Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020)
IEEE Conference Record # 48766; IEEE Xplore ISBN: 978-1-7281-5371-1

Career Recommendation Systems using Content based Filtering

Tanya V Yadalam
Undergraduate Student
B.M.S. College of Engineering
Bengaluru, India
yadalamtanya@gmail.com

Vaishnavi M Gowda
Undergraduate Student
B.M.S. College of Engineering
Bengaluru, India
vaishnavi1997@gmail.com

Vanditha Shiva Kumar
Undergraduate Student
B.M.S. College of Engineering
Bengaluru, India
vandithashiv97@gmail.com

Disha Girish
Undergraduate Student
B.M.S. College of Engineering
Bengaluru, India
dishagirish98@gmail.com

Namratha M
Assistant Professor
B.M.S. College of Engineering
Bengaluru, India
namratham.cse@bmsce.ac.in

Abstract—Machine learning is a sub-field of data science that concentrates on designing algorithms which can learn from and make predictions on the data. Presently recommendation frameworks are utilized to take care of the issue of the overwhelming amount of information in every domain and enables the clients to concentrate on information that is significant to their area of interest. One domain where such recommender systems can play a significant role to help college graduates to fulfil their dreams by recommending a job based on their interest and skillset. Currently, there are a plethora of websites which provide heaps of information regarding employment opportunities, but this task is extremely tedious for students as they need to go through large amounts of information to find the ideal job. Simultaneously, existing job recommendation systems only take into consideration the domain in which the user is interested while ignoring their profile and skillset, which can help recommend jobs which are tailor made for the user. This paper examines existing career recommendation system and highlights the drawbacks of these systems, such as cold start, scalability and sparsely. Furthermore, proposed implementations of career recommendation system using machine learning have been researched in order to identify how the recommender systems introduce features of security, reliability and transparency in the process of career recommendation. In addition, possibilities for improvements in these systems have been explored, in order to design a career recommendation system using the content based filtering approach.

Index Terms—natural language processing, cosine similarity, content based filtering

I. Introduction

At present the proposal frameworks are utilized to tackle the issue of data overload in numerous regions enabling clients to concentrate on significant data dependent on their advantage. One of the zones where such frameworks can assume a significant job is in helping understudies accomplish their vocation objectives by creating customized employment and aptitude suggestions. This is primarily done using collaborative filtering where automatic predictions are made about the interests of a user by collecting preferences from many users. Collaborative filtering, also referred to as social filtering, uses algorithms to filter data from user reviews to make personalised recommendations for users with similar preferences. The content of each product is represented as a set of characteristics or terms, ideally the words that occur in a document. The client's profile is represented with the same terms and is created by examining the data of the items which have previously been seen by the user.

During the most recent couple of decades, recommender frameworks have become a critical piece of our lives. These frameworks are calculations that are planned for recom-mending pertinent things to the clients. One of the regions where such frameworks can assume a significant job is in helping understudies accomplish their vocation objectives by producing customized occupation and expertise proposals. The expectation is to actualize a vocation suggestion framework which enables the understudies to pick the profession most appropriate for them. The framework additionally enables clients to rate the employments dependent on their experience on various parameters. These methods help the clients settle on increasingly educated choices. Existing employment suggestion frameworks just think about the client's field of interest, yet don't contemplate the client's profile and aptitudes, which can produce progressively applicable vocation proposals for clients. In this report, the proposed system is a Career Recommendation system, which tends to such deficiencies. Utilizing ML and content based filtering methods the framework first outputs the client's profile and resume, recognizes the key abilities of the up-and-comer and produces customized work proposals.

