End Semester Examination - Report

April 28,2017
BTech (ICT) Sem VI
Winter 2017
Algorithms and Optimizations for Big Data
Ahmedabad University
Report by: Mihir Gajjar - 1401076

***Abstract*—The Web, is leaving the era of search and entering one of discovery - when something wonderful that you didn't know how to ask for, finds you. Social network like LinkedIn contains a lot of professional data about millions of users across the globe. The aim of this report is to build modules for suggesting career progression path to its registered users. Using the user's profile, relevant suggestions on how the user should consider next set of skills to be acquired and a career path based on the user's career goal is recommended.**

***Index Terms*—Inferring attributes, Recommender system, Collaborative filtering, Content-based filtering.**

I. Introduction

An online social network like LinkedIn consists of a collection of profiles that represent a member's professional identity. A profile contains a member's current and past work history, education, and projects, among other pieces of his identity. For many members an important part of that identity is the set of skills and talents for which they are known. On LinkedIn, the "Skills and Expertise" section is a part of the profile that allows the member to display this information. The user may add skills to his/her profile either by manually entering them, in which case a type-ahead prompt is provided, or by accepting recommendations of inferred skills. This paper aims at the following two modules:

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.

A. Recommender Systems

Recommender systems work by analyzing characteristics of clients such as their historical buys, age, gender, etc (also known as features) and trying to infer future interests by identifying what other clients with similar characteristics have purchased. In this sense, the more data, the better the predictive power. Hence, Big Data are actually a good thing for recommender engines and machine learning, in general. Recommendation Algorithms are based on two basic filtering techniques: Collaborative Filtering, Content-Based Filtering.
The Collaborative filtering approach, exploits either user-to-user or item-to-item similarities to produce the recommendations. The user-to-user approaches searches for what other "like-minded" users found interesting or purchased; the item-to-item approaches identified which products were frequently bought together.
Content based filtering makes recommendations based on user choices made in the past. (e.g., if a user rated an action movie positively over a movie recommendation site, the system would probably recommend more of recent action movies that he has not yet seen or rated).

II. Data Manipulation

The data manipulation can be performed as follows:

1) Reading the data: Convert the data from the txt files(in JSON format) to files with CSV format.
2) Cleaning the data: Clean the files with CSV format by removing the data impurities.
3) Structuring the data: Organize the data in the files with CSV format in a structured form.

A. Reading the data

The given data of the candidate profiles is in a JavaScript Object Notation (JSON) format as can be verified by this online tool[7]. To bring the data in a suitable format, first the text files are converted to CSV files using this online tool[8].

B. Cleaning the data

Now we have the data in the CSV format, however, this data is not clean. These are the following impurities present in the data:

- There are special characters present in the data. For example: ✓, ã , ç , etc.
- There are spelling mistakes in the data. For example: In "Senior Web Developer.txt" file in the Skills section of candidate 18, M.S. Office is written as M.S. Ofiice.
- The data in some of the text files is in a different language. For example: data in "Senior IT Architect.txt" file is in Portuguese language.
- In the data many of the potential skills are duplicate. For example: "Java programming" and "Java development" are both equivalent to the skill "Java" .

Hence, we need to take care of these impurities. By using the tools provided by Microsoft Excel the special characters in the CSV files are removed and the spelling mistakes are corrected. Currently, in the implemented solution of modules the data which is in English language and does not have any data duplication are used. However the language problem can be solved by using a translating tool[9]. The data duplication problem can be solved by using clustering algorithms and by giving a name to each cluster. Hence, the cluster for a skill can be decided using the clustering algorithm and the name of that assigned cluster can now be used as an alternative name for that skill.

C. Structuring the data

Now the cleaned data is available in the CSV files, but the data is still not in a structured form. The data entry in different sections(like Skills, Company, etc.) has been done by using different delimiters like ,(comma), |, &&, etc. Hence the entries in these sections(like Skills, Company, etc.) needs to be separated. Hence new CSV files are created using a python script which separates all the skills by using only ,(comma) as a delimiter. In the new CSV files we have two columns, 'Candidate ID' and 'Skills' (in which the entries are separated by using ,(comma) as a delimiter.) . Hence, now the data is available in a structured form.

