcad	Knovel
Database Administrator	Information System	SQL	Oracle	Cross Platform	Clipper
Mechanical Engineer	Biomedical, Transportation	MATLAB	Production Planning	CAD drawing	FoxPro

Table 4. 2: User Feature Matrix

	User Domain	User Language	User Framework	User Platform	User Database
User ID 1	Network System	Python, Perl	Flask, Flex	SharePoint	Greenplum
User ID 2	Mobi	Java, C#, MATLAB	Ionic, Corona	Microsoft Azure	MySQL
User ID 3	Design, Development	C++, Arduino	Power eSim	Orcad	Knovel
User ID 4	Information System	SQL	Oracle	Cross Platform	Clipper
User ID 5	Biomedical, Transportation	MATLAB, C++	Production Planning	CAD drawing	FoxPro, MySQL

Job feature matrix and user feature matrix will help out in finding of similarity by using cosine similarity.

Similarity= (0.7*cosine_similarity (Domain, User Domain)) + (0.3*(cosine_similarity (Language, User Language) + cosine_similarity (Framework, User Framework) + cosine_similarity (Platform, User Platform) + cosine_similarity (Database, User Database)))

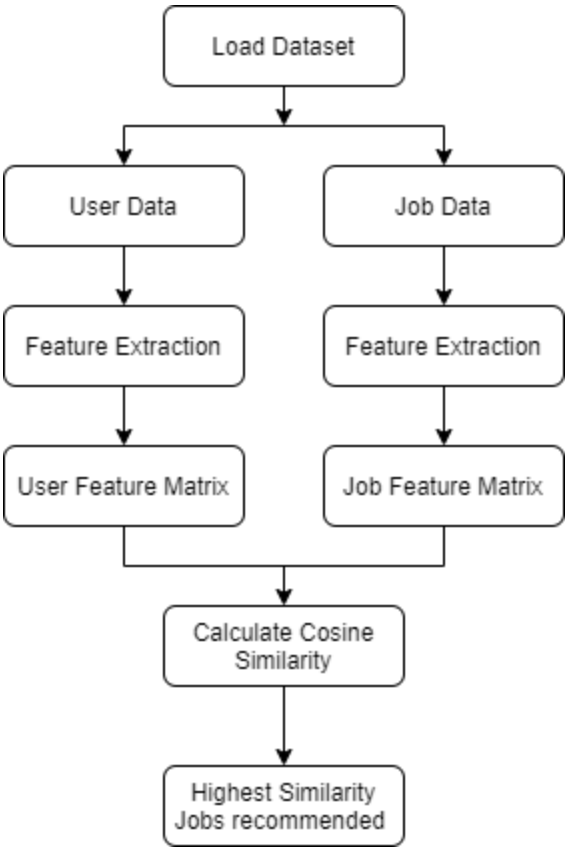

Figure 4.3: Content based Career Recommender System

4.7.2 Basic architecture of the Deep NLP based network

Convolutional Neural networks used for text processing generally takes a sentence as an input. Models like word2vec or Glove used to get low dimensional representation of input. Sentence is then divided into words by using word2vec model. Features extracted from these words and then passed to a convolutional layer. The results obtained from convolutional layer fed into pooling layer, by using max pooling to get a maximum representative number. This maximum number is then passed to a fully connected neural network. This network assign weights to each feature of the text. Finally Neural network makes prediction/ classification decision based on these weights. Basic architecture of the Deep NLP based network is shown is figure 4.4