Along these lines, the framework not just enables its clients to look into a lot of data, yet in addition extend their portfolio and resume have the option to propel their careers further. The system also allows users to rate the jobs based on their expe-rience on different parameters. These reviews help the users make more informed decisions. The existing solutions lack in the use of students' skills, interests and academic performance towards filtering a possible recommendation.[1][2]

II. LITERATURE SURVEY

A recommender system algorithm is created to help both client and companies in the many different ways. Methods used are Content based filtering dependent on a correlation be-tween the content of the jobs and a client profile. The content of every job is presented as a lot of descriptors or terms, ideally the words that are present in a report Collaborative filtering: It is an algorithm that helps make automatic predictions depending on the preferences of the client which are gathered from their likes and dislikes and client information or taste data from numerous clients Hybrid filter-ing: It is the mix of the possibility of both collaborative filtering and content-based filtering to give viable proposals of related The principle advantage of utilizing Hybrid Recommendation framework is that the precision is very high. The reason is mainly due to the absence of data about the domain dependencies in collabora-tive filtering, and the people's choices in content-based system. Hence, the amalgamation of both leads to common data increase, which helps enhance the recommendations.[3][4] Recommender frameworks (RS) are utilized to assist clients with finding new products or services, for example, books, music, transportation or even individuals or the suggested thing, depending on the data collected about the client, a similar approach is used here to find job recommendations according to the users profile. collaborative filtering approach is implemented in this paper. Collaborative filtering is mainly used when recommendation of jobs is done to users based on similar user profiles with a similar skill set. Content based filtering on the other hand is used when the user specifically states that he wants a job for a particular sort.[5] In order to get sufficient data about the user there is a need to gather data about them explicitly through a form. Other data can also be collected implicitly by monitoring their activity on the application. Hence data collection can be done explicitly or implicitly. Explicit user data gathering happens when clients know that they are giving you with information by actively filling forms about their details and preferences. For example, while enrolling for online assistance, clients for the most part fill in a form that asks their details. In this paper a similar technique is used by asking the users to create a profile on website.[6] In the Content based Filtering algorithm – the very first task is to obtain similarity evaluation. A plan to use the Tanimoto coefficient method. It gives the similarity between two sets. It is a ratio of intersections. Consider that set M is A,B,C and set N is X, Y, Z. The Tanimoto coefficient T of two set M and N is 0.6. This value does not take into consideration the user feedback, but in the case of a limited data set it is efficient. After obtaining the similarity and evaluating it, then generating closest neighbourhoods. To improve execution, numerous strategies have been proposed. The strategies for choosing closest neighbourhoods incorporate arrangement utilizing K-means, a threshold for the number of similar rating items and a graph algorithm is used. In general, it selects users with a similarity higher than a given threshold. K-means clustering is used to group similar items.

K- means is a clustering has a broad range of applications in data mining, statistics and machine learning. The input to k-means is the pair-wise distance between the items to be grouped, where the distance means how different the items are. The number of groups, k is also one of the parameters involved in clustering or grouping. It is a looping algorithm and starts with a randomly dividing the items into k different groups. To search for other users who have similar interests as the user, the models partitions the customer base into various groups and treats this as a classification problem. The algorithm's main objective is designate the user to the group having the most similar user profiles. [7][8] The proposed recommendation system is based on content based filtering. It has following steps: Data preparation, Similarity coefficient using Cosine similarity, Clustering using K means algorithm, Web page recommendation The main drawbacks of collaborative filtering are given below.

Cold start is a case where it is hard to offer recommen-dations to fresh users as their profiles haven't been created and they haven't evaluated any products yet, so the system is not familiar with their taste. This complication is solved with surveys while generating a fresh profile. Things can have a cold start when they're new to the system and haven't been evaluated previously. Therefore, when a user newly joins the career recommendation system, do not know anything about his preferences or his feedback. So, base our recommendations entirely on his information entered in the survey. The issue of trust originates towards evaluations of a specific customer.The opinions of users with a short history may not be considered that important as the opinions of those who have rich history in their profiles.This problem can be addressed by distribution of priorities to all users. This is overcome by using collaborative filtering which bases the job recommendation using the users own profile. Protection has been the most significant issue. So as to get the exact and accurate recommendation, the system must gather all the information about the user, including demographic data, and data about the location of a specific user. Normally, there are concerns about the reliability, security and confi-dentiality of the given information. Numerous online shops offer protection of privacy of the users by using particular algorithms and programs. This dilemma is overcome using collaborative recommendation. Content-based filtering main disadvantage is Content analysis is important to describe the characteristics of an item. The excellence of the product can't be evaluated. The similarity estimation is inadequate to the product description.[9]

NLP is performed on the feedback that has been given by the users. It has different steps included like sentence segmentation, tokenization, normalizing, data cleaning and sentimental analysis. First is sentence segmentation where the bit of content is broken in different sentences. Subsequent stage is Tokenization in which the sentence is separated into singular words called as to-kens and can tokenize them at whatever point.

