FIT5145 Assignment – 4

Case Study on Data Science at Indeed

Jaimon Thyparambil Thomas
Student ID : 29566428
Email : jthy0001@monash.student.edu
Monash University

October 14, 2018

Contents

Contents ………………………………………………………………. 2

Project Information ..…………………………………….............. 3 - 4

Introduction ..…………………………………………………. 3

What is indeed? …………………………………………….. 3

Relation to Data Science …………………………………... 3

Data Roles ……..…………………………………………….. 3 - 4

Business Model ……………..……………………………………. 4 - 5

Challenges Faced …………………………………………... 5

Data Characteristics …………………………………………………. 5 - 8

Data Sources ………………………………………………… 5 - 6

Volume ………………………………………………………….. 6

Velocity ………………………………………………………….. 6

Variety …………………………………………………………… 6

Veracity ………………………………………………………….. 6

Variability ……………………………………………………….. 6

Storage …………………………………………………………... 6

Data Processing and Technologies used …………………. 7 - 8

Data Analysis ……………………………………………………….... 8 - 9

Conclusion ……………………………………………………………. 10

References……………………………………………………………. 10

Project Information

Introduction

This report is an attempt at analysing the data models and the technologies employed in 'Indeed', a popular online job portal for handling huge amount of data.

What is Indeed?

Indeed, is one among the most popular site for job portal spread across 60 plus countries available in 28 different languages covering [1]. Their motto is helping people get jobs. Which involves gathering all available jobs worldwide in one website, by crawling over 100+k websites and based on the information provided by the employers. Through this site they help job seekers to search for jobs, post their resumes and research on various companies. Thus, every day they try to connect millions of people to new opportunities.

Relation to Data Science

In indeed about 9.8 jobs are added every passing second with around 200 million unique visitors. Given the enormity of data handled, data science is included at its very core. They use data science in how they rank search results, recommend positions and various skills for job seekers, in estimating salaries for jobs, in providing tips for employers for creating great job descriptions. [2]

Data Roles

Some main roles performed are listed below [3]

Data scientist

- We might have to analyse, visualize and model job search related data.
- To improve performance, we must build machine learning models
- Should have the skills and knowledge of a statistician and machine learning expert
- Should be proficient in Python/R, Spark etc and their scientific libraries.

Data Engineer

- Skilled in extracting, transforming, and loading data
- Integrate with diverse APIs
- Will have work along with analyst, data scientists as well as the data supply side
- Will have to work with relational database

Product Scientist

- Use data mining, and machine learning techniques to understand how job applicants and suppliers are interacting on Indeed, and how it is reflected in data.

- Execute statistically sound tactics for evaluating web pages across business related metrics.
- Work in expanding Indeed's set of tools and techniques used for manipulating and interpreting huge amount of product data.

Business Intelligence Analyst

- Should be able to create dashboards, visualize data, find algorithms, and business tools to use across the organization.
- Should Also have profound knowledge in Python (particularly Pandas and NumPy), R, SAS programming experience.
- Should have an aptitude to analyse trends, compile data into comprehensive reports, and making recommendation based on data.

Business Model

Nowadays unemployment is an acute problem that societies face. In this scenario, job portals like Indeed makes transformative changes by recommending the most eligible job seeker to the best possible employers. From a business perspective, indeed aims to increase click rates in general and conversion of profile matches to successful placements. To achieve these aims, they decided to go beyond basic job searches and added functionalities like job recommendation as a new mode of interaction. As they found out that

- One fourth of the searches in Indeed specified only location without any keywords. As there are many job seekers who don't know what keywords to search for.
- We also know that as we provide recommendations users start feel more personalised experience and find more desired results

As per basic work flow, they create lots of classifiers based on jobs description and resume. For instance, one of the classifiers tries to find minimum years of experience. Also, to differentiate between skill sets like expert and beginner.