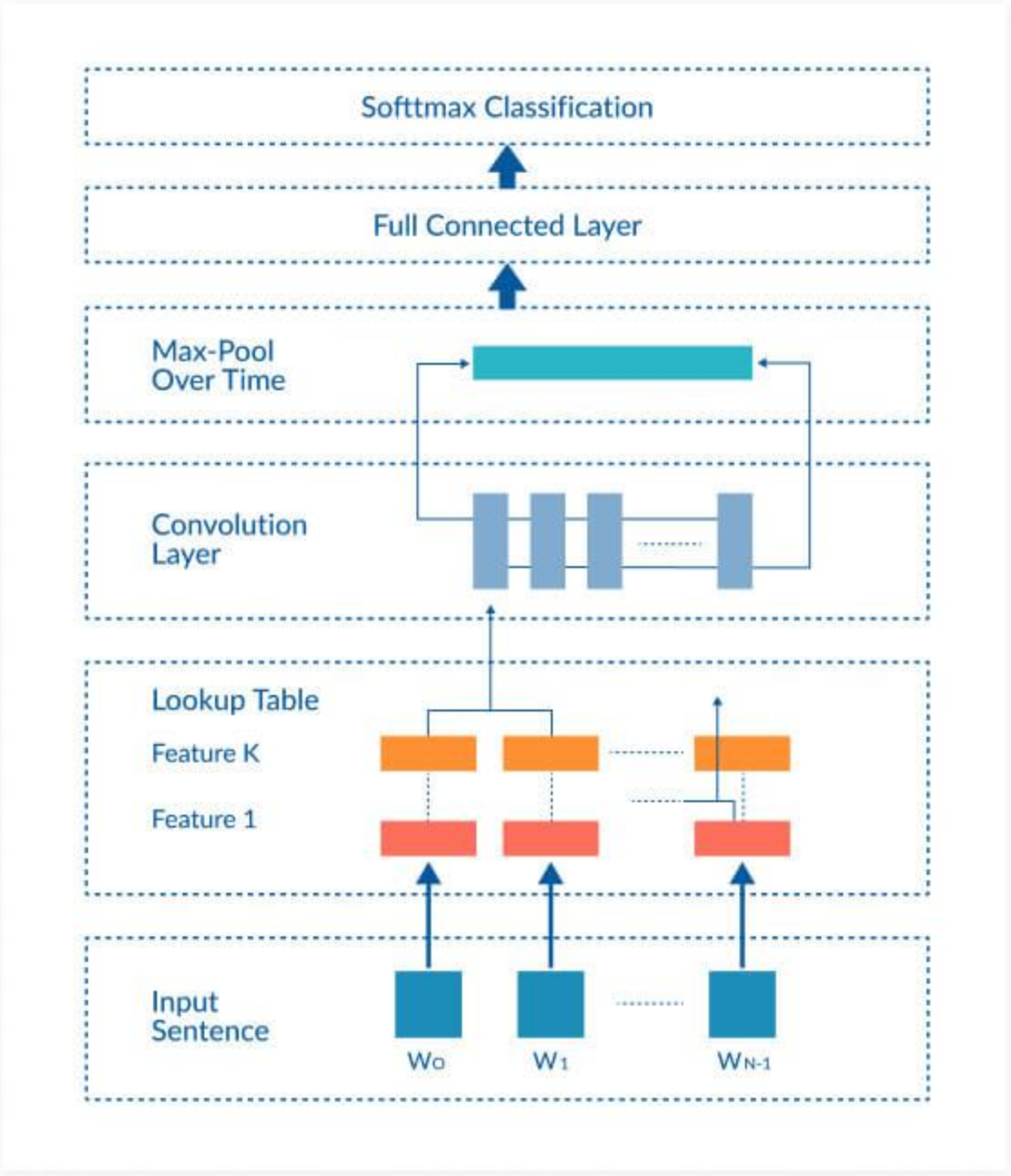

Figure 4.4: Basic architecture of the Deep NLP based network

4.7.3 Deep NLP based Career Recommender System

Due to the advances in the machine learning and the artificial neural networks many applications in the natural language processing folds have also advanced. A natural language processing, computer vision all these fields have been the ones that were traditionally attributed to human perception and were quite hard to solve using the more traditional machine learning techniques .

So due to this advances in the field during the past couple of years so many tasks have been actually solved or we have the breakthroughs in them so that's the basic reason behind making of career recommender system using deep NLP technique CNN through which different career options will be recommended to engineering candidates.