III. Module - I

A. User - based Collaborative filtering

Collaborative Filtering is used to recommend a career path - in terms of skillset - to be acquired by reading the user's profile. It is based on the assumption that an active user preferences would be in accordance with other similar user preferences. In particular, **User - based Collaborative filtering** is implemented where in neighborhood of similar-taste people is selected and their opinions are used for making predictions.

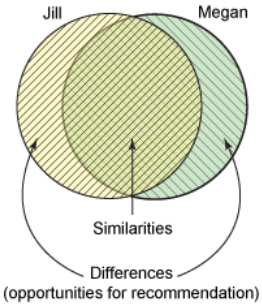

Figure: 1 Similarities and Differences used in User - based Collaborative Filtering.

B. Feature Extraction

Now the data is available in a structured form. The features(skills) from the comma separated entries is extracted and a feature vector for each user is created which indicates the skills he/she possess.

C. Recommendation System

Now, the feature vector for each user is available which indicates the skills that particular user possess, a user to user correlation matrix will be created. Based on this correlation matrix, the similarity between users is calculated using k-nearest neighbour. Then, based on the user neighbour's skill-set, one can predict the skills for a user(who does not possess this skill) . Then, based on user's similarity with other users promising skills for recommendation can be selected. For example, a user possesses skills similar to five other users, hence, we can recommend that particular user those skills which other five users possess, but that particular user does not possess. Based on the feature vectors the similarity/relationship of one particular user with other users is calculated using k-nearest neighbour. In this way collaboration is achieved.

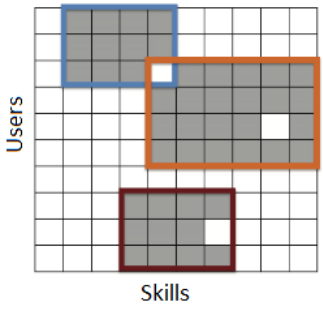

Figure: 1 User - skills Matrix Representation.

In the above figure(matrix), the rows represent the users and columns the skills. A gray box indicates that the client possesses that skill. The three large gray areas are the clusters which indicates the users having similar skills. The white boxes represent good candidate recommendations of skills to the users.

D. Proposed Algorithm

The collaborative filtering based recommendation algorithm for recommending skill-set to be acquired to the user is as follows:

ALGORITHM: (Collaborative Filtering Based Recommendation Algorithm)

---

1) Manipulate the data.
2) Extract the features(skills) from the data and prepare feature vectors for each user.
3) Create a user to user correlation matrix.
4) Based on the correlation matrix, calculate the similarity scores between users using k-nearest neighbours.
5) Predict the skill to be suggested to an active user (who has not acquired this skill earlier) by using the user neighbour's skill-set.
6) Select promising skills for recommendation based on user's similarity with other users.

IV. Module - II

A. Content-based Filtering

Content - based filtering is used to recommend a career path based on the user goal and other related information about the user. Content - based filtering makes recommendations based on the user choices. A user model is created based on the career goal he chooses and from the other attributes of the user. This model is applied to predict which kind of Job, Education, Company etc. would be suitable for a user based on his preferred career goal.

B. Feature Extraction

From the files available in the CSV format, new CSV files are to be created which contains the job title 'Job Title','Skills','Education','Company', etc. by using only

,(comma) as a delimiter. Now, from these comma separated entries, the attributes of the user are to be extracted. Hence, the features used will be 'Job Title','Skills','Education','Company', etc. A feature vector for every career goal is created which indicates the features related to that particular career goal. In the feature vector, the entries in every attribute, will be sorted based on the number of occurrences of that entry in the available data. Hence, on the top we will have the entries in the attributes which are most relevant for that career goal.

C. Recommendation System

Now, from the available data we will have the sorted feature vectors of the career goals. Hence, when a user enters his career goal, from the feature vector of that particular career goal, we can recommend: jobs, education required for that career goal, companies which provide jobs related to the career goal, skills required, etc. As the feature vector is sorted, the most relevant attributes for that career goal will be recommended.