Normalizing which incorporates Lemmatizing and Stemming. The objective of both stemming and lemmatization is to de-crease inflectional structures and in some cases derivationally related types of a word to a typical base form.With that being stated, stemming or lemmatizing causes us lessen the quantity of in general terms to certain ”root” terms. Stemming is an unrefined method for lessening terms to their root, by simply characterizing rules of cleaving off certain characters toward the finish of the word, and ideally, gets great outcomes more often than not. Lemmatization is similarly a progres-sively orderly methodology of doing likewise which stemming does, however includes some vocabulary and morphological examination. Next comes Data cleaning which incorporates spelling and grammar revision and at last sentimental analysis is performed where characterization of feelings, for example, positive and negative words are done inside content informa-tion utilizing text analysis methods which is later fed to the database. NLP is performed on the feedback that has been given by the users. It has different steps included like sentence segmentation, tokenization, normalizing, data cleaning and sentimental analysis.

The use of natural language processing is done to give feedback on our recommender system. A text box of a limit of 10 words allows the user to enter a comment on which perform a sentiment analysis by training it against a database and then segregate them into positive, negative and neutral comments. First, is sentence segmentation where the bit of content is broken in different sentences. Subsequent stage is Tokenization in which the sentence is separated into singular words called as tokens.. Tokenization carries out this responsibility by finding word limits. Ending point of a word and start of the following word is called word limits. These tokens are valuable for finding such ex-amples just as is considered as a base step for stemming and lemmatization. Next comes Normalizing which incorporates Lemmatizing and Stemming. The objective of both stemming and lemmatization is to decrease inflectional structures and in some cases derivationally related types of a word to a typical base form.With that being stated, stemming or lemmatizing causes us lessen the quantity of in general terms to certain ”root” terms. Stemming is an unrefined method for lessening terms to their root, by simply characterizing rules of cleaving off certain characters toward the finish of the word, and ideally, gets great outcomes more often than not. Lemmatization is similarly a progressively orderly methodology of doing like-wise which stemming does, however includes some vocabulary and morphological examination. Next comes Data cleaning which incorporates spelling and grammar revision. It deals with removing the ”noise” in the data. Grammar checking is significantly learning based, tremendous measure of legitimate content information is found out and models are made with the end goal of grammar correction. A spellchecker focuses to spelling blunders and conceivably recommends options. Last sentimental analysis is performed where characterization of feelings, for example, positive and negative words are done inside content information utilizing text analysis methods. Sen-timent analysis is a text analysis method that identifies polarity inside content, regardless of whether an entire record, passage, sentence, or condition. Sentiment analysis is performed in light of the fact that there is an estimation that 80 percent of the world’s information is unstructured, as it were it’s chaotic. Huge volumes of content information is made each day yet it’s difficult to examine, comprehend, and sort through, also tedious and costly. Sentiment analysis in any case, enables the site to comprehend this unstructured content via naturally labeling it. Then the final words are fed back to the database. [10][11]