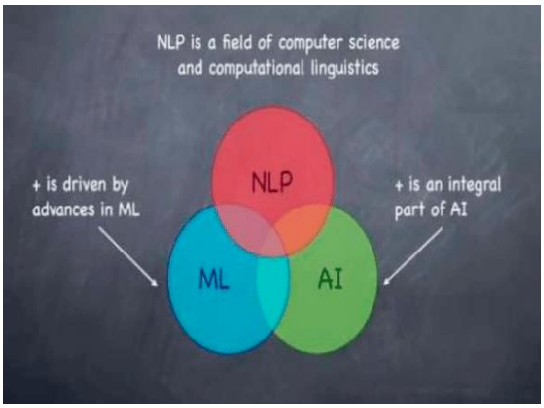

Figure 4.5: Deep NLP

Deep NLP based Career Recommender System have following these steps:

Load the Dataset. Split the dataset into training and testing data.

X_train = train ['user_profile'], Y_train = train ['Jobs'], X_test = test ['user_profile'], Y_test = test ['Jobs']

Define model parameters. Parameters involve, Vocab_size = 1000 this will create a dictionary consisting of 1000 words. Sequences length = 1200, this indicates actually user profile in which there are 1200 words. Embedding Dimensionality = 64, it represents feature matrix dimensionality. Max_features = 2000, maximum number of features are 2000. Num_lables = outcome, Lables represents jobs. Batch size = 32 ,Batch size means in one iteration 32 training samples will be passing through the network at a time. Num_Epoch = 2000 to update the weights of the training samples, 2000 times training will be carried out in this model. Num_filters = 200, 200 filters will be used in the convolution layer. Kernel_size = 16, Convolutional filter will be of dimensionality 16. Hidden_dims = 512, 512 hidden Layers will be involved in the network.

Apply tokenization on user profile. Tokenization is a method of converting each and every word in the user profile to their respective indexes present in the vocabulary. Apply Pre/Post pad sequencing. This will equalize the length of each sentence in the user profile by

appending zeros. Initialize sequential model, sequential means all the layers of CNN model will be connected sequentially in the network.

Add Embedding layer and its parameters (Vocab_size, Dimension, Sequences length) in the network. Add convolution layer of 1D and its parameters (Num_filters, Kernel_size and Activation Function). 'Relu' activation function will be used in the model as it will reduce the linearity from the user profile. Add pooling layer. This layer will use GlobalMaxPooling function to extract the maximum value from the vector and will reduce the vector size and as a result pooled feature map will be retrieved.

Add 512 hidden layers in the network, dropout of 0.3 and activation function 'relu'. Add output layer i-e labels. Apply 'softmax' function to the labels to distribute their probabilities uniformly.

Compile the model. Generate the model's summary. Top jobs/ career options will be recommended.

