Hierarchical Visualization of Naukri Dataset

Sethuram Ramalinga Reddy

STUDENT NUMBER 17311993

Abstract

In this paper, a visualization is created for the Naukri job dataset in the form of a hierarchical tree to represent the total number of jobs available for each skill in every industry located in India's major cities. This gives the detailed explanation on how the data is pre-processed, and various visualization task and encoding techniques are incorporated in the visualization using D3.js. This visualization helps in analyzing the current trends in the job market and also to identify the most in-demand skills that would help the professionals to take career decisions.

1. Introduction

Visualization is the graphical presentation of information such as data, with the goal of providing the viewer with a qualitative understanding of the information contents. The dataset visualized is the Naukri Dataset which is taken from the Kaggle website [nau17], and this has not been visualized previously. The Naukri website gives only the information about the Job description and availability in each sector, but they doesn't give simple summary statistics like total jobs available based on the cities, industry, and skills. Therefore this visualization presents the number of jobs available for each skill in every industry located in India's major cities. The tool that is used for visualizing the dataset is D3.js which is a javascript library used for creating the interactive and dynamic visualization, and it makes use of a tree dataset which is a transformed from original data. The data preprocessing is done in Java and python which includes data cleaning and framing a tree structure that represents the hierarchy.

2. Dataset Description

Naukri.com is an Indian Job Portal that holds a massive database of the jobs, resumes and recruitment consultants. This website is considered as a common platform for the job seekers and hiring people to come together. This website is a common platform that facilitates interactions between job seekers and employers. A dataset was obtained by pre-crawling this website, and it is a subset of a larger dataset which contains more than 9.4 million job listings, created by extracting from the site. The resultant dataset has got 22K records of job descriptions. This dataset is the recently updated dataset that conveys the latest information about the availability of the jobs. This dataset has never been visualized which sets the foundation for the creativity in visualizing the data.

3. Data Preparation

The dataset consists of many columns like Job Id, company, experience etc. which are not suitable for the case study that has been proposed. So, a subset was obtained from the original by selecting only the columns that are mentioned below:

- Job Location
- Industry
- Skills
- Total Number of jobs

Once the features are selected, there are certain records where a certain job is posted for more than one location, and this is represented as “Chennai, Delhi, Hyderabad, Bengaluru". In this case, this particular record is duplicated into four records but with different job location details to avoid the loss of information. The names of the job locations where not standard and they were many different names for the same city that would result in the repetition of information while visualizing it. For, E.g. Chennai and Madras are the new and old name for that city, so they are standardized only to the new name. The records are selected in such a way that the all the four features are present and then if the number of jobs field was empty then it is replaced by one to be suitable for calculation. This Data cleaning process was done in java. Now the cleaned data has to be transformed into a hierarchical tree structure with the Label Jobs in the root layer followed by the cities, industries and the total number of jobs as nested children. This transformation was done using python. The transformation was done without using any packages. This was purely constructed using highly nested JSON data structure. The transformation component has two parts. The first part involves the search function which determines when the new level has to be inserted and when the nested level has to be inserted. The new level can be inserted only when that level is not found in the tree or else it has to do the nested level insertion. This function returns location where the insertion should happen.

Second, a recursive function that takes the input from the search function and starts building a new level and further nested levels. Data is inserted during this process. The entire operation happens dynamically.

4. Visual Tasks

The dataset was visualized in the form of a tree [Tre] so that it can be easy to drill down each level according to the need. The visualized tree has got four levels nested in it. They are

- *First Level represents the cities*
- *The second Level represents the Industries*
- *The third Level represents the Skills*
- *The fourth Level represents the Total number of Jobs*

The visualization was made interactive by giving the drill down flexibility to the user by clicking on every node so that they can avoid looking into the details which are not required. Some animation effects like collapsing the sublevels and expanding the size of the last levels were given so that the user's attention gets increased on that sublevel or last level. The visualization falls under the discover, search and derive [Mun14]. In discover task we have stated the hypothesis which is the claim made by the ministry, and this is verified true or not from the visualization. In search task, the user picks the aspect which they are interested and proceed further. In derive task, the variable total number of jobs is derived from the number of jobs variable.

5. Visual Encoding

All the nodes were indicated using a mark such as a *circle* with a label(text) describing the node and the paths represented in the form of *lines*(mark) were used to indicate the connection between the nodes. All the levels are differentiated with the increase in *size* and also the *brightness* of the *color* with respect to the depth of the levels. The last level which indicates the total number of jobs is encoded as the *motion* which is in the form of an expanding circle with count embedded within it and the size of this circle is large which compared to other levels. The path which they traversed through the tree has been highlighted using the green color which makes the user easier to traverse back and forth since the visualization is pretty huge and this helps in differentiating from the other paths [WGK10].

6. Interaction

Suppose if we have to know the total number of jobs available in the Kolkata in the IT-Software Industry for the application programming then we have to click on Kolkata which displays all the industries available in Kolkata. Now the IT- Software Industry has to be selected, and then it displays the total number of jobs which is 2 in this case. The entire path which we have traversed through the tree for the current scenario is highlighted in green color as mentioned earlier.

6.1. Illustrations, graphs, and photographs

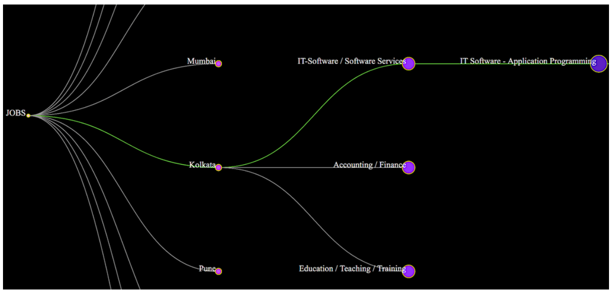
JOBS
Mumbai
Kolkata
Pune
IT-Software / Software Services
Accounting / Finance
Education / Teaching / Training
IT Software - Application Programming

Figure 1: *Visualisation of jobs in kolkata and IT-services in hierarchical tree structure*

7. Strengths and Weaknesses

The strength of the visualization is the nuanced way of representing the data that makes the user interaction more accessible and also attractive. Because of its hierarchical structure, understanding the data gets easier. The weakness of the visualization is when the dataset comes under the scale of Big Data; its complexity gets increased because of the number of depths in the hierarchical structure.

8. Accessing Visualization

Youtube video link: `https://youtu.be/UoCzRPCDQeU`
Code link: `https://github.com/sethuram975351/NaukriViz`

9. Conclusion

The summary statistics like total jobs available based on the cities, industry, and skills which is derived from the dataset is visualized. This visualization provides the good picture of the job market in India for the professionals to get succeeded in their job search phase. This visualization is advantageous for potential employers to analyze the job market. For Naukri, this will help them make informed business decisions on targeting specific sectors (cities and industries) to increase their market share. This helps in analyzing the trends in current job market across all the sectors. From this analysis, the most in-demand skills can be identified that would help the professionals to take career decisions. As a result, Naukri will be recognized as a major player in the recruitment business.

References

[Mun14] MUNZNER T.: *Visualization analysis and design*. CRC press, 2014. 2

[nau17] Jobs on naukri.com. `https://www.kaggle.com/PromptCloudHQ/jobs-on-naukricom`, 2017. 1

[Tre] Collapsible tree. `https://bl.ocks.org/mbostock/4339083`. 2

[WGK10] WARD M. O., GRINSTEIN G., KEIM D.: *Interactive data visualization: foundations, techniques, and applications*. CRC Press, 2010. 2