A Research Report

on

Career Path Prediction using Machine Learning:

An application in Indian setting

Prepared by-

Ayush Agrawal - 2017B3A70599P

Sarthak Sehgal - 2017B3A70452P

Mayank Singh - 2017B4A80769P

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

ACKNOWLEDGEMENT

Motivation and encouragement remain pillars of successful completion of any venture. We would like to express our gratitude towards Dr. Tanu Shukla for trust and constant guidance in preparing this research report. We would like to acknowledge research scholars Ms. Mounika and Ms. Divya for constantly reviewing, encouraging and helping us in finalizing and editing this report.

ABSTRACT

Career counselling has become an important aspect due to the emergence of various professions. A country like India, facing favourable demographic dividend could gain a lot if career counselling is effectively and efficiently implemented. Realizing this scenario and understanding the non-trivial task of career counselling, we have put the hat of a career counsellor on machine learning algorithms. In this paper, we apply machine learning algorithms such as k-means clustering, Markov chain to model the career path of an individual through the collection of profiles and features extraction from descriptive information available in LinkedIn. We solve the problem that given a person's current position and his/her goal, what is the best career path recommended for him/her. The model is applied to an Indian setting separately for privately funded technical institutions and government funded technical institutions. Further, differences and similarities are discussed between the outcomes of the models. As a conclusion, analysis of results and discussion of possible improvements in the model are presented.

BACKGROUND

Several career options make it confusing for college students or current employees to choose whether to pursue higher studies, change jobs or which sector to put legs upon. While making such decisions one gets overwhelmed thinking about the consequences of making the decision. They tend to go for counseling or getting advice from some acquaintances. Since people having experienced such scenarios tend to give proper advice to students/employees, they act as perfect counselors. So, accumulating these experiences from a large number of people and using machine learning algorithms so as to make a consistent career path prediction system could be a source of help to many.

Career path model is presented with a sequence of his/her education or working experiences in time order. Each segment of experience regarding work or education is denoted as a node. For education, defining factors are college and degree. For working experience, defining factors are position and employer (company).

Each person's profile and career path is a sequence of multiple nodes (steps). Each node will be represented by various multinomial features, specifically - profession, industry and company size. Given a person's current career path and his/her goal, we try to discover the optimal career path. An optimal career path is defined as a path that has the highest probability to reach the goal node. In case multiple paths exist having the same probability, the paths with shortest length (least number of jobs) are optimal.

This model is being applied to an Indian setting. To come up with such a model several challenges were faced in the form of how to account for such a heterogeneous population which is characterized by ethnic, cultural, economic, geographical and other extrinsic as well as intrinsic factors; how to translate qualitative information of one’s profile into features; how to solve lack of job specificity problem and choosing algorithms for estimating path recommendations.

LITERATURE REVIEW

Factors affecting career choice

The factors have been classified under four broad categories - Demographic (age, gender, geography); Socio-economic (economic status, financial constraints, economic stability); Personal (personality, interest, academic ability, aptitude, beliefs); and Socio-environmental (family, peer group, school).

1. **Demographic factors**

   Omari et al. (2017), studied factors affecting career choice among university students in the school of Business and Economics, Kisii University. The findings revealed that age influences career choice decision. According to them, this also concurs with Owie (2003) who argues that age is the most critical factor in career choice, the most significant rationale for a person opting for a specific career is the natural zest in the area which comes with age. Three characteristics - personal attainment, educational attainment, and career experience were taken as a measurement of age. Their study also revealed that gender was the most influencing factor affecting career choice. According to Dick and Rallis (1991), women continued to be disproportionately underrepresented in science and engineering arenas. In their findings, which was collected from nine high schools of Rhode Island showed that salary was a more important factor for men while genuine interest was a more important factor for women not choosing careers in engineering and science fields. Also, there remained a marked disparity in the proportion of men to women planning careers in engineering or science. Mtemeri's (2017) study contradicted these findings stating that gender didn't influence career choices among high school students. According to this research, both male and female students compete for the same careers. There is no difference between the subjects done by both that lead them to the choice of different careers. It also revealed that both male and female role models influence students in their choices of careers. An interesting finding in this study was the influence of the geographical location of the school in career choice.

2. Socio-economic factors

For choosing a career, there are intervening variables such as financial constraints (the ability of a student to pay his/her tuition fees). Moving forward, expected financial reward motivated the students in choosing a specific career based on the salary they were expecting at the time of employment (Omari et al., 2017). The same is supported by Ahmad et al. (2017). Their study was based on factors affecting career choices among business students. Both financial constraints and financial outcomes determined career choice to a great extent. Students whose parents own and operate small businesses may want or feel obligated to follow their parents' businesses as students may consider the ease that the job would be available to them right out of school, they could hold a high position within the business, and there is a possibility that they might own and operate the business one day (Zody et al, 2006; Fizar, 2013). financial aspects that students consider include high earning potential, benefits, and opportunities for advancement (Beggs et al., 2008; Fizar 2013).

3. **Personal Factors**

Olamide and Olawaiye (2013), studied factors affecting career choice in secondary schools of Ogun state and found that personality largely affects career choice among secondary school students. Descriptive variables taken for personality were - grades, doing career research on their own, being willing to work in a job traditionally held by the opposite sex and being the type of student who would choose his or her own secondary school subjects. Dick and Rallis (1991), stated that students make career decisions on the basis of their beliefs about themselves and their own abilities and their beliefs about the relative values of different careers. A career's perceived value is determined by intrinsic factors that include intellectual interests. Empirical studies have determined that there is a positive and significant relationship between the interests of students related to a particular subject and career choice (Ahmad et al. 2017). In a research study the factor "match with interest" rated over job characteristics, major attributes, and psychological and social benefits in importance when students choose a major (Beggs et al.,2008; Fizar 2013). The difference arises in choosing a more or less work-intensive career path depending upon the academic ability of the students i.e. the ability to handle majors with greater workloads or choosing path requiring more education (Fizar 2013).

4. **Socio-Environmental Factors**

It was found that peer group influences career choice decisions (Omari et al. 2017). In this study three characteristics of peer group were taken into account- peer group age, peer group gender, and peer confidence. Out of the three, peer group gender and age came out to be most influential. From the above study, it was also concluded that parental advice is also a major contributing factor for the same. Weldernfael and Dodge (2014), studied the factors influencing career choice in South African Township High School students. In their study, peer pressure accounted for fifty-three percent of social factors and lack of parental support accounted for twenty percent of social factors and was the major barrier leading to poor career path choice. In Mtemari (2017) study, parents influence on students' career choices comes in various forms- parental actions, values and beliefs, connectedness, expectations, education, and careers. School also has a major influence on career choice. Career guidance lessons students learn from teachers have a bearing on students' choices of career. From the study, it also emerged that peer advice has a significant influence on students in choosing career paths. This advice may come formally or informally during interactions with other students. The students may unknowingly define their thinking based on ideas and suggestions of their support group (Olamide and Olawaiye, 2013). Some students have more opportunities in the form of scholarships, guidance, etc which affects their decision on choosing a particular career path (Fizar 2013). Olamide and Olawaiye (2013), used nine descriptors for measuring opportunity and found that it had an influential role in affecting career choice.

Problems faced and related algorithms/models developed

Career path prediction is a non-trivial task due to a variety of problems ranging from lack of specificity of job titles to difficulty of accounting several intrinsic and extrinsic factors including time which affects career choice and ultimately path. This part of the literature review presents various problems faced in modeling career path and related algorithms and models developed to curb these problems.

1. Data Profiling

For training, a lot of data is required. Collecting the data of several peoples' profiles that is generalizable or analyzable is a challenge. Generally scraping of the data from social networking sites such as LinkedIn, Twitter and Facebook is done. Liu et al. (2016) collected data from About.com in making career path prediction model, which encourages its users to list their multiple social accounts explicitly in their personal profiles. Social accounts were from LinkedIn, Facebook, and Twitter. According to them, there is no available benchmark dataset suitable for career path modeling. Li et al. (2017) in their 'Next career move prediction model' took data from LinkedIn profiles. While Mimno and McCallum (2008) collected 9722 resumes for modeling. Subahi (2018) reviewed recruitment sites data and took data from three such sites. Lou et al. (2010) collected 67000 profiles from LinkedIn to model career path.

2. Lack of job specificity

Because of vocabulary variants, people state the same job titles in different words e.g. product managers and marketing coordinators are referring to the same job title. Because of these variations, it becomes tough to use such data without pre-processing. Liu et al. (2016); Mimno & McMallum (2017); Subahi (2018) manually mapped all the semantically similar titles to the same career stage. Job titles heterogeneity leads to the problem of generalization and embedding inspired by Natural Language Processing can be used to solve this problem (Li et al., 2017). Lou et al. (2010), used K-means clustering in order to club similar titles together based on the measure of average semantic similarity distance.