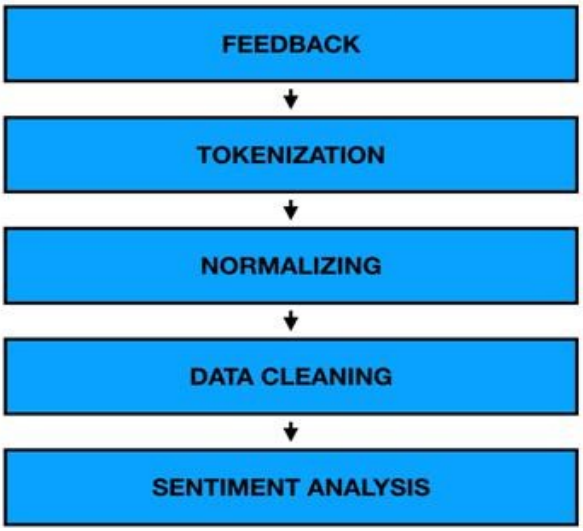

Fig. 1. Flowdiagram of NLP

Recommendation System algorithms that work on an E-Commerce website must cater to customers based on their personal interests of a product just like how career recom-mendation systems need to cater to the interests and strengths of the user. In these websites, Once the purchase is done, or the product is added to the card or wish list- similar products that similar customers purchase is enlisted below as ‘People also buy’. The models and methods used are cluster model - This model treats divides the user base into many parts and each user is assigned to the segment containing more similar customers, collaborative filtering is based on the products that are present in the shopping cart, it displays recommendations for the user and the searched based uses the clients purchase history and based on the items rated by him, the algorithm performs a search query to find other products which are popular by the same writer, artist or director or it uses keywords and subjects similar to the items to give results. Similarly, in the proposed system planned to implement content based filtering techniques on the inputs given by the user and then display the top most feasible career paths for him. In this paper, the method used to for rating of

recommender system is a star-based system that helps analyze the efficiency of the system built and to check how effectively the system recommends. [12][13]

The Video Recommendation based on ML With the ad-vancement of people watching, streaming and downloading small scale videos. The video recommendation system tracks the user's list of viewed videos and lists their preferences. The recommendation system that is being implemented also ensures user preferences while recommending an apt career. These suggestions are usually calculated from data sets using content-based proposals that in turn improve the suggestions made by the framework. Finally, clustering is performed classifying users into segments. Division of the users into segments enhances the recommendation based on similarity algorithm. The drawbacks of this system were an effective way to locate accurate video watchers, making accurate suggestions of what videos that the system recommends and the framework developed must deal with a large number of customers making real time requests to watch videos. This paper implements content based filtering and recommends careers based on user preferences. The career recommendation system also recommends based on user preferences and strengths but the method adopted is content based filtering.[14]

The desired features of a Research Paper Recommender System are accuracy - information needs differ among users due to different backgrounds and knowledge, inclinations and objectives, and settings consequently the framework needs to fulfill the client's individual information need. The content must also be relevant to the user. This is exactly same as a career recommendation system where the profiles of the users are extremely diverse and the content needs to match each users profile. User satisfaction - The recommendations made to users based on the research paper must be optimal. Recommender systems must also vary on demographics of the user i.e the age group and qualifications of the user to provide optimum user satisfaction. This feature is also included in this paper where user details are taken as a form-based input and ensures maximum user satisfaction. This method validates user satisfaction by considering user feedback using a star-based rating system that ensures feedback and accuracy of the recommendation system and comment box for user feedback. The comment box is further processed using NLP and sentiment analysis is performed. All these steps ensure user satisfaction as well as an accurate recommendation of a career domain most preferred for a user.[15][16]

Content based filtering algorithm predicts users' choices and preferences and hence a list or description of most accurate items or predictions are recommended to user. A collaborative filtering-based recommender system for an e-commerce website has following: User registration: On the user-side the client can enter all the details. After this, they can login using specific user credentials and password. All the newly added and modified items are included into the product list. Then they can buy the necessary products. Data retrieval system: The server will keep track of and save all the user's choices and preferences in their database and confirm them if necessary. It has to set up the connection to communicate with the clients and modify each client's activities in its database. It will validate every client before they access the application. So, this will help us prevent will prevent illegal users from accessing the system. This career recommendation system also uses user credentials and passwords to validate the user. A similar natural language processing algorithm is used to identify good and bad words which helps us identify if our recommender system has a positive or negative feedback – a feature that is included in the comment box.[17]

Nowadays, automated recommendation systems have com-paratively become better at suggesting items, their perfor-mance declines when they have limited data regarding the clients' predilection and choices. This usually happens when the clients have joined recently and hence there is limited information about their like and dislikes. So in this research, they have attempted to use natural language processing (NLP) to make suggestions when information about the users' choice is not within reach.[18]

The algorithm which is primarily used is word-space simi-larity. This uses a standard word-space similarity metric. Each plot summary is converted into a vector of binary character-istics, one per word, where the value of each characteristic shows whether a particular word is used in the plot summary. The semblance between any two films is just the cosine between those two characteristic vectors.[19]

