Algorithms and Optimization for Big Data - Final Exam Report

Akash Soni, Roll Number : 1401047
BTech (ICT), Semester : VI
$28^{th} April, 2017$
School of Engineering and Applied Science, Ahmedabad University

***Abstract*—Finding a career path has been a difficult, tiresome and time consuming process, because people looking for a new position had to collect information from many different but relevant sources. Career path recommender have been proposed in order to automate and simplify this task, also increasing its effectiveness. However, current approaches rely on scarce manually collected data that often do not completely reveal people skills. The aim is to find out relationships between career paths and people skills making use of data from LinkedIn users' profiles. Semantic associations arise by applying Latent Semantic Analysis (LSA)[2]. This paper aims in preparing two modules, first that reads user's profile and suggest a career path in terms of skillset to be acquired and second in which user enters a career goal and based on this career goal and other related information the platform suggest a career path.**

***Index Terms*—Collaborative filtering, Latent Sementic Analysis, Recommender systems**

I. Introduction

With the rapid developing of the Internet technology, more and more job seekers release their own personal information whereas enterprises post the jobs on the Internet. Because of the advancement of Web 2.0 technology, there is a dramatic increase in job seekers' personal information and enterprises' recruiting information. As a result, the information becomes overloaded, which lead to the low utilization rate. Some platform like LinkedIn uses the career path recommender system, which is the system that suggests the user their respective career path with the help of features like "skills" mentioned in their profile.

Recommendations are a matching problem. Given a set of users and a set of items, we want to match users to their preferred items. There are two high-level approaches to this type of matching: content based and behavior based. They each have pros and cons, and there are also ways to combine them to take advantage of both techniques. Recommender system is a system that has the ability to predict whether a particular user would prefer an item or not based on the user's profile. As a recommender system, the career path recommender system is capable of retrieving a list of career positions that suggest the user's desire, or a list of talent candidates that meet the requirement of acquiring a particular position or the career goal. For example, content-based recommender and collaborative filtering recommender which have shown success in different recommender systems.

Based on the papers being studied during of preliminary research, some issues that have been paid much attention in the career path recommender system.[2]

- How to extract the information of career position and people and contribute the user profile for matching the career paths and people well?
- Which recommendation technology is used in the career recommender system based on user profile?
- How to build a career recommender system based on the real data with a certain application background?

II. Block Diagram

The block diagram of the approach that is being taken and implemented is shown below:

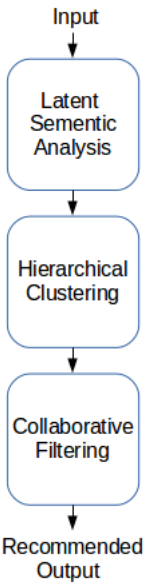

Figure: 1 Flow of the problem

III. Working

Let $X = x_1, x_2, ...$ be a set of user profiles, each being a description of a specific individual. To each profile X is associated a set $\theta(x_i)$ of the skills, which the corresponding individual has declared to possess. The same skill maybe associated to more than one profile. The set of all the distinct skills is denoted by $\theta = \theta_1, \theta_2, ...$ To each profile x_i is also associated a career goal i.e. job position $p(u_i)$. The same job position may be the current one for more than one profile. We denote by $Y = y_1, y_2, ...$ the set of all distinct job positions.

A. Hierarchical Clustering using Latent Semantic Analysis

The most important job after having the data is to clean the data and extract the necessary data from the entire dataset. Initially, the dataset is being merged into a single file. Then the focus is on obtaining a categorization of possible job positions i.e. career goals from the available data. In order to do so, a hierarchical clustering of positions in Y is performed. We extract a vector-based representation, where skills are used as features. Specifically, from the set X of known user profiles, we build a $|\theta| \times |Y|$ matrix C counting the co-occurrences between skills and positions across them. Values in C are computed as follows[3]:

$$c_{i,j} = |x \in X| : \theta_i \in \theta(x) \wedge y_j = y(x) \tag{1}$$

$c_{i,j}$ is the number of profiles having both θ_i among skills and y_j as position. Each position is then represented as a weighted mix of different skills, according to those possessed by persons employed in that position.

