Career path and Skill Recommendation System
End Semester Evaluation Report
Algorithm and Optimization for Big Data (AOBD)

Instructor : Prof. Ratnik Gandhi
Authors : Vidit Shah
Roll No. :1401078

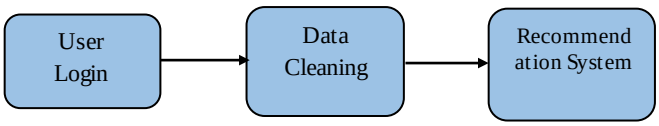
***Abstract*— The LinkedIn, the world's largest professional network uses horizontal collaborative filtering known as browsemaps. This is used for job suggestions. In this report, we propose our algorithm for skill recommendation. The approach use in this paper is first the data cleaning for fetching the skills from the given dataset and then suggest a career progression skills using collaborative filtering.**

Keywords—Career progreesion path ;Data cleaning; collaborative filters.

I. INTRODUCTION

The task given is we have to design the modules for LinkedIn user to suggest the set of skills which are to be required for their career development. So the modules we have to design are: (1) A module that reads user's profile and suggest a career path – in terms of skillset - to be acquired. (2) A module in which user enters a career goal and based on this career goal and other related information the platform suggests a career path. This are the two modules which are supposed to be design but before implementing these modules the bigger problem is How can I take my relevant data (which is useful to design the modules) from big data which is given.

II. OVERALL APPROACH



The main task is to make Recommendation system for the user that this are the skills that you should acquire for the career development. So as we can see in above diagram this is the overall approach which I have use. The detailed description of this approach is given below.

III. ORGANIZATION OF REPORT

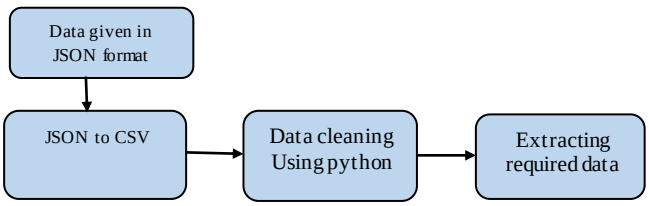
So the report organization is as in section 1 and 2 there are introduction and abstract in 4 tells about Data cleaning, 5 is about collaborative filtering,6 is Implementation methods of modules, In Section 7 there is Code output and Time complexity, 8 is for "Proof of Correctness" and 9 is for conclusion 10 is for future work and 11th section is references.

IV. DATA CELANING

Data cleaning is the process of removing corrupt or inaccurate entries from the given dataset. The main purpose for data cleaning are Eliminate the Redundancy, Increase data Reliability.

So the data about the candidate is given in text files. Which is in "json" format.

The method which I use for "Data Cleaning":



So, the process of data cleaning is shown in the above flow diagram.The text files which are given is as a data all are in JSON format. So for implementation of any method I have convert it in the ".csv" format. This part I have done using online tool. Now I have all ".csv" files then in data there are special characters like comma,bullet,etc. so first I will remove that special characters using the python script.Then taking the candiadtate ID and skills and store it in the ".csv" format.so here is output of this process of data cleaning:

CandidateID	Skills
0	Selenium, Protractor, Cucumber JVM, Rest Assured, SoapUI NG Pro
4	Test Automation
7	Vba, Visual Basic, Testing, Project Management, Project Planning, Quality Assurance, Microsoft Office, Agile, Software Testing, Software Development
11	Automation Testing, API Testing, Functional Testing, Database Testing, Cross Browser Testing
14	Manual, System, Automation Testing, Embedded System
15	TOSCA, Functional Testing, Regression Testing, Manual Testing, Automation Testing
17	Siemens S7-300, Win CC, Epam, Codesys, Microsoft Office, Microsoft project, ECS
24	C#, PowerShell, Python, Perl, Bash script, Tcl
27	Quality Engineering Automation, Web Development, Graphic Design, Animation, Music
28	NFC validation, Bluetooth Testing, Android, Python
36	Electrically trained Automation Engineer, Welding and fabricating, DHF Gate Safety, Safety Passport trained, Faac Gate Safety Courses, HSE Site Safety Trained

A. Issue and possible solution for data cleaning:

Here I have to do some stuff manually, i.e. In given dataset there are some words which are written in other languages. So in order to remove that words which are written in other

language I should have solid knowledge of "Natural Language Tool Kit (NLTK)". The other thing also in "Job Description" there are whole sentences written for data cleaning I have to find and extract the keywords for that also NLTK is required.