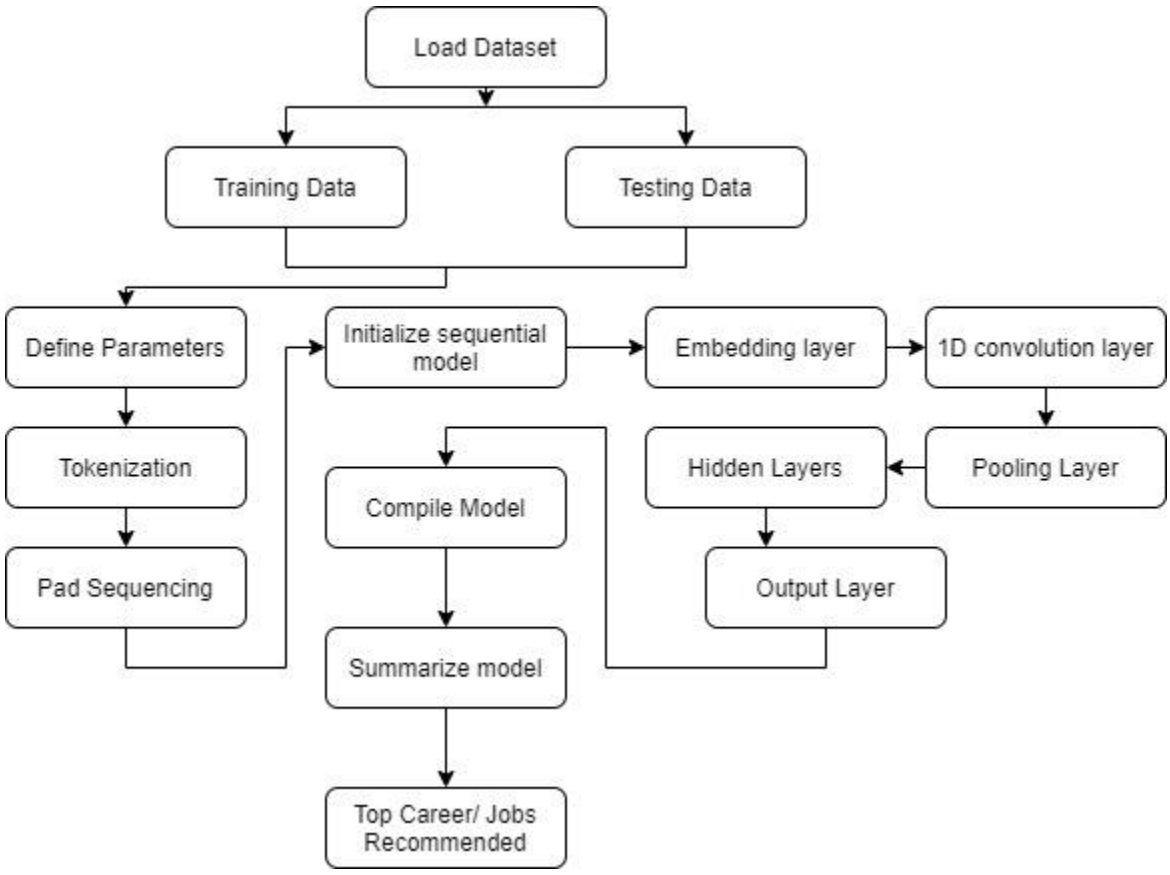

Figure 4.4: CNN based Career Recommender System

CHAPTER 5: RESULTS AND COMPARISON

5.1 Results

Our Proposed model CNN based career recommender system suggests different career/job options to engineering graduates based on their skills by using deep learning technique as compared to traditional content-based filtering recommender systems.

Figure 5.1: Result

The figure 5.1 demonstrates the web-based application which gives the user an interface through which he/she can enter the text data.

There are four engineering disciplines are given. User has to select one of them.

After selecting the engineering field, major subjects of that engineering discipline are listed down. Relevant subjects are then selected according to the discipline as shown in Figure 5.1. Gender information is provided.

Figure 5.2: Result

The Figure 5.2 demonstrates the academic qualification of the user is asked if he/she has done B.Sc/MS or Ph.D. GPA of his/her last degree program is provided by the user. The interests are also being given like sports, book reading, etc. as shown in Figure 5.2.

Click the submit button.

Figure 5.3: Result

The best job recommendation as shown in Figure 5.3 is now at your hands. This framework produces the job role recommendation according to the skills required for a particular job.

For better performance evaluation, same dataset is implemented on traditional content based filtering algorithm. Content based filtering technique recommends jobs to the candidates based on the cosine similarity calculations between user and job profiles.

```
Respondent 3 was recommended jobs from content based filtering in
['ICONMA' 'Cypress Group' 'SolTech, Inc' 'Saicon Consultants Inc.'
 'Samsung SDS America Inc' 'Strivector' 'Aria Systems' 'Aria Systems'
 'Strivector' 'Universal Software Corporation']
Respondent 3 was recommended job titles from content based filtering
['Python Developer' 'Full-Stack Python Developer'
 'Full Stack Javascript Engineer' 'Full Stack Developer'
 'System Administrator' 'System Administrator' 'Application Administrator'
 'Application Administrator' 'System Administrator' 'Full Stack Engineer']
```

Figure 5.4: Result

The figure 5.4 demonstrates content based job recommendations to user ID 3. First three rows indicates top 10 companies for recommended job titles. Last three rows shows top 10 job titles recommendations to user ID 3.