The skills within the set θ can be semantically similar to each other or even be synonyms. A well-known technique in this context is Latent Semantic Analysis (LSA), which employs Singular Value Decomposition (SVD) to obtain a lower-rank approximation of the matrix. We apply LSA to the skill-position matrix C to obtain a transformed matrix C'. We first decompose C into the following,

$$C = U * \Sigma * V^T \tag{2}$$

These matrices define a latent vector space, where skills and positions are represented by rows of U and V, while values in Σ indicate the importance of each dimension of this space. By initializing all the values of Σ except the r highest ones to 0 and multiplying back the three components, we obtain the transformed matrix C_0 , which is a denoised approximation of C with rank r. For the r parameter, we choose the minimum value for which the sum of retained eigenvalues is at least 50% of the total. The transformed matrix C_0 , as the original one C, contains for each position p_k a column vector p_k representing it. We evaluate the distance between two positions as the inverse of their cosine similarity.[3]

$$d(y_a, y_b) = 1 - cos(y_a, y_b) = 1 - \frac{y_a * y_b}{||y_a|| * ||y_b||} \tag{3}$$

The mutual distances between positions are finally given in input to a complete-linkage agglomerative clustering algorithm, which extracts a dendrogram of all positions. This dendrogram has the form of a binary tree with positions as leaves.

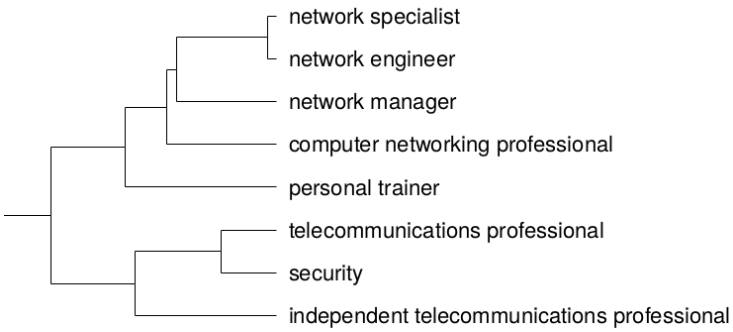

Figure: 2 Hierarchical Structure

B. Collaborative Filtering

The collaborative filtering algorithm has a very interesting property known as feature learning, that can learn for itself what features it needs to learn.[1]

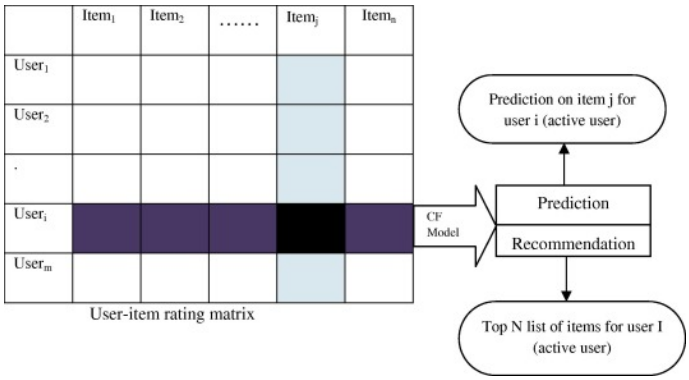

Figure: 3 Collaborative Filtering Process

Formalizing the collaborative filtering problem
We can more formally describe the approach as follows:

- Given $(\theta^1, \theta^2 ..., \theta^{nu})$ (i.e. given the parameter vectors for each users' preferences)
- Given $(x^1, x^2 ..., x^{nu})$ (i.e. given the parameter vectors for each users' skills)
- We must minimize an optimization function which tries to identify the best parameter vector associated with a skill.

$$min_{\theta^{(1)}} = \frac{1}{2} \sum_{j:r(i,j)=1} ((\theta^{(j)})^T (x^{(i)}) - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} (x_k^{(i)})^2 \tag{4}$$

- The summation is done over all the indices j considering the data for user x_i. Doing so, the cost function gets minimized and the squared error reduces.

Estimation functions
The estimation functions are as follows:

- If we're given the user's skills we can use that to work out the career path.

$$min_{\theta^{(1)},....,\theta^{(n_u)}} = \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} ((\theta^{(j)})^T (x^{(i)}) - y^{(i,j)})^2$$

$$+\frac{\lambda}{2}\sum_{j=1}^{n_u}\sum_{k=1}^{n}(\theta_k^{(j)})^2 \tag{5}$$

- If we're given the user's career goal we can use them to work out the user's skills.

$$min_{x^{(1)},....,x^{(n_m)}} = \frac{1}{2}\sum_{i=1}^{n_m}\sum_{j:r(i,j)=1}((\theta^{(j)})^T(x^{(i)})-$$

$$y^{(i,j)})^2 + \frac{\lambda}{2}\sum_{i=1}^{n_m}\sum_{k=1}^{n}(x_k^{(j)})^2 \tag{6}$$

Algorithm Structure

1) Initialize $\theta^1, ..., \theta^{nu}$ and $x^1, ..., x^{nm}$.As we do in neural networks, we will be initializing all parameters.