3. **Translation of qualitative information into quantitative**

Liu et al. (2016) based on prior knowledge divided the career path into four parts, where each part will represent a milestone. To account for work experience, they made time stamps and for quantized milestones in each career path from one to four. So each step contained time stamp and milestone reached. In this, the problem might arise when there is heterogeneity among the path such as a person diverting his interests from one branch to another. Li et al. (2017) used context vectors involving time stamps and features discussed in factors section to quantize the profiles. A context vector is basically a quantized form of words based on semantic operations. Through this, they were able to map several features onto one's profile which made modeling possible. Lou et al. (2010) quantified each industry name by company size and industry for easy application in the model. They used university ranking from US News and divided universities into four categories. Although rankings do not act as perfect criteria for student reputation, it may act as a good indicator for application in the model.

4. **Influential factors identification**

Any model is critical to the inputs intake. So, the correct identification of factors becomes an important task. Learning the stage-sharing and stage-specific features in each career path presents another crucial challenge since different career paths have different influential factors and within each career path also, these factors can vary. Liu et al. (2016), Li et al. (2017) accounted for the time factor in work experience while Lou et al. (2010) has not accounted for the time factor. Liu et al. took demographic factors, LIWC (Linguistic inquiry and word count for accounting personal factors and social traits), user topic features as career-oriented factors. Li et al. (2017) accounted for skills sets, titles, current position, education institute for the NEMO model. Lou and Li considered the current position as a factor which is highly correlated with the next move which became the basis of their model. Subahi’s (2018) research was limited to computer science degrees so factors he took were specific to this arena. He took KA (knowledge area) as the major factor for his model framework. Jamsandekar and Waghmode (2015) did a comparative study on several models and stated that current prediction systems take only some factors into account.

5. **Choosing a model/algorithm**

   Taking reliability and validity factors into account, several prediction models and algorithms have been developed. In general, all of them involve machine learning techniques. Liu et al. (2016) used Multi-source learning framework with a fused lasso penalty (MSLFL). According to this research, this model sought to solve three major challenges - (1) Source Fusion, (2) Temporal Relatedness Modelling, (3) Influential Factors Identification. While Yang et al. (2017), focused on two sources of predictive signals: profile context matching and career path mining and propose a contextual LSTM (Long Short-Term Memory) model, NEMO, to simultaneously capture signals from both sources by jointly learning latent representations for different types of entities (e.g., employees, skills, companies) that appear in different sources. Mimno and McCallum (2008) applied the Topical sequence model. These models are capable of learning underlying hidden topical components in the presence of polysemy and synonymy using the approach of probability transitions. Subahi (2018) work involved a new artificial neural network (ANN) approach for career path prediction (CPP) based on analyzing computer science's body of knowledge (BoK) in degree programs. Lou et al. (2010) used time-homogenous Markov chains and Dijkstra's algorithm to predict the optimal career path. The review of the literature showed that for India, there is clearly visible research gap for career path prediction models. Career development and theorizing could have a major contribution in uplifting the individuals' capacities to operate in the local, national, and global economies; at the proximal level of personal intervention upward to the distal level of influencing corporate and government policy. (Patton et al. 2009). In India, students from rural as well as urban areas are unaware about career selection and it is not economically viable to get career guidance to decide their career paths and for testing their skills, intelligence and interests. Less number of counselors is also a problem. (Jamsandekar, 2015). Moreover, models discussed above could not be directly applied to India because models developed and tested are not generalizable because of specificity of the university, lack of specificity of job titles, varied family cultures across nations, differences in choosing explanatory variables (factors), etc.

This model uses modified K-means clustering to solve lack of job specificity problem, Markov chain to calculate transition probabilities from t-1 to t states, and Dijkstra's algorithm for identifying maximum probability path. For effective application in an Indian setting, appropriate training of the model is done so that factors accounting for career choice other than current position can be treated with minimum variance.

RATIONALE

Some students don’t explore real career possibilities. Technical colleges might play an important role by implementing factors (information, knowledge, skills) that could be applied in their daily studies so as to support them for appropriate career choice. Serious consideration of many alternative choices becomes necessary for career selection. Industries see where, why, and when it could be beneficial for them so as to efficiently invest resources for the purpose of training. Moreover, efficient career planning makes students follow a career plan endowed with informed decision-making. (Michael, 2002).

According to the survey conducted by the Council of Scientific and Industrial Research (2017), about 40% of students are confused about their career options. This may lead to wrong career selection and then reducing the productivity of employees. (Gorad et al, 2017) Machine learning practices can provide a significant contribution by providing support to users for opting the right education domain to shape their career. (Mundra et al, 2014). For effective career selection, we believe that knowing the career path that leads to a particular goal can contribute a lot since it imbibes positions and industries which are encountered and thus can be related to the interest of students/employees.

Existing career path prediction models are specific to the university domain, further fixed by cultural background and other specific environmental factors. In India, to the best of our knowledge, there is no such significant study related to modeling of the career path. Choosing an optimal career path can smoothen the overall development of human capital. Career path modeling could help institutions and policymakers understand patterns and trends in the job market in order to set policy and focus training resources where they can be most effective (Mimno & McCallum, 2008). Realizing the importance of such a model in India, we present the application of a career path prediction model developed by Lou et al., predicting the path which has the maximum probability to be attained.

OBJECTIVES

1. To come up with the career path prediction model which can be applied to an Indian setting.
2. Training the model taking samples independently from privately funded institutions and government-funded institutions.
3. Observe differences/similarities in the career path taken by students from these institutions

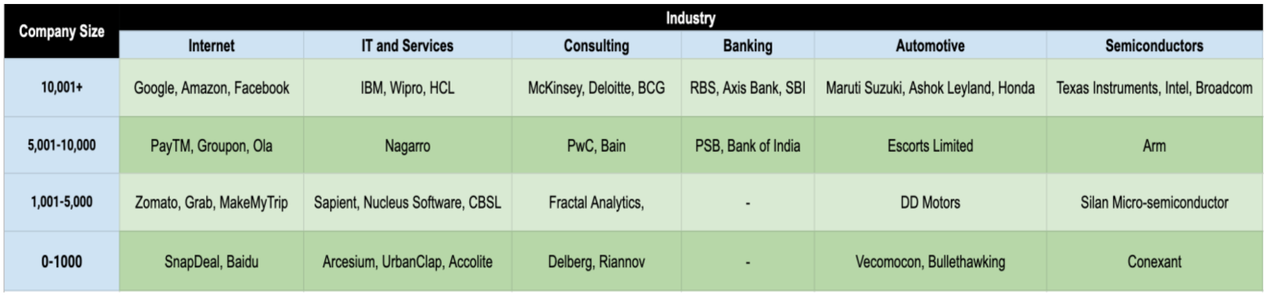
METHODOLOGY

Data Collection

Two privately funded technical institutions and two government funded technical institutions profile data is taken for model development and analysis. These colleges could be taken as a representative of heterogeneous population because these colleges can be considered as independent of the region they are situated in. Moreover, since colleges are in accordance with NIRF ranking, student personal factors (aptitude, knowledge) can be taken with minimum variance across these institutes. Although, institute rank does not necessarily reflect the reputation of student's program, it works well as an indicator and for easy application. We obtained the data of around 2000 alumni from LinkedIn as our data source. Each profile contains some working experience which includes name of company, position, time period, and optional description. Further, for each company, data was obtained which included the industry and the company size. The data was pre-processed in the following way.

Data Pre-Processing

Companies have been classified on the basis of the industry and the company's size (given by number of employees). For instance, as per the data obtained from LinkedIn, Google India lies under the industry "Information Technology and Services" and has "10,001+ employees". For position titles, our major challenge was lack of specificity of job titles. This could lead to problem of inconsistency. Moreover, it is impossible for us to handwrite rules for classifying those positions into reasonable number of categories. Therefore, we ran modified K-means clustering algorithm.