Based on these classifiers they try to find out matching cases. That is when job seekers' resume is consistent with the pre-requisite skill sets expected by an employer, a 'match' is obtained. In that case [4]

- For both employer and the job seeker a text 'match' (as a highlighted text) will be shown for matched applications.
- Similarly, the same recommendation system is used for suggesting recommended jobs for the job seeker based on his resume.
- Also, while sending mail for the job seeker suggesting that there is a job vacancy.

As the recommendations becomes better the click rates will increase. It will also help in providing more successful interviews

Some of the main factors that they took into consideration while processing the data are [5]:

- Recommendation set keeps on changing as new jobs keeps on coming.
- Everyday 200+ million unique users visit the site. As a result, they should be able to recommend with limited amount of user data.
- Contents freshness matters as compared to the newer jobs the older jobs has more chance of already been filled

- Jobs will have only a limited supply of seats. So, if we keep on recommending the same job for everyone. The employer will be flood with applications.

Challenges Faced

Scaling into different countries

Initially, indeed was available only for US job market. Now it is available in more than 60 countries and 28 languages. To achieve this, they provided different domain based on countries. Like indeed.com for US and indeed.ca for Canada. Within each server there is an application part and configuration part. In the application they make checks based on domain to show country specific data. Thus, they implemented a system such that only one product is there which is provided in all countries [6]

Job seekers might not know correct job description

To overcome this issue, indeed used the concept of stemming of the words that the job seeker used to search. Like rather than just searching for the exact word match they also search for synonyms etc.

Detecting language and trying to Tokenize data

While crawling the sites not all sites may have provided the language nodes properly. Similarly, while the employers of the employees have entered the job description or uploaded resume they may not have described the language properly. So, they had to come up their own language detector. They implemented their language detector based on naïve Bayesian model. They had to find the correct language because for each language tokenising key words were different as in English if there are white spaces between words Japanese sentences might not have white space. So, they had to find an algorithm to tokenise properly with more recall and precision of the tokens formed.

Localisation may not always provide correct data.

For example, if the user is searching for jobs in Paris but most of the jobs in their description they have mentioned just France. So, these results also must be displayed in the search results. So, they should have to handle cases like that also

Data Characteristics

The various characteristics of the data is been provided below

Data sources

The various data sources were

- crawling through web
- Data Provided by the employer

- Data provided by the job seeker.

Volume

The overall data present in peta bytes with around

- 100 million resumes
- 72 million company ratings
- 500 million salaries

Velocity

The velocity at which data is been added are

- 200 Million unique monthly visitors
- 9.8 jobs added per second globally

Variety

They must face a variety of data like

- Resume in pdf
- Company logos and images as JPEG, PNG etc
- portability in 28 different languages

Veracity

Veracity of the recommendation made may affect due to various factors which has to be handled such as

- When searched for architect the result can be for building architect and software architect thus there are chances of less precision in data.
- When searched for hr it should also show results for human resources. There is a chance of less recall for such searches.
- When they try to tokenize data for job description or search query or resume, we might not be able to tokenise data properly due to the various characteristics of the language. Which might not result in desired search results

Variability

The various types of data that we need to handle are

- Content in different languages like for websites, resumes, job description etc
- Not all logos and image have same resolution and size
- the way user has uploaded resume like pdf etc

Storage

- Large Volume of data is handled using distributed platforms such as HDFS
- Existing structured data is managed using relational database such as MySQL

Data Processing and Technologies used

They use a modified bit torrent kind of system called RAD (resilient artifact distribution) for transferring data or updates between their servers. Which for efficiency transfers only the changes made from the previous iteration by using a modified concept of bit torrent and thus distributing the load among the servers itself.

The search engine is built on Apache Lucene

For analytics they mainly use their own large-scale analytics platform known as Imhotep

One of the main product development principle that they use is measure everything like clicks, memory usage etc. based on these we can analyse and create new features and models.

The way their software architecture is been designed is such that the release of once feature in one internal team is not depended on the other teams. So, this results in taking changes into production much faster. They implemented this by building a box car framework

Initially their recommendation engine was based on initial minimum viable product (MVP) build with Apache mahout from which they shifted to a hybrid offline + online pipeline.