D. Proposed Algorithm

The content-based filtering based recommendation algorithm for recommending career path based on the career goal is as follows:

ALGORITHM: (Content-based Filtering Based Recommendation Algorithm)

---

1) Manipulate the data.
2) Extract the features related to a particular career goal from the data and prepare feature vectors for each career goal.
3) Sort the entries in the feature vector based on the number of occurrences of that entry in the available data.
4) Relate the career goal entered by the user with the career goals available in the data and obtain the feature vector for that career goal.
5) Recommend the career path in terms of skills, jobs, companies, etc to the user (which the user currently does not possess) based on the feature vector of the career goal.

V. Implementation Results

After the manipulation of data is done on all the txt files containing data in the JSON format, the data is provided to the recommendation algorithm to provide recommendations - in terms of skill-set - to be acquired.

```
======= RESTART: C:\Users\Mihir Gajjar\Desktop\BX-Dump\recommender
>>> r = recommender(users)
>>> r.recommend('0')
['Unix', 'Perforce', 'IBM HTTP Server', 'SUN OpenSSO']
>>>
```

Figure: 1 Recommendations - in terms of skill-set - to be acquired.

In the above figure, the recommender is initialized with the data available in the structured CSV file "Computer Systems Manager.csv" . And then the recommendation system is asked to give a recommendation for the candidateID '0', and the skills which are recommended by the system are as displayed. The value of k for k-nearest neighbour is 1 and the maximum number of skills to recommend - n is 5.

VI. Proofs of Correctness

A. User - based Collaborative filtering

The basic assumption and idea behind Collaborative filtering is as follows:

- Users are ready to enter the skills they possess.
- Users who had similar tastes in the past, will have similar tastes in the future.

Under these assumptions, Collaborative Filtering is a very successful and promising technique for building recommender systems[10]. Collaborative filtering has been very successful in both research and practice, and in both information filtering applications and E-commerce applications[3][4][5][6]. Hence, the problem of recommending skills can be addressed by using the collaborative filtering based recommender systems.

B. Content-based Filtering

Content-based recommendation systems try to recommend items similar to the user's preferences[11]. Here, the career goal entered by the user is selected as the user preference, and career paths in terms of jobs, skills, company, etc. are recommended based on the career goal. Hence, the problem of recommending career paths to a user based on his career goal can be addressed by using the content-based filtering based recommender system.

VII. Efficiency of the Proposed Algorithm

The computational complexity of the proposed algorithm is measured to be O(n×s), where n is the number of users and s is the maximum number of skills a user can possess.

References

[1] Bastian, Mathieu and e. al., "LinkedIn skills: large-scale topic extraction and inference," in Proceedings of the 8th ACM Conference on Recommender systems, 2014.

[2] Priyanka Rastogi and Dr. Vijendra Singh, "Systematic Evaluation of Social Recommendation Systems: Challenges and Future" International Journal of Advanced Computer Science and Applications(ijacsa), 7(4), 2016. http://dx.doi.org/10.14569/IJACSA.2016.070420

[3] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. (1997). GroupLens: Applying Collaborative Filtering to Usenet News. Communications of the ACM, 40(3), pp. 77-87.

[4] Shardanand, U., and Maes, P. (1995). Social Information Filtering: Algorithms for Automating 'Word of Mouth'. In Proceedings of CHI '95. Denver, CO.

[5] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of CSCW '94, Chapel Hill, NC.

[6] Hill, W., Stead, L., Rosenstein, M., and Furnas, G. (1995). Recommending and Evaluating Choices in a Virtual Community of Use. In Proceedings of CHI '95

[7] "The JSON Validator". Jsonlint.com. N.p., 2011. Web. 24 Apr. 2017.

[8] "JSON To CSV Converter Online". Json-csv.com. N.p., 2012. Web. 24 Apr. 2017.

[9] "Google Translate". Translate.google.com. N.p., 2006. Web. 26 Apr. 2017.

[10] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," Advances in Artificial Intelligence, vol. 2009, Article ID 421425, 19 pages, 2009. doi:10.1155/2009/421425

[11] Poonam B Thorat, R M Goudar and Sunita Barve. Article: Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System. International Journal of Computer Applications 110(4):31-36, January 2015.