5.2 Comparison

The quality of a recommendation algorithm can be evaluated using different types of measurement which can be accuracy or coverage. The type of metrics used depends on the type of filtering technique. Accuracy is the fraction of correct recommendations out of total possible recommendations. The suitability of each metric depends on the features of the dataset and the type of tasks that the recommender system will do.

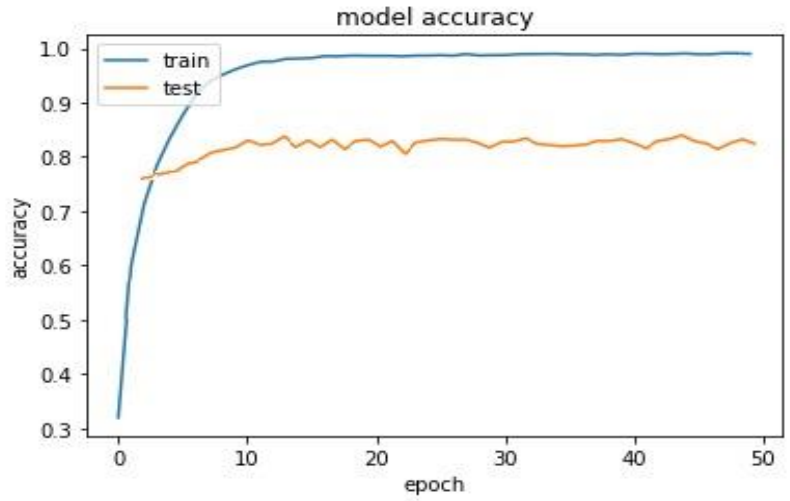

Figure 5.5: CNN model Accuracy

The Figure 5.5 demonstrates accuracy of CNN based career recommender system. X- Axis shows no. of epochs and accuracy is shown on y-Axis.

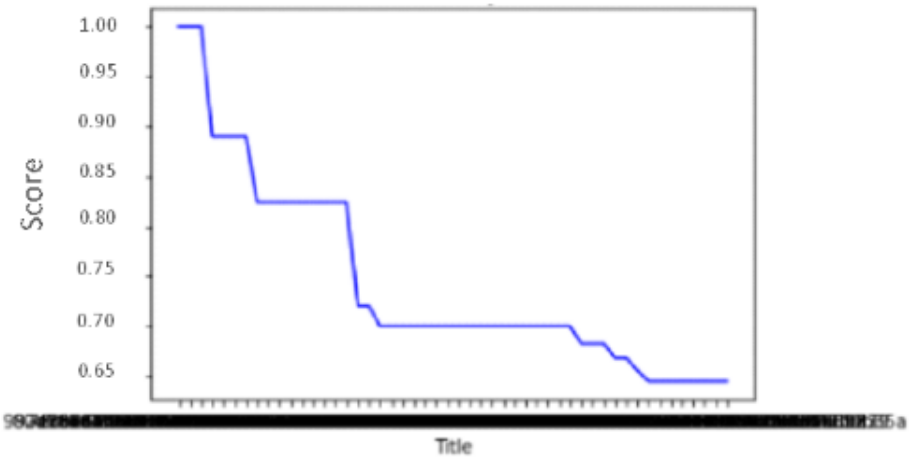

Figure 5.6: Content based filtering model Accuracy

The figure 5.6 shows the similarity accuracy of content based career recommender system. X-axis shows the job titles and y-axis shows Cosine Similarity b/w User and job features.