Company Size	Industry					
	Internet	IT and Services	Consulting	Banking	Automotive	Semiconductors
10,001+	Google, Amazon, Facebook	IBM, Wipro, HCL	McKinsey, Deloitte, BCG	RBS, Axis Bank, SBI	Maruti Suzuki, Ashok Leyland, Honda	Texas Instruments, Intel, Broadcom
5,001-10,000	PayTM, Groupon, Ola	Nagarro	PwC, Bain	PSB, Bank of India	Escorts Limited	Arm
1,001-5,000	Zomato, Grab, MakeMyTrip	Sapient, Nucleus Software, CBSL	Fractal Analytics,	-	DD Motors	Silan Micro-semiconductor
0-1000	SnapDeal, Baidu	Arcesium, UrbanClap, Accolite	Delberg, Riannov	-	Vecomocon, Bullethawking	Conexant

Figure 1: Examples of companies taken in particular category denoted by industry and company size

Clustering

1. **Semantic Similarity**

   WordNet::Similarity project (Wordnet, 2006) is a Perl module that implements a variety of semantic similarity measures based on information from lexical database WordNet (Lou et al., 2010). The similarity between any two position titles is given calculated by WordNet and defined by a number between 0 and 1.

2. **Algorithm**

   k was chosen to be 40 and words with semantic similarity index 0.5 were clustered together. A Python script was run which clustered the position titles into 40 categories. Each category was assigned a unique job title manually on the basis of the jobs it contained.

3. **Clustering results**

   After clustering, semantically similar positions were clubbed together. For example, “Software Developer” and “Software Development Engineer” have been grouped with

“Programmer”.

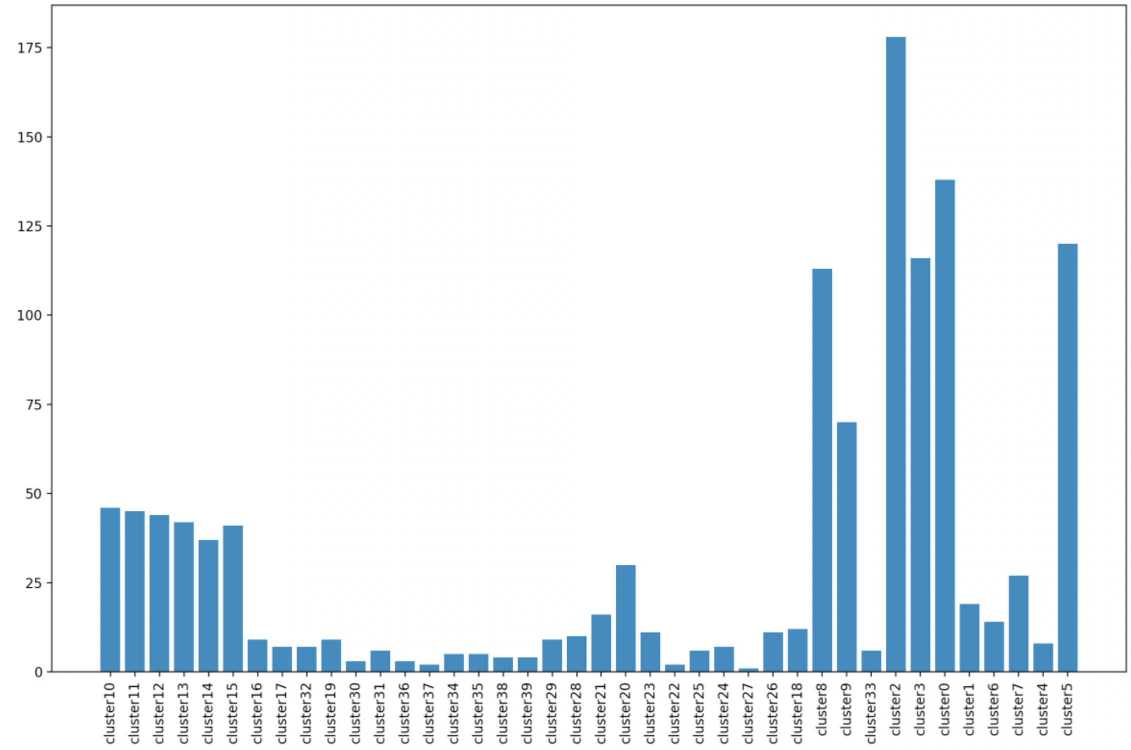


Figure 2: Sample clustering result (k=40, similarity=0.5)

Model

1. Markov Chain

For each person, a series of jobs has been obtained. We assume that a person’s current job depends only on his/her previous job. This is a simplified model of the real world scenario but it can be argued that previous job influences the current job by a greater degree. Thus, we can model each person’s career path as a Markov chain where the next stage only depends on the current state.

$$\Pr(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} = x \mid X_n = x_n)$$

We further assume that the career path is a time homogeneous Markov chain, so that the process is described by a single, time-independent transition matrix (Lou et al., 2010).

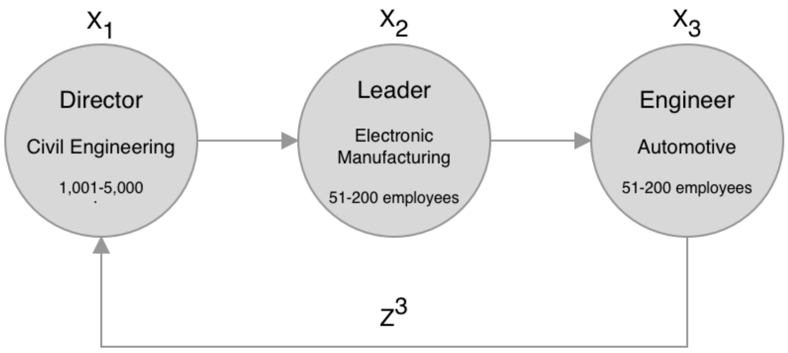
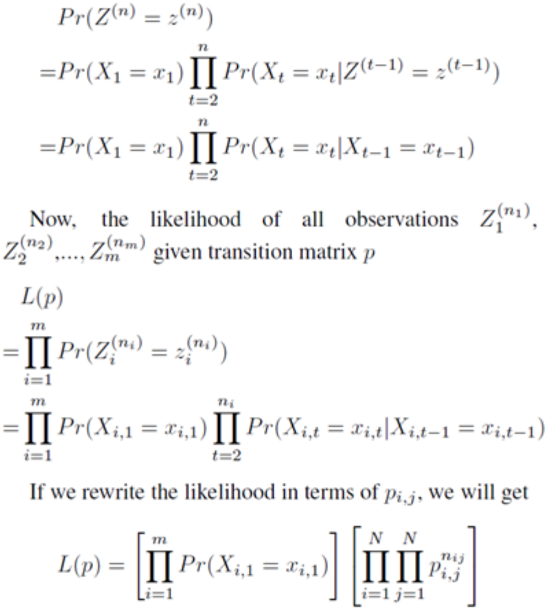


Figure 3: Sample Markov Chain

2. **Probability Estimation**

Transition matrix is denoted by p(N,N) , where N is the total number of different states and entry p(i,j) represents the probability of going from state i to state j. Observation z(n) of a person's profile comprising of a sequence of states x_1,x_2,………x_n, and we also think a career path of length n is a random variable Z(n). Then the probability of Z(n) taking value of z(n).



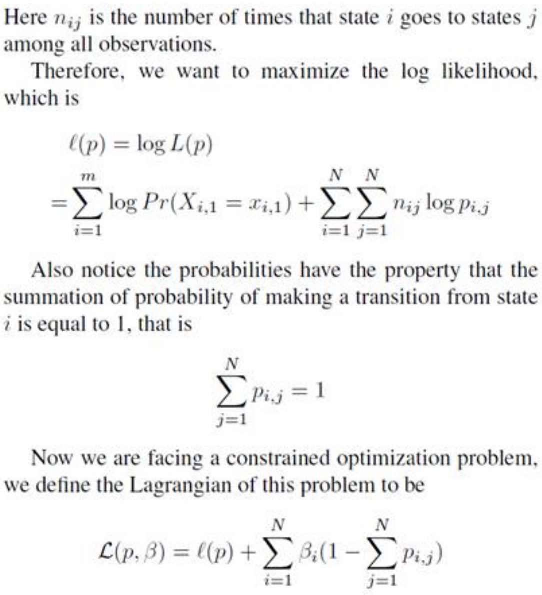
$$
\begin{aligned}
&Pr(Z^{(n)} = z^{(n)}) \\
=&Pr(X_1 = x_1)\prod_{t=2}^{n} Pr(X_t = x_t | Z^{(t-1)} = z^{(t-1)}) \\
=&Pr(X_1 = x_1)\prod_{t=2}^{n} Pr(X_t = x_t | X_{t-1} = x_{t-1})
\end{aligned}
$$

Now, the likelihood of all observations $Z_1^{(n_1)}$, $Z_2^{(n_2)}$,..., $Z_m^{(n_m)}$ given transition matrix p

$$
\begin{aligned}
&L(p) \\
=&\prod_{i=1}^{m} Pr(Z_i^{(n_i)} = z_i^{(n_i)}) \\
=&\prod_{i=1}^{m} Pr(X_{i,1} = x_{i,1})\prod_{t=2}^{n_i} Pr(X_{i,t} = x_{i,t} | X_{i,t-1} = x_{i,t-1})
\end{aligned}
$$

If we rewrite the likelihood in terms of $p_{i,j}$, we will get

$$
L(p) = \left[\prod_{i=1}^{m} Pr(X_{i,1} = x_{i,1})\right]\left[\prod_{i=1}^{N}\prod_{j=1}^{N} p_{i,j}^{n_{ij}}\right]
$$



Here n_{ij} is the number of times that state i goes to states j among all observations.