III. Proposed Methodology

The existing systems have Collaborative Filtering ap-proaches which suffer from three problems: cold start, trust and privacy. These three drawbacks are very problematic at times and crucial opportunities can be missed because of this. The aim is to solve this problem and build a recommendation system in Python since it is easy and efficient to implement algorithms on different operating systems. Python also has many libraries and functions that allow us to perform actions with variations in code. The aim is to use various python libraries to apply Machine Learning techniques. The dataset used for training the model is the most important part of the project. The dataset has been created by pre processing and combining multiple databases. Our dataset contains 17 columns and 20,000 entries. The columns are Percentage in Operating system, Percentage in Algorithms and design, Average percentage in programming, Percentage in software engineering, Percentage in computer networks, Percentage in electronics, Percentage in computer architecture, Percentage in mathematics, Rate communication skills on 100 , Hours you want to work every day , Logical quotient rating, Number of hackathons won, Rate your coding skills, Rate your public speaking ability, Do you want to work for a long time with the system? ,Self-learning capability,Willing to take up Extra courses. CSS and HTML to style is used in our website and have also ensured that the website is user-friendly and intuitive. There is a Home and About page and then have a Feedback page that allows users to give feedback. The user credentials are stored in a database. If it is a new user -

they are made to register. The register page collects details about the user and stores them in the database. After this the user is redirected to the log-in page where he has to re-enter his/her credentials. The recommendation takes a form-based input from the user and recommends an accurate and apt career. It recommends the three best career options that the person can take up. For the backend, have used Python. To connect frontend to backend, have used Python Flask and predict the career of the user based on the data filled by them and performed cosine similarity on that. Cosine similarity is the similarity between two vectors. It can be applied to things accessible on a dataset to register closeness to each other by means of keywords. Closeness between two vectors (A and B) is determined by taking the dot product and dividing it by the magnitude value. To get our data into the format required for cosine similarity used pandas and numPy files and have used pandas is used which is a fast and effective data frame used to manipulate the data. It provides various tools reading and writing data between in-memory data structures and different formats. It also is extremely effective in merging and joining of various data sets. It also has an intelligent label-based function. With regards to repeated reading of the identical data from a local disk storage, Numpy offers .npy file format. This file format makes reading data extremely fast from simple CSV documents.Finally In our project, need to deal with an extremely large datasets of 20000 entries which contain many labels in the suggested job role column. These labels are in the form of words or numbers. This is basically done to convert words into numbers that can be understood by the machine learning algorithm. Algorithms can then determine in a better way on how the labels should be operated. It is an important pre-processing step for the structured dataset in supervised learning and have applied Label Encoding on the job data set, on the target column which is suggested job roles. It contains about 50 domains.

Instead of forcing a user login initially, requesting the user to answer a few questions based on preferences initially to generate a non-detailed but an example output. The process starts by cleaning and building the datasets, then get the numerical features from data, after that cosine similarity function is applied to get the similarity between previous user preference and the available jobs and finally get the top recommend jobs according to the score of the similarity. The aim is to implement a feedback and a comment section that has a textbox of 10 words. NLP is performed on the given feedback and determine whether it's a positive, negative or a neutral comment to provide better results to the students using the recommender system. Block diagram of the proposed system:

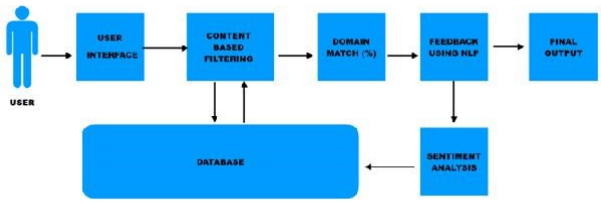

Fig. 2. High level block diagram of career recommendation system

IV. Conclusion

The principle undertaking of Career Recommendation frameworks is to consider what the user proposes his in-clinations and qualities are and map this with the accurate professional choice. It thinks about the client's field of interest and furthermore the user's profile interest and abilities, which can produce applicable career suggestions for the user. The framework provides extra abilities to undergraduates required for related employment opportunities. Users can extend their portfolio and resume to have the option to propel their vo-cations further. Users can give their feedback and criticism dependent on their experience on various parameters. This system is developed for engineering students, it is possible to include other streams such as business, arts. The student's profile can be handled in a more secure way by providing data encryption. This system can be implemented using collabora-tive approach.