Indeed's production application is run in many data centres. All the information from these data centres such as click stream and other events is been replicated into a central HDFS repository located in our Austin data centre. It is from this data centre that they build their own machine learning models.

The data recommendation system can mainly be implemented based on two types one is content based that's is based on the user's resume's keywords etc next one is behaviour based on user's activity like previous searches etc.

Initially they attempted on a personalised recommendation based on Apache Mahout's user to user collaborative filtering implementations. They did this by feeding clickstream data into a Mahout builder that ran in their Hadoop cluster and thus produced a map of users to recommend jobs. To access this list of recommended jobs for multiple clients they build a new service to provide access to this model during run time.

The various roadblocks that were faced due to using mahout on their traffic where.

- The builder took around 18 hours on Indeed's 2013 click stream which is far smaller than the current data.
- They ran the builder only once a day which meant that the new users joining won't be able to those recommendations until the next day.
- New jobs which were added is not visible as recommendations until the builder ran again
- The data produced after modelling had a size around 50GB, so it was difficult to transfer over a WAN from one data centre to another.

They solved this issue by using the concept of min Hash by finding Jaccard similarities between two users clicking in O(n) complexity.

After transitioning to MinHash they shifted to a hybrid recommendation model which utilised the offline component that builds daily in Hadoop and an online component implemented in memcache which is made up of the current days click activity. Both these models are been combined to create the final set of recommendations per user. The main benefit of this feature is that the recommendation system became more dynamic because now it also takes into consideration the click activity per user.

Data analysis

The type of analysis that indeed mainly uses is a recommender-based analysis. It mainly tries to recommend jobs for the job seeker and tries to find the best possible match for the employer.

One of the main sources of its incomes is paid slots for companies. So, whenever a user enters a relevant search related to a job provided by the paid slot companies then those companies job is been listed in the recommended jobs list. Similarly, they also earn by suggesting jobs for job seeker based on the previous job searches. Where the user can register to notify when there is any new job openings

So, the basic flow on how the data flows is

Initially the job results got by crawling web and from the details entered from the employer they get the list of available jobs. Then these jobs are been classified into various tokens based on requirements for the job and by whom the job is been provided. Like min no of years of experience required. Also, in classifying in expert, beginner etc

Similarly, the resume uploaded by the job seeker is also passed through a classifier which converts the details provided in the resume by the job seeker into tokens.

Then if the requirement tokens from the employer is matched with the tokens from resume uploaded by the job seeker. Then it is considered as a match case and it is informed to the job seeker by highlighting that it is a match case. So that it is easier for the employer to select the resume from the list of available resumes. Similarly, when a job seeker watches a job description if it is a match case then it is highlighted to the job seeker saying that it is a match case thus provides the job seeker some confidence for applying the job as now the chances of acceptance is been increased. If it is a match case and if the job seeker has registered for mail regarding job openings, then a mail to job seeker is been send suggesting this job if the job seeker has searched for the job description keywords before.