Therefore, we want to maximize the log likelihood, which is

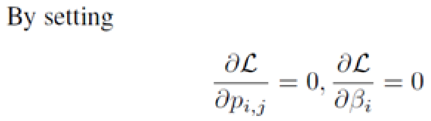
$$\begin{aligned}\ell(p) &= \log L(p)\\ &= \sum_{i=1}^{m} \log Pr(X_{i,1} = x_{i,1}) + \sum_{i=1}^{N}\sum_{j=1}^{N} n_{ij} \log p_{i,j}\end{aligned}$$

Also notice the probabilities have the property that the summation of probability of making a transition from state i is equal to 1, that is

$$\sum_{j=1}^{N} p_{i,j} = 1$$

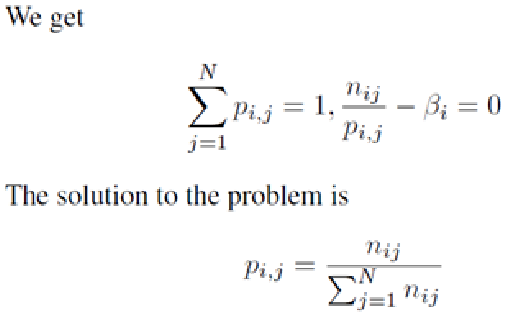
Now we are facing a constrained optimization problem, we define the Lagrangian of this problem to be

$$\mathcal{L}(p, \beta) = \ell(p) + \sum_{i=1}^{N} \beta_i (1 - \sum_{j=1}^{N} p_{i,j})$$



By setting

$$\frac{\partial \mathcal{L}}{\partial p_{i,j}} = 0, \frac{\partial \mathcal{L}}{\partial \beta_i} = 0$$

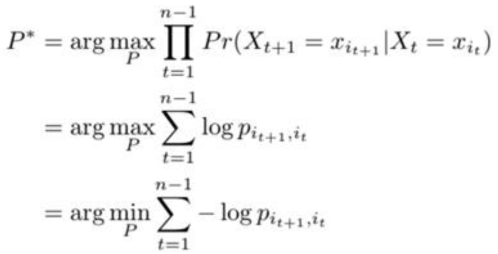


We get

$$\sum_{j=1}^{N} p_{i,j} = 1, \frac{n_{ij}}{p_{i,j}} - \beta_i = 0$$

The solution to the problem is

$$p_{i,j} = \frac{n_{ij}}{\sum_{j=1}^{N} n_{ij}}$$



$$\begin{aligned}P^* &= \arg\max_P \prod_{t=1}^{n-1} Pr(X_{t+1} = x_{i_{t+1}} | X_t = x_{i_t})\\ &= \arg\max_P \sum_{t=1}^{n-1} \log p_{i_{t+1}, i_t}\\ &= \arg\min_P \sum_{t=1}^{n-1} - \log p_{i_{t+1}, i_t}\end{aligned}$$

So, p(i,j) ,the probability of going from state i (time: t-1) to state j (time: t) will be no. of times transition from i^th^ state to j^th^ state is observed divided by no. of transitions from i^th^

state to all the states(1 to Nth).

3. **Path Prediction**

Once we have the probability estimation model ready, we can use it to predict the probability of each transition (from one job description to another). A graph is constructed where each job description is taken as a node. If a transition from one job to another exists, an edge is created.

4. **Shortest Path**

Weight of each edge is given by the probability as defined by Markov chain. The objective is to find the shortest path having maximum probability between two given job descriptions. To use Dijkstra's shortest path algorithm, we have to convert this problem into a path minimization problem. We can do it in the following way (Lou et. al, 2010):

$$
\begin{aligned}
P^* &= \arg\max_P \prod_{t=1}^{n-1} Pr(X_{t+1} = x_{i_{t+1}} | X_t = x_{i_t}) \\
&= \arg\max_P \sum_{t=1}^{n-1} \log p_{i_{t+1}, i_t} \\
&= \arg\min_P \sum_{t=1}^{n-1} - \log p_{i_{t+1}, i_t}
\end{aligned}
$$

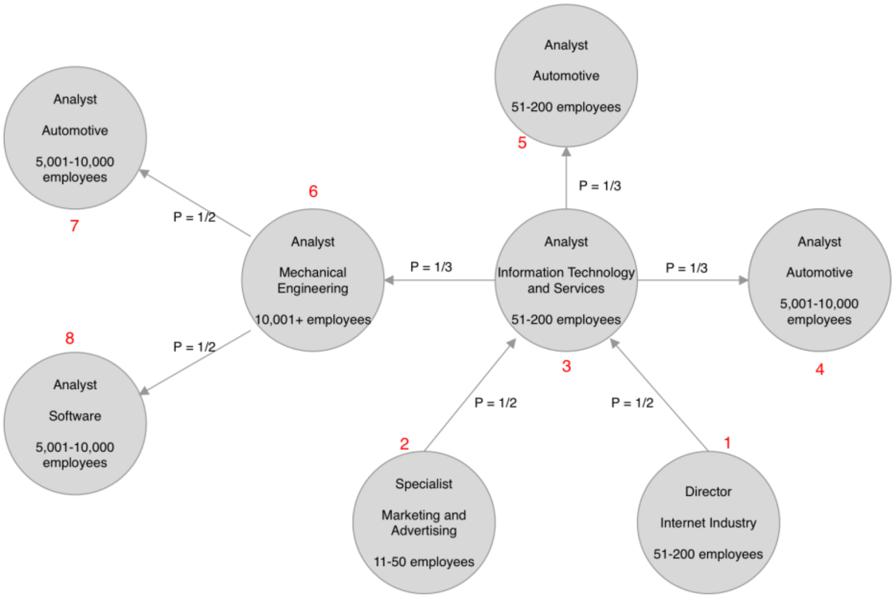


Figure 4: Career path graph with conditional probabilities

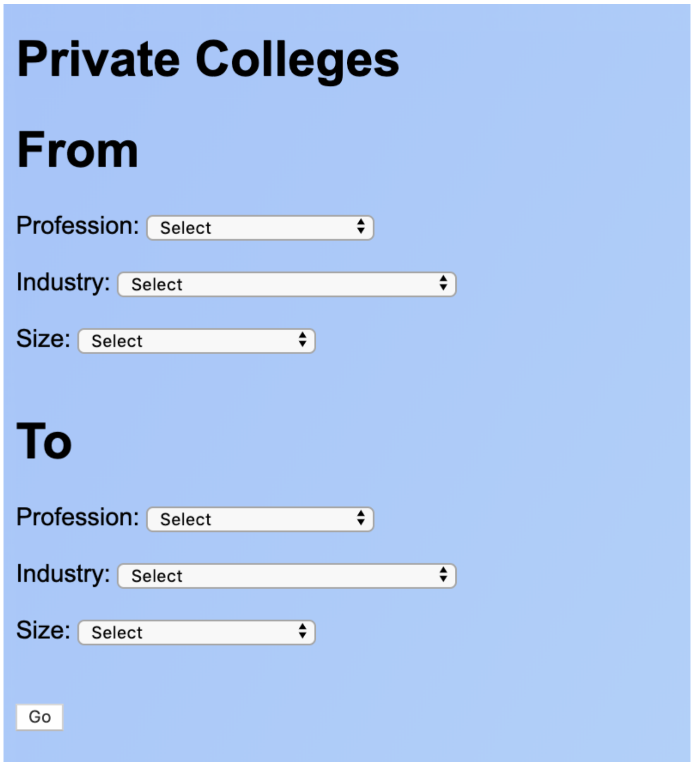
The above graph (Figure 4) is modelled using career paths of 3 people. If someone wants to envision the probability and path of going from job description 1 to job description 4 (or 7), two paths exist: 1 to 3 to 4 and 1 to 3 to 6 to 7. Path 1 yields a probability = ⅙ and path 2 yields a probability = $^{1}/_{12}$. Clearly, path 1 should be chosen over path 2 due to high probability and shorter length. In a case when probabilities come out to be same, the path with shortest length should be opted.

5. Web app development

A web app was developed to visualise the shortest career paths of private and government institutions, given the “from” position and the “to” position - specified by profession, industry and company size.