2) Minimize cost function (J($x^1, ..., x^{nm}, \theta^1, ..., \theta^{nu}$) using gradient descent. The updated rules look like the below one, where the top term is the partial derivative of the cost function with respect to ${x_k}^i$ while the bottom is the partial derivative of the cost function with respect to ${\theta_k}^i$

$$x_k^{(i)} := x_k^{(i)} - \alpha\left(\sum_{j:r(i,j)=1}((\theta^{(j)})^T(x^{(i)}) - y^{(i,j)})\theta_k^{(j)} + \lambda x_k^{(i)}\right) \tag{7}$$

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha\left(\sum_{i:r(i,j)=1}((\theta^{(j)})^T(x^{(i)}) - y^{(i,j)})x_k^{(i)} + \lambda\theta_k^{(j)}\right) \tag{8}$$

3) Having minimized the values, given a user with parameters θ and career path with learned skills , we predict a career goal of $(\theta^j)^T x^i$. This is the collaborative filtering algorithm, which give pretty good predictions.

IV. Entire Approach for implementation of Module 1

The input given are the the user's profile. After reading of the data is done, identifying the skills that the individual possess and recommending a career path to be acquired based on those skills is done. The hierarchical clustering of the career goals and the associated skills would have been already done using the approach mentioned above in the subsection $III - A$.

Now as the recommendation to be made is of the career path, initialization of the θ parameters is done as the parameters x are already given that basically represents the skills. Afterwards, the cost function is minimized iteratively by keeping x fixed and by making the use of update process. We need to learn all the skills for all the users - so we need an additional summation term, "The regularization parameter" λ which will be preventing over-fitting. Thus using the logistic regression approach, it will classify the user to a particular career path based on the possessed skills by formulating $\theta^T x$. The mapping of the career paths to the skills is already done. So, with this approach, the skills of the user which maps most to the career path and the same is our required output.

V. Entire Approach for implementation of Module 2

In this module, the user's career path is already known. The problem is to suggest the skills that should be acquired in order to fulfill the the career goal. The hierarchical clustering of the career goals and the associated skills would have been already done using the approach mentioned above in the subsection $III - A$.

Now as the recommendation to be made is of the skills to be acquired, initialization of the x parameters i.e. the skills is done, as the parameters θ are already given i.e. the career path. Now, using the above equation the cost function is to be minimized iteratively by keeping the θ fixed using the update process. We need to learn all the skills for all the users - so we need an additional summation term, "The regularization parameter" λ will prevent over-fitting. Thus using the logistic regression approach, it will recommend the user to acquire the skills in order to achieve the career goal by formulating $\theta^T x$. The mapping of the career paths to the skills is already done. So, the career path is analysed and the skills are recommended which the users lacks and it is our desired output.

VI. Outputs

First Module that reads user's profile and suggest a career path in terms of skillset

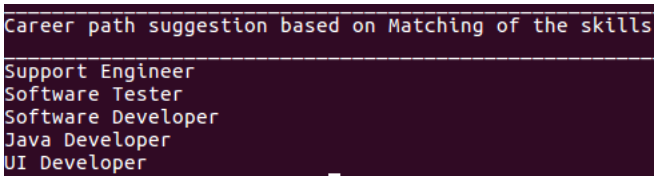

Figure: 4 Output for Module 1

Second Module in which user enters a career goal and based on this career goal, other related skills are suggested.

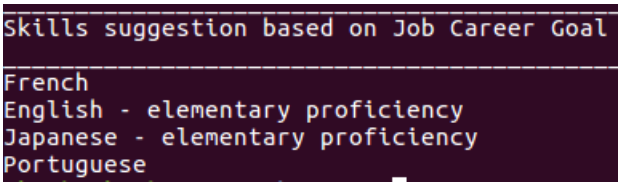

Figure: 5 Output for Module 2

References

[1] F.O. Isinkaye, Y.O. Folajimi and B.A. Ojokoh "Recommendation systems: Principles, methods and evaluation", doi.org/10.1016/j.eij.2015.06.005, "Egyptian Informatics Journal (2015).

[2] Zheng Siting, Hong Wenxing, Zhang Ning and Yang Fan "Job Recommender Systems: A Survey", The 7th International Conference on Computer Science and Education (ICCSE 2012) July 14-17, 2012. Melbourne, Australia

[3] Lili Wu, Sam Shah, Sean Choi, Mitul Tiwari and Christian Posse "The Browsemaps: Collaborative Filtering at LinkedIn", Proceedings of the 6th Workshop on Recommender Systems and the Social Web (RSWeb 2014), collocated with ACM RecSys 2014, 10/06/2014, Foster City, CA, USA.

[4] Giacomo Domeniconi, Gianluca Moro, Andrea Pagliarani, Karin Pasini and Roberto Pasolini "Job Recommendation From Semantic Similarity of LinkedIn Users' Skills", DISI, University degli Studi di Bologna, Via Venezia 52, Cesena, Italy