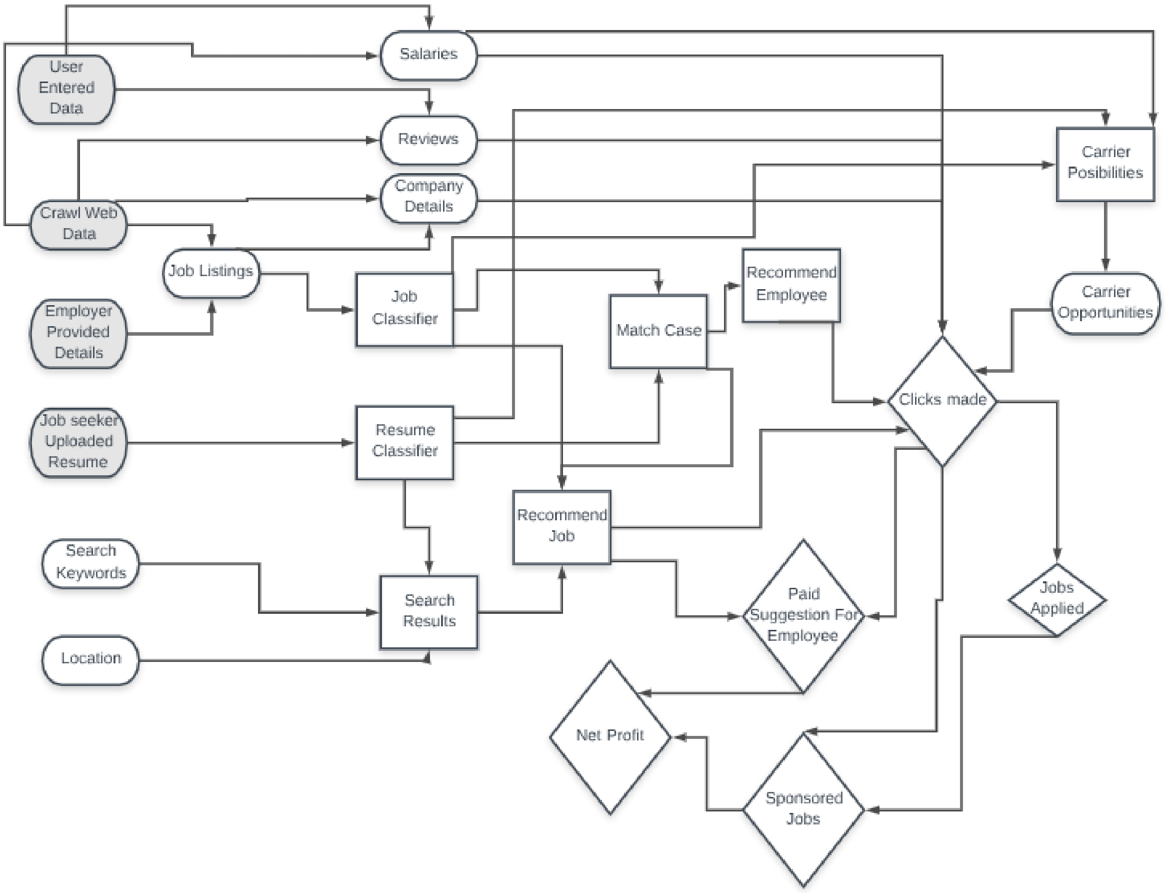
Similarly based on the information from the job seekers resume and search keyword and location entered they can provide a better search result which can result in more no of clicks and thus a greater number of successful conversions

Also, the salaries, reviews and company details such as ratings may influence the search results. Which eventually results more no of clicks and conversion.

Another Suggestion model which I have added to their actual model that's is based on the job classifier and the resume classifiers data they can suggest various carrier paths like what are the various job opportunities currently available what are the extra qualifications I can build over my current qualification. What will be the job opportunities if the job seeker achieves those qualifications? Similarly, what are salaries for the new jobs results. This provides a clear idea to the job seeker that what are the next steps that he must take and thus provides more click rates and in the long run more conversions. They can even charge for those suggestions if they wish or they can do it as a charity service to help for the better future of people. Through which they can gain more publicity.

They also use a predictive analysis by using machine learning for calculating the salaries for various jobs

The following influence diagram shows the some of the main variables, decisions and the data models used to measure the click rates and net profit in Indeed.



Conclusion

Implementing a data science project involves a lot of people with various skill sets and knowledge and usage of various technologies. In this case study I have explained the basic things that they have taken into consideration while implementing the website. I have also proposed a new model which is basically an upgrade of the existing model with few additions. I believe more promising results can be obtained from this model.

References

- [1]https://au.indeed.com/about/our-company?hl=en
- [2] https://medium.com/indeed-data-science/introducing-the-indeed-data-science-blog7e2985fe1e92
- [3] https://www.indeed.jobs
- [4] https://engineering.indeedblog.com/talks/data-to-deployment/
- [5] https://www.oreilly.com/ideas/algorithms-and-architecture-for-job-recommendations
- [6] https://engineering.indeedblog.com/talks/internationalize-success/