Private Institutions: https://sarthak-sehgal.github.io/TSR/webapp/private.html

Government Institutions: https://sarthaksehgal.github.io/TSR/webapp/government.html



Private Colleges

From

Profession: Select

Industry: Select

Size: Select

To

Profession: Select

Industry: Select

Size: Select

Figure 5: Web app to determine probable career path

To support the web app and visualize the career paths followed by all the individuals, a graph visualizer was also developed (see Figure 8 and Figure 9).

Private Institutions: https://sarthak-sehgal.github.io/TSR/webapp/privateGraph.html

Government Institutions:

https://sarthaksehgal.github.io/TSR/webapp/governmentGraph.html

RESULTS AND DISCUSSIONS

Objective 1: To come up with the career path prediction model which can be applied to an Indian setting

As evident in literature review, we looked upon several models and the way they solved several problems relating to career path modelling. To apply in an Indian setting, we chose Lou et al. (2010) model involving use of Markov chain and Dijkstra’s algorithm. To solve the problem of lack of specificity of job-titles, modified k-means clustering is used involving concept of semantic similarity. Current state is taken as the sole indicator of the next state and thus career path is understood as a representation of time-homogeneous Markov chain. Factors such as region, personal, college type have been considered as extrinsic and appropriate sampling and hence appropriate training of data is assumed to nullify the variance effect of these factors. Other factors under socio-environmental and socio-economic heads do affect career paths to a greater extent. But to reduce complexity of the model, these factors are ignored. After obtaining transition probabilities through Markov chain, Dijkstra's algorithm is used to obtain optimal path. In this way we came up with the model to apply in India.

Objective 2 and 3: Training the model taking samples independently from privately funded institutions and government-funded institutions and observing the differences/similarities among the two.

Above model was developed separately for privately funded institutions and publicly funded institutions taking sample size approximately 1000.The model has generated a graph containing all the possible career paths in the data collected. On a given query, the model suggests logical career path recommendations.

Following graphs show visualised data:

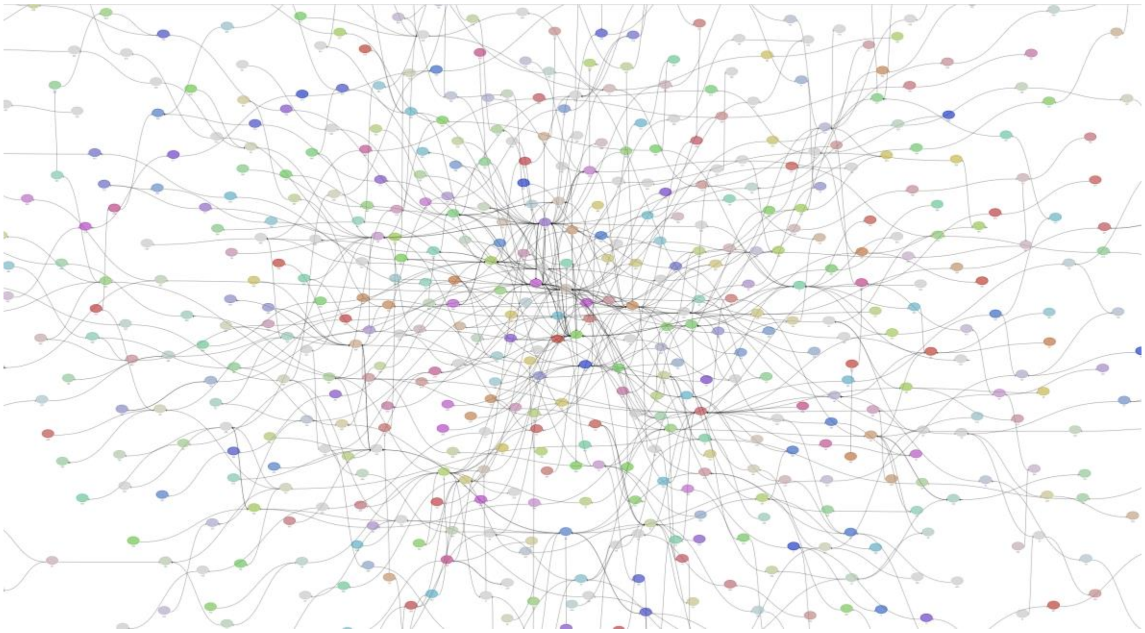


Figure 6: Private institutions' individuals' career paths visualised

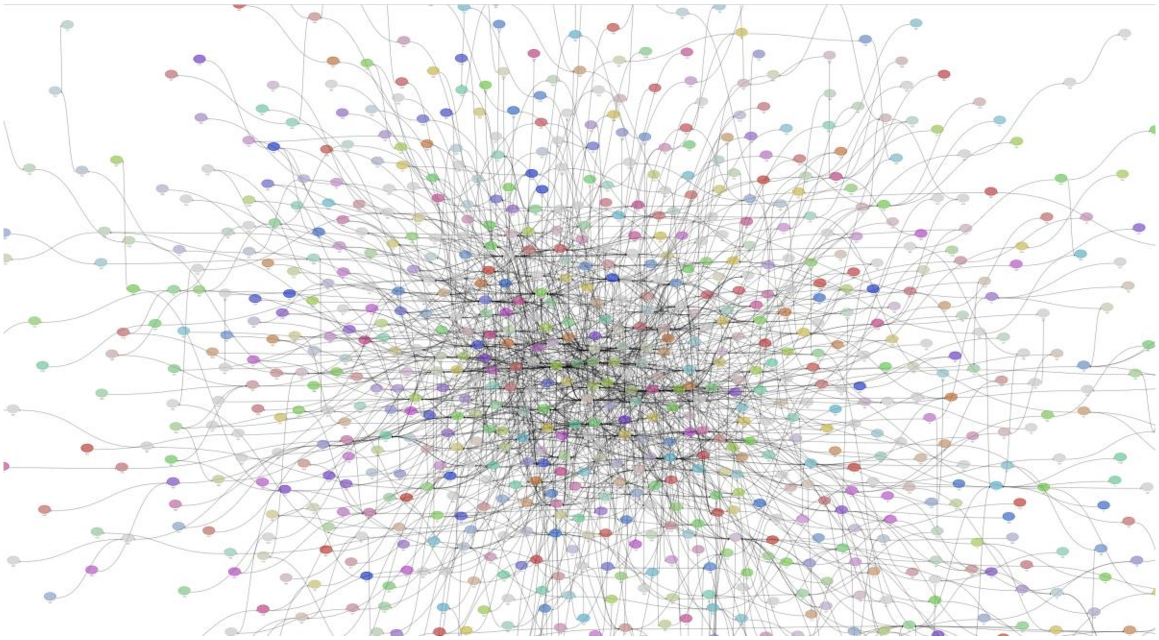


Figure 7: Government institutions' individuals' career paths visualised

Above graphs show data collected from respective institutions visualized through nodes connected paths. It can be observed that private graph is more sparsely located than government graph, which is densely located.

Analysis of data

1. On the basis of industry

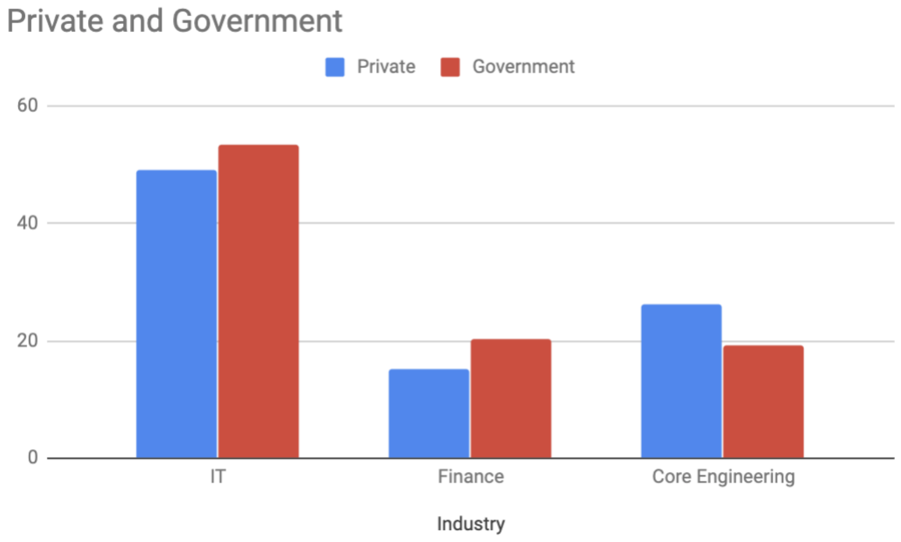


Figure 8: Comparison between private and government institutions' data w.r.t. industry