A comparison of different techniques discussed in Table 5.1.

Table 5.1: Comparison of Different Techniques

Models	Predictor	Accuracy
Proposed Model	CNN based Career recommender system	84 %
Comparative Model	Content based career recommender system	81 %

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Conclusion

In the past couple of years, we have seen a big change in the recommendation domain which shifted from traditional matrix factorization algorithms to state-of-the-art deep learning-based methods. Recommender systems provide easiness to each individual because it predicts suggestions based on personal users interest. Our proposed approach comes up with suggestions about jobs for engineering candidates.

Engineering graduates find a lot of difficulty in terms of the best possible career according to their qualifications and interest. They don't know much about their relevant available jobs. They have to drop CV's at different places to get hired and this process requires a lot of time and cost. To overcome these problems we develop a recommender system named as career recommender system (CRS). We use different parameters such as Engineering Discipline, Gender, Qualification, GPA, interest and skills which will provide the best suitable job according to the individual.

6.2 Future Work

Our proposed technique covers only four to six engineering disciplines. In the future, we will enhance the scope for our CRS which will include other engineering disciplines as well as other educational areas like medical, art, etc. Currently, this technique was only applied to a single dataset. To show generalizability, it is desirable to replicate this research on further datasets with different properties. Due to the ease of prediction and potential upside in this task, it is a valuable extension.

REFERENCES

[1] Karbhari, N., Deshmukh, A. and Shinde, V.D., 2017, August. Recommendation system using content filtering: A case study for college campus placement. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 963-965). IEEE.

[2] Almalis, N.D., Tsihrintzis, G.A., Karagiannis, N. and Strati, A.D., 2015, July. FoDRA—A new content-based job recommendation algorithm for job seeking and recruiting. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1-7). IEEE.

[3] Paparrizos, I., Cambazoglu, B.B. and Gionis, A., 2011, October. Machine learned job recommendation. In *Proceedings of the fifth ACM Conference on Recommender Systems* (pp. 325-328).

[4] Liu, R., Ouyang, Y., Rong, W., Song, X., Tang, C. and Xiong, Z., 2016, July. Rating prediction based job recommendation service for college students. In *International conference on computational science and its applications* (pp. 453-467). Springer, Cham.

[5] Al-Otaibi, S.T. and Ykhlef, M., 2012. Job recommendation systems for enhancing e-recruitment process. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)* (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[6] Zhang, Y., Yang, C. and Niu, Z., 2014, December. A research of job recommendation system based on collaborative filtering. In *2014 Seventh International Symposium on Computational Intelligence and Design* (Vol. 1, pp. 533-538). IEEE.

[7] Hong, W., Zheng, S. and Wang, H., 2013, April. Dynamic user profile-based job recommender system. In *2013 8th International Conference on Computer Science & Education* (pp. 1499-1503). IEEE.

[8] Siting, Z., Wenxing, H., Ning, Z. and Fan, Y., 2012, July. Job recommender systems: a survey. In *2012 7th International Conference on Computer Science & Education (ICCSE)* (pp. 920-924). IEEE.

[9] Diaby, M., Viennet, E. and Launay, T., 2013, August. Toward the next generation of recruitment tools: an online social network-based job recommender system. In *2013*

IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013) (pp. 821-828). IEEE.

[10] Aggarwal, C.C., 2016. An introduction to recommender systems. In *Recommender systems* (pp. 1-28). Springer, Cham.

[11] Gedikli, F., Jannach, D. and Ge, M., 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, *72*(4), pp.367-382.

[12] Shah, K., Salunke, A., Dongare, S. and Antala, K., 2017, March. Recommender systems: An overview of different approaches to recommendations. In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* (pp. 1-4). IEEE.

[13] Bobadilla, J., Ortega, F., Hernando, A. and Gutiérrez, A., 2013. Recommender systems survey. *Knowledge-based systems*, *46*, pp.109-132.