V. COLLABORATIVE FILTERING

The purpose of this report is that we want to recommend the skills that should be acquired to the users of the LinkedIn for their career development. So the recommendation is a suggestion or the proposal as the best course of action. The method which I use in this paper is the collaborative filtering.

This is called collaborative because it makes recommendations based on other people in effect, people collaborate to come up with recommendations. LinkedIn makes extensive use of item-based collaborative filtering, which showcases relationships between pairs of items based on the wisdom of the crowd. It works like this. Suppose the task is to recommend a book to you. I search among other users of the site to find one that is similar to you in the books she enjoys. Once I find that similar person I can see what she likes and recommend those books to you.

VI. IMPLEMENTATION METHOS OF MODULES

The definition of Module (I) is as reads user's profile and suggest a career path in terms of skillset to be acquired. We can say that we have to design same for skills recommendation. The abstract idea be like suppose LinkedIn has 5 users in the system. So they will enter their skills and we will see that suppose the four users have skills like C, C++, Java, other programming skills. And the 5th user have C, C++, and other skills so our program should be able to tell that user that for your career progression this skill: JAVA will be needed.

For the implementation part there will be major roll of the Machine learning concepts and Collaborative filtering. The Machine learning concept which is K-nearest neighborhood (KNN) algorithm. So the initial step is first I will be cleaning the data then from that I will extract the required data then I will check the skills of the current user after that I will check similarity among the all users and after that I will suggest the skill. So the flow of algorithm can be like:

- Taking Input data.
- Data cleaning.
- Fetching out the required data.
- K-means neighborhood.
- Finding the similar users.
- Suggest the Skills to user.

So as the flow of the algorithm is explained above for suggesting skills. After Completion of data cleaning I have use the K- nearest neighbor (KNN) because the similarity between the users can be calculated using KNN algorithm (k = 1).

In the implementation of next Module(II) the task given is when user enters a career goal and based on this career goal and other related information the platform suggests a career path. So in this module user will give his/her career goal based on that we have to give the path of the career. The algorithm is

- Input: desire/career goal ()
- From users u(base) profile recognize w1 = industrial work experience.
- Find users in the domain of input goal ()
- For users u(i) in ().csv file, find the jobs they have done in the work-job-titles column
- Process the job duration column to find the job first done by the users u(i)
- Extract keywords from job-titles to find what the user did in that job (eg. 'internship')
- From that suggest career path.

So suppose a user enters career goal as 'software engineer' and their work-profile shows that they have 0 experience, algorithm will process work_experience and learn that most users start with internship so the first step of career path for this user will be internship, followed by entry-level programmer, etc.

A. Limitations of the implementation:

In this Implementation I am not able to do it for complete data set so for that reason my implementation is based on limited users set.

VII. CODE OUTPUT AND TIME COMPLEXITY

So here is the output of the implementation module, here as explained above the implementation is done on limited data set and giving appropriate results.

```
'''
===== RESTART: C:\Users\vidit_shah\Desktop\AOBD_finalppr\recommender.py =====
>>> r = recommender(users)
>>> r.recommend('0')
This are the skills that user should acquired
Python MongoDB Hadoop
>>>
```

In r.recommend('0') the 0 is the "Candidate ID". The total **"Time Complexity"** of code is **O(n*s) ,** where n is the number of users and s is the skills.

VIII. PROOF OF CORRECTNESS

In "Proof of correctness" we have to prove mathematically and by the output that the solution which we are providing is correct but in ML and Data mining which are open problems the proof of correctness is that "Why you have use this method ?" like suppose there are variables like a,b and if answer is 3 then we have to prove that addition is the best operator so in

same manner I have used Collaborative filtering so I have to prove that “Why we are use collaborative filtering method ?”, because this approaches building a model from a user's past behavior as well as similar decisions made by other users. And by the research CF has been proved to be effective for solving the information overload problem.

IX. Conclusion

The problem was given is the problem of “Machine Learning” and “Data Mining”, by the above implementation I am able to give user the skills that to be required using on limited dataset which mentation above, So the overall time complexity is O(n*s) where n is number of user and s is skills.

X. Future Work

The work need to be done is that try to apply this algorithm on the whole dataset. The other remaining work is about Nature Language Tool Kit(NLTK) for data cleaning the reasons are mention in section(IV).

XI. References

[1] Browsemap: Collaborative Filtering At LinkedIn, Lili Wu, October ,2014.

[2] http://ieeexplore.ieee.org/document/7270736/

[3] http://www.guidetodatamining.com/

[4] http://ieeexplore.ieee.org/document/7176109/