Analysing the data industry wise, three sectors - IT, Finance, and Core Engineering were looked upon. IT sector comprised of Internet, Computer Software, Computer Hardware, Computer Networking, Computer and Network Security, Information Technology and Services, and Information Services industries. Finance sector comprised of Management Consulting, Financial Services, Banking, Investment Management, Investment Banking, Venture Capital and Private Equity, Insurance. Core Engineering comprised of Civil Engineering, Electrical/Electronic Manufacturing, Biotechnology, Mining and Metals, Mechanical or Industrial Engineering, Machinery, Industrial Automation, Textiles, Railroad Manufacture, Aviation/Aerospace, Architecture and Planning, Oil and Energy, Construction, Automotive, and Semiconductors. IT sector can be seen to be chosen most which might be due to inclination of students in India towards IT (Gopu 2016) and also since the sample is taken from technical institutions. IT sector constitutes larger share in government institutions (53%) than private institutions (49%). Interestingly, private institutions have higher contribution in Core Engineering (26%) than government institutions (19%). For financial sector, private institutions have lesser contribution (15%) when compared with government institutions (20%). From this, we can observe that students have chosen financial and IT sector more in government institutions while students have chosen core engineering sector more in private institutions.

2. On the basis of profession

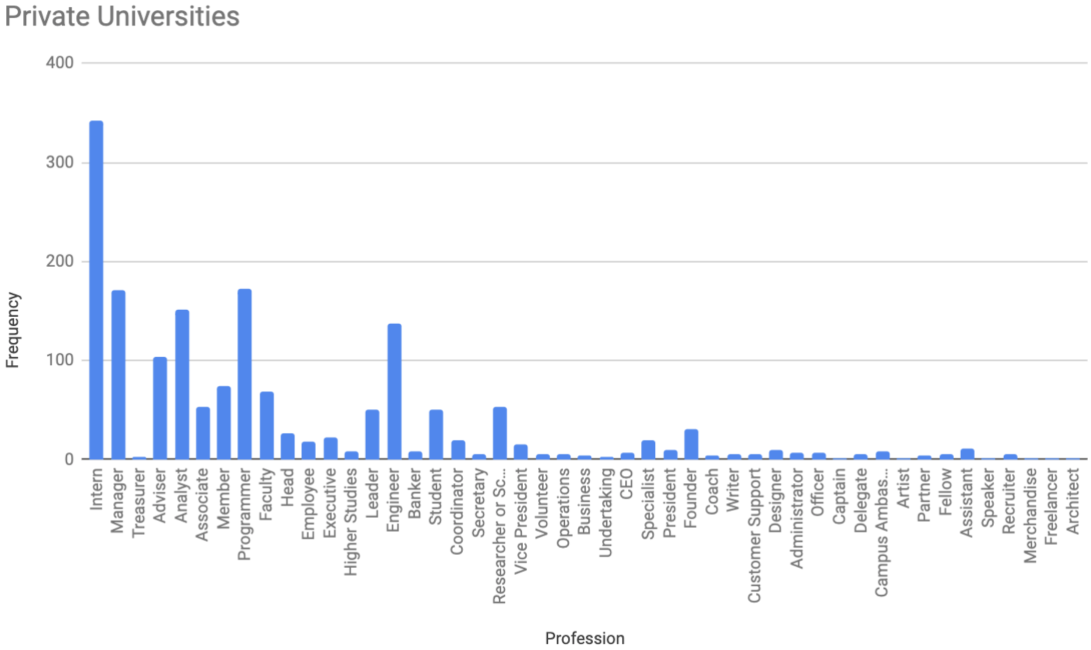


Figure 9: Clustering result - Private Institutions

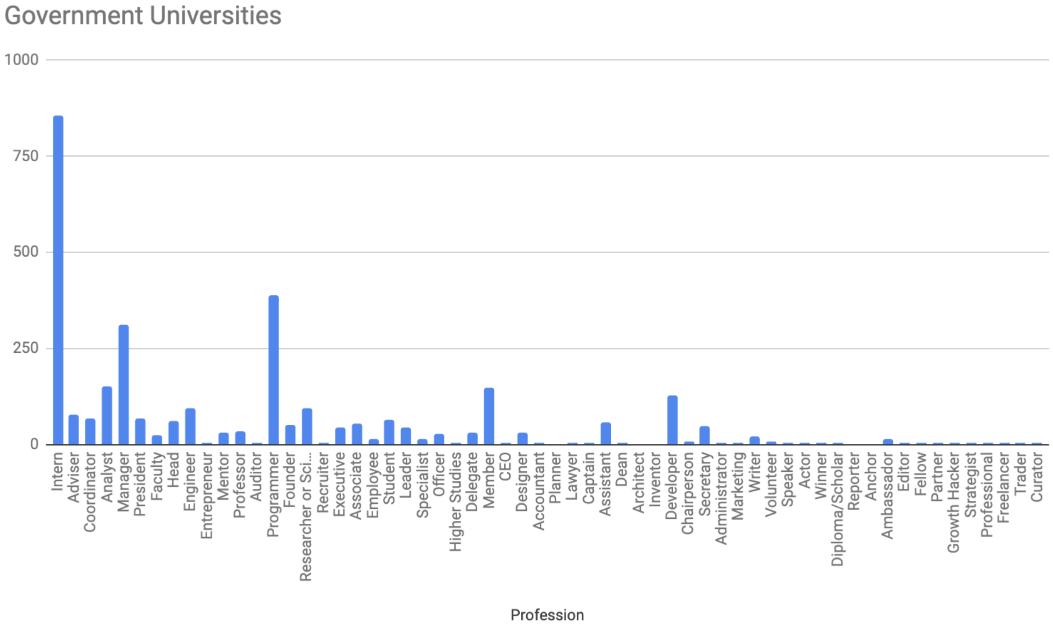
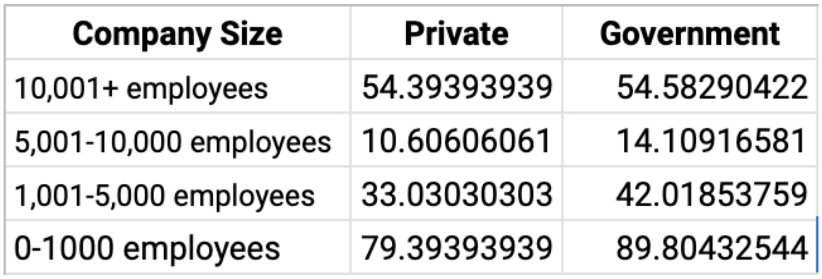


Figure 10: Clustering results - Government institutions

From both the graphs, we can observe intern constitute the largest section of position in career path for both public and private technical institutions which is logically consistent since doing internship prior to the job is very common. Profession of programmers is also observed to be involved in most career paths in both the graphs showing inclination towards this sector.

3. **On the basis of company size**



Company Size	Private	Government
10,001+ employees	54.39393939	54.58290422
5,001-10,000 employees	10.60606061	14.10916581
1,001-5,000 employees	33.03030303	42.01853759
0-1000 employees	79.39393939	89.80432544

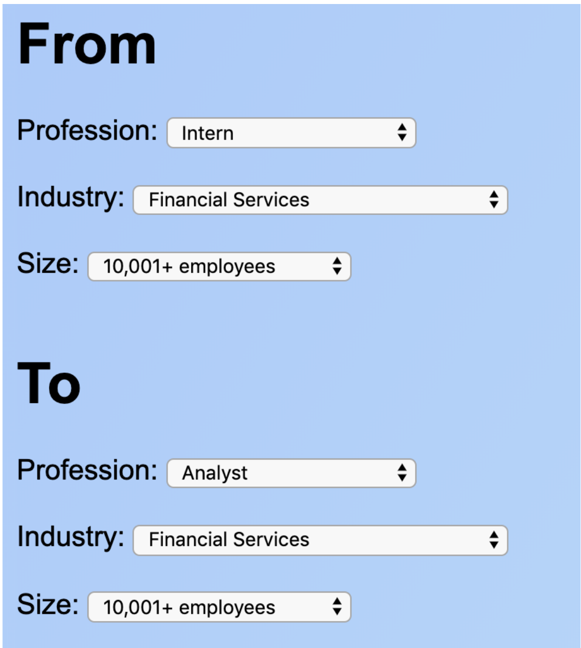
Figure 11: Jobs classified on the basis of company size

From above data, 1000+ employees company had similar contributions in both private and government institutions, while difference was observed in company size of 0-1000 employees whose contribution was more in Government institutions (89%) than in Private institutions (79%).