References

[1] E K Subramanian, Ramachandran." Student Career Guidance System: Recommendation of a course". In "International Journal of Recent Technology and Engineering", Volume 7, Issue 6S4,2019.
[2] Ivens Portugal, Paulo Alencar, Donald Cowan." The Use Of Machine Learning Algorithms In Recommender Systems: A Systematic Review".
[3] Manish Kumar Singh, Dr. Dinesh Prasad Sahu." Research Aspects Of The System". In International Journal For Research In Applied Science Engineering Technology (IJRASET), Volume 5 Issue XI November 2017.
[4] Bhumika Bhatt, Prof. Premal J Patel, Prof. Hetal Gaudani. "A Review Paper On Machine Learning Based Recommendation System. In "International Journal Of Engineering Development And Research" Volume 2, Issue 4 ,2014.
[5] Kaveri Roy, Aditi Choudhary And J. Jayapradha."Product Recommendations Using Data Mining And Machine Learning Algorithms ". In ARPN Journal Of Engineering And Applied Sciences Vol. 12, No. 19, October 2017.
[6] Kaustubh Kulkarni , Keshav Wagh, Swapnil Badgujar, Jijnasa Patil , A Study Of Recommender Systems With Hybrid Collaborative Filtering , Volume: 03 Issue: 04 — Apr-2016.
[7] Priyanka." A Survey Paper On Various Algorithm's Based Recommender System". In IOSR Journal Of Computer Engineering, Volume 19, Issue 3, (May - June 2017).
[8] Mukta Kohar, Chhavi Rana."Survey Paper On Recommendation System". In (IJCSIT) International Journal Of Computer Science And Information Technologies, Vol. 3 (2), 2012.

[9] Joeran Beel, Stefanlanger, Marcel Genzmehr, Bela Gipp, Corinna Breitinger, Andreas Nurnberger,¨ "Research Paper Recommendation System: A Quantitative Literature Survey", International Journal On Digital Libraries (2015).

[10] Effectively Pre-processing the Text Data Part 1: Text Cleaning Ac-cessed January 30, 2019. https://towardsdatascience.com/effectively-pre-processing-the-text-data-part-1-text-cleaning-9ecae119cb3e

[11] Introduction to Natural Language Processing Accessed Septem-ber 10, 2019. https://www.geeksforgeeks.org/introduction-to-natural-language-processing/

[12] Greg Linden, Brent Smith, Jeremy York," Amazon.Com Recommendation", IEEE Computer Society, 2003.

[13] N. Divya, S. Sandhiya, D.R. Anita Sofia Liz , P. Gnanaoli ,"A Collaborative Filtering Based Recommender System Using Rating Prediction", International Journal Of Pure And Applied Mathematics Volume 119 No. 10 2018, 1-7.

[14] Kousthubha Balachandra, Mr.Chethan R, The Video Recommendation Based On Machine Learning, International Journal Of Innovative Research In Computer And Communication Engineering, Vol. 6, Issue 6, June 2018.

[15] Michael Fleischman And Eduard Hovy ,Recommendations Without User Preferences: A Natural Language Processing Approach.

[16] Zhe Yang, Bing Wu, Kan Zheng, Senior Member, IEEE, Xianbin Wang, Senior Member, IEEE, And Lei Lei, Member, "Recommender System For Mobile Internet Applications". In IEEE,2016.

[17] "Exploring Collaborative Filtering on Implicit Data for Job Recommendations". Accessed October 4, 2019. https://www.welcometothejungle.com/en/articles/collaborative-filtering-job-recommendations

[18] Jan, Teichmann."How to build a Recommendation Engine quick and simple" Medium. Accessed October 4, 2019. https://towardsdatascience.com/how-to-build-a-recommendation-engine-quick-and-simple-aec8c71a823e

[19] Armand, Olivares."Building NLP Content - Based Recom-mender Systems" Medium. Accessed September 10, 2019. https://medium.com/@armandj.olivares/building-nlp-content-based-recommender-systems-b104a709c042