After training the model, similarities / differences in career paths between private and public funded institutions are observed through web application developed. Probability estimated is interpreted as the possibility to reach the goal, given the present position.

Analysis on the basis of predicted paths by the model:

1. Financial Sector



From

Profession: Intern

Industry: Financial Services

Size: 10,001+ employees

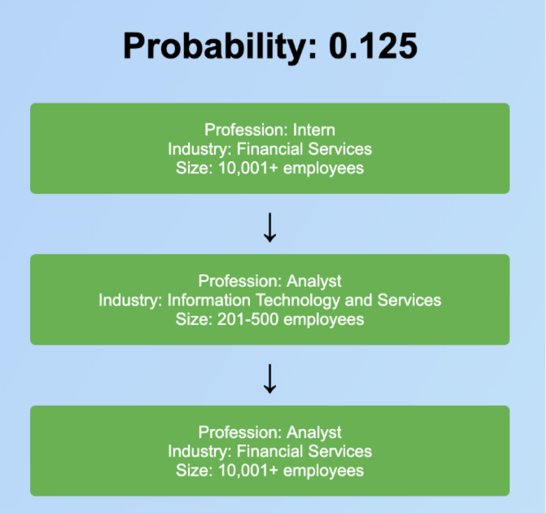
To

Profession: Analyst

Industry: Financial Services

Size: 10,001+ employees

Figure 12: Current position and goal



Probability: 0.125

Profession: Intern
Industry: Financial Services
Size: 10,001+ employees

↓

Profession: Analyst
Industry: Information Technology and Services
Size: 201-500 employees

↓

Profession: Analyst
Industry: Financial Services
Size: 10,001+ employees

Figure 13: Private institutions result

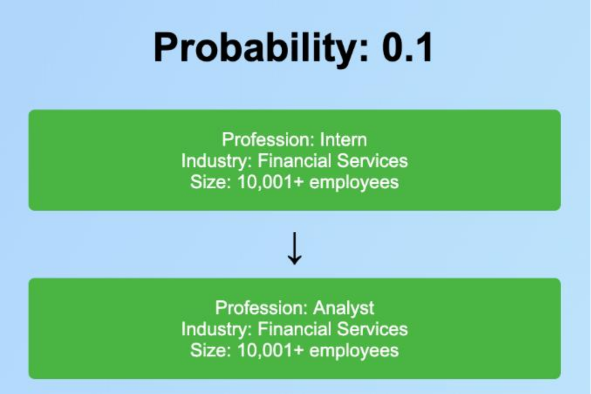
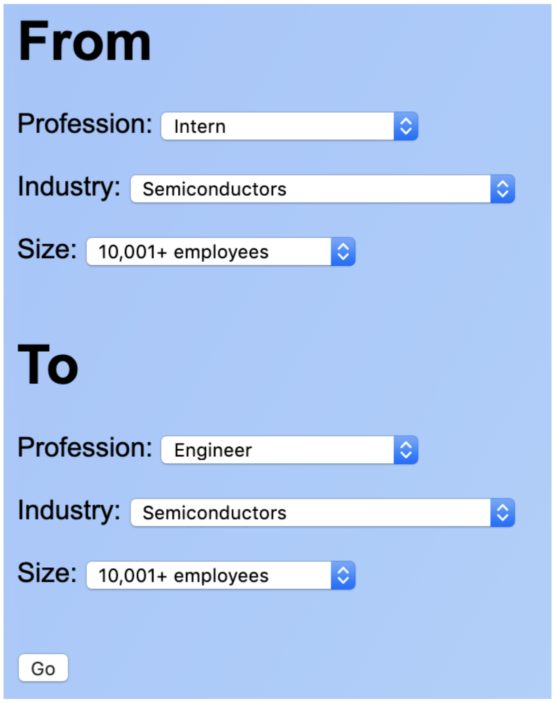


Figure 14: Government institutions' result

Comparing both the probabilities, 0.1 came for government institutions whereas 0.125 came for private institutions. These are almost similar showing the same trend in both the college types. On observing the path there was addition of one node in between for private institutions, node of analyst in industry of Information Technology and Services. Moreover, probability is little more for private institutions. Probably, working as an analyst provided an edge and increased chances of being an analyst in financial services.

2. Core Engineering Sector



From

Profession: Intern

Industry: Semiconductors

Size: 10,001+ employees

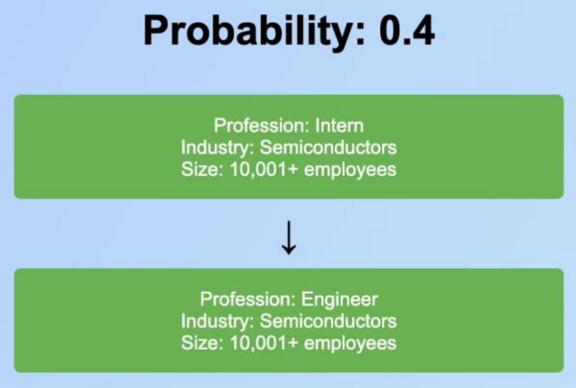
To

Profession: Engineer

Industry: Semiconductors

Size: 10,001+ employees

Figure 15: Current position and goal for core engineering sector



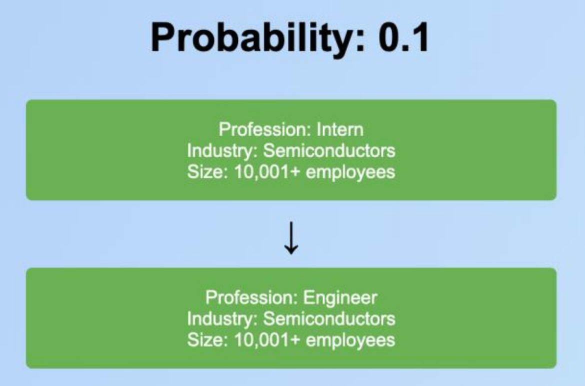
Probability: 0.4

Profession: Intern
Industry: Semiconductors
Size: 10,001+ employees

↓

Profession: Engineer
Industry: Semiconductors
Size: 10,001+ employees

Figure 16: Private institutions' result for core engineering sector



Probability: 0.1

Profession: Intern
Industry: Semiconductors
Size: 10,001+ employees

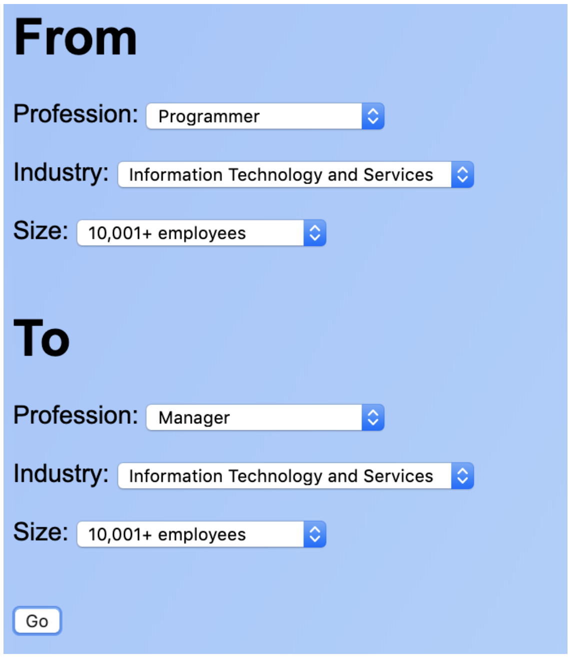
↓

Profession: Engineer
Industry: Semiconductors
Size: 10,001+ employees

Figure 17: Government institutions' results for core engineering sector

On observing the probabilities, 0.1 came for Government Institutions while 0.4 came for Private Institutions. Optimal paths recommended by the model are same. We can infer that people from Private institutions have chosen core branches more (here engineering in semiconductors industry) as compared with people from Government Institutions which led the model to conclude that there are increased chances for a person belonging to private institution to reach his/her goal stated above.

3. IT Sector



From

Profession: Programmer

Industry: Information Technology and Services

Size: 10,001+ employees

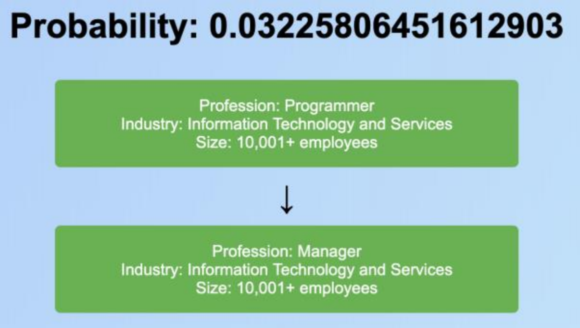
To

Profession: Manager

Industry: Information Technology and Services

Size: 10,001+ employees

Figure 18: Current position and goal for IT Sector



Probability: 0.03225806451612903

Profession: Programmer
Industry: Information Technology and Services
Size: 10,001+ employees

↓

Profession: Manager
Industry: Information Technology and Services
Size: 10,001+ employees

Figure 19: Private Institutions' result for IT sector

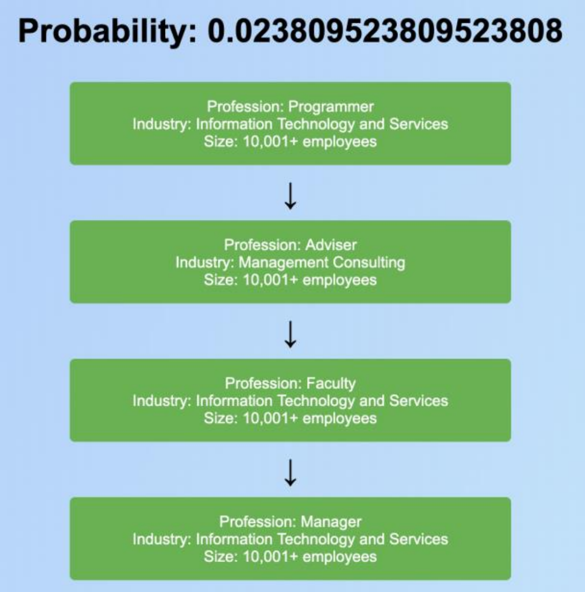

Probability: 0.023809523809523808
Profession: Programmer
Industry: Information Technology and Services
Size: 10,001+ employees
Profession: Adviser
Industry: Management Consulting
Size: 10,001+ employees
Profession: Faculty
Industry: Information Technology and Services
Size: 10,001+ employees
Profession: Manager
Industry: Information Technology and Services
Size: 10,001+ employees

Figure 20: Government institutions' result for IT sector

On comparing the probabilities, both institutions showed similar probabilities (0.02 - 0.03). However, recommended path is longer for government institutions. Adviser in consulting and Faculty in IT nodes are added for the government institutions, though the probabilities are coming similar. This shows that additional working experience is recommended for government Institutions.

CONCLUSIONS AND IMPLICATIONS

1. On comparing the career paths followed by the individuals of private and government institutions, we found some relevant results with promising probabilities which could help in choosing the correct path in a real-world scenario. For instance, as mentioned above, there is a 40% chance of an alumni of private institution currently working as an Intern in Semiconductors industry to land up as an Engineer in the same industry. Whereas, for the same source and destination, an individual of government institution is likely to follow a different path with probability 10%. This is in accordance with the statistics mentioned earlier that alumni of private institutions have worked in core industries more than alumni of government institutions.

2. Accuracy and consistency of the model majorly depends upon sample size. We tried to develop a basic model to set up the stage for future improvements and research in this field especially in India. So, collecting more data, taking sampling style (type of college and other extrinsic factors) into account could be of great help.
3. Time factor apart from probability could also be included in this model so as to give user an idea of an approximate/optimal time one needs to get through the path.
4. Several socio-economic factors (financial constraints, economic stability) , personal (beliefs), and socio-environmental factors (family, peer group, schooling) have been assumed to show negligible effect, though this assumption doesn’t seem appealing. So, improvement can be done in this model by introducing these features (through appropriate measuring techniques) in the nodes.

5. Raw data collected from LinkedIn is represented in natural language format which added to the task of data pre-processing. The entries in the raw data highly vary from person to person depending upon their input. Many position titles contain more details than what we desire for example- Computer science engineer and CEO. So, improvement in quality of transition from raw data in natural language format to the training input in mathematical format is needed.
6. During education to job transition, our model doesn’t account branch of the student involved in the transition which can affect the accuracy and quality of the model. Although, this problem is partially solved by the industries they work in as an intern, but this model calls for improvement in this arena.
7. Higher positions in the same title are treated as similar so as to perform clustering task accurately. For example, Associate professor, Assistant professor etc. are taken to be as professor. Therefore, including these positions is another improvement we suggest.
8. This study is based on quantitative data and it can be extended further to include qualitative data by employing interview method. Further, to strengthen the results, parametric statistics could be used and tested.

REFERENCES

- Afaq, K., Sharif N., Ahmad, N. (2017). ‘Factors Influencing Students’ Career Choices: Empirical Evidence from Business Students ‘. *Journal of Southeast Asian Research*. Vol. 2017,pp. 1-15.
- Beggs, J.M, Bantham, J.M. ,Taylor, S.(2008).’ Distinguishing the factors influencing college students’ choice of major’. *College Student Journal*. Vol. 42, Issue 2,p. 381. Gopu, M.. (2016).’ A review paper: Student attitude towards computer science’. *International Journal Of Pharmacy & Technology*. Vol. 8, Issue 3, pp. 4653-4666.
- Borchert, M. (2002). ‘ Career Choice Factors of High School Students ‘, University of Wisconsin-Stout.
- Dick, T.P., Rallis, S.F. (1991).’Factors and Influences on High School Students' Career Choices’. *Journal for Research in Mathematics Education*. Vol. 22, No. 4 , pp. 281-292.
- Dodge, A. , Welderufael, M. (2014).’ Factors that Influence Career Choice in South African Township High School Students’. *Graduate Master's Theses, Capstones, and Culminating Projects*.
- Fizar,D. (2013).’Factors Affecting Career Choices of College Students Enrolled in Agriculture’,University of Tennessee, Martin.
- Gorad, N., Zalte, I., Nandi, A., Nayak, D. (2017). ‘Career Counselling using Data Mining’. *International Journal of Innovation Research in Computer and Communication Engineering*. Vol. 5, Issue 4.

- Li, L., Jing, H., Tong, H., Yang, J., He, Q., Chen, B.C.(2017). ' NEMO: next career move prediction with contextual embedding'. *Proceedings of the 26th International Conference on World Wide Web Companion*.
- Liu,Y.,Zhang,L.,Nie,L.,Yan,Y.,Rosenblum,D.S. (2016). 'Fortune Teller: Predicting your Career Path', *Proceedings of Association for the Advancement of Artificial Intelligence*, pp. 201-207.
- Lou, Y.,Ren, R., Zhao, Y. (2010), 'A Machine Learning Approach for Future Career Planning'. Technical Report, Stanford University.
- Mimno, D. and McCallum, A. (2008). ' Modelling career path trajectories ', University of Massachusetts,
- Mtemeri, J. (2017).'Factors Influencing the Choice of Career Pathways Among High School Students in Midlands Province, Zimbabwe', University of South Africa, Muckleneuk, Pretoria, South Africa.
- Mundra, A., Soni, A., Kumar Sharma, S., Kumar, P., Chauhan, D. (2014). 'Decision Support System for Determining: Right Education Career Choice'. *Elsevier Publication*.
- Olamide, S.O., Olawaiye, S.O. (2013). 'The factors determining the choice of career among secondary school students'. *The International Journal of Engineering and Science.* Vol. 2,Issue 6, pp. 33-44.
- Ooro, O.H., Omari, S., Mong'are, O. (2017).'An Assessment of Factors Influencing Career Choices Among University Students: A Survey of Students in The School of Business And Economics, Kisii University'. *IOSR Journal of Humanities and Social Science*. Vol. 22, Issue 11,pp. 82-91.

- Patton, W., & McIlveen, P. (2008). ‘Annual Review: Practice and research in career counseling and development -2008’. *The Career Development Quarterly.* Pp. 1-80.
- Patwardhan, S., & Pedersen, T. (2006). ‘Using wordnet-based context vectors to estimate the semantic relatedness of concepts’. *In Proceedings of the Workshop on Making Sense of Sense at the 11th Conference of the European Chapter of the Association for Computational Linguistics* , pp. 1–8.
- Subahi, A.F. (2018). ‘Data Collection for Career Path Prediction Based on Analysing Body of Knowledge of Computer Science Degrees’. *Journal of Software.* Vol. 13, No. 10, pp. 533-546
- Waghmode M. L and Jamsandekar P.P (2015). ‘ A Study of Expert System for Career Selection: Literature Review ‘ . *International Journal of Advanced Research in Computer Science and Software Engineering.* Vol. 5 , Issue 9, pp. 779-785.
- Zody, Z ., MacDermid, S., Schrank, H., Sprenkle, D. (2006). ‘Boundaries and the functioning of family and business systems ‘. *Journal of Family and Economic Issues*. Vol. 27, Issue 2,pp. 185-206.