Asian Journal of Pharmaceutical and Clinical Research

Innovare Academic Sciences
Knowledge to Innovation

Special Issue (April)

Online - 2455-3891
Print - 0974-2441

Full Proceeding Paper

CAREER RECOMMENDER: A NOVEL APPROACH TO SUGGEST JOBS AND POST-GRADUATION STREAMS

PRAFFUL NATH MATHUR, ABHISHEK DIXIT, SAKKARAVARTHI RAMANATHAN

School of Computing Science and Engineering, VIT University, Chennai, Tamil Nadu, India. Email: asha.s@vit.ac.in

Received: 19 January 2017, Revised and Accepted: 20 February 2017

ABSTRACT

To implement a novel approach to recommend jobs and colleges based on résumé of freshly graduated students. Job postings are crawled from web using a web crawler and stored in a customized database. College lists are also retrieved for post-graduation streams and stored in a database. Student résumé is stored and parsed using natural language processing methods to form a résumé model. Text mining algorithms are applied on this model to extract useful information (i.e., degree, technical skills, extracurricular skills, current location, and hobbies). This information is used to suggest matching jobs and colleges to the candidate.

Keywords: Text mining, Information retrieval, Natural language processing, Résumé extraction, Web crawler.

DOI: http://dx.doi.org/10.22159/ajpcr.2017.v10s1.19758

INTRODUCTION

Our country is on the path of development. Technology sector of our economy is developing by leaps and bounds by the support of various government schemes such as make in India, digital India, and start-up India. As a result, the job opportunity in the various technological domains such as software industry, chemical industry, manufacturing industry, and information security, etc., is increasing day by day. The employers are looking for rightful candidates who are eligible for the job with suitable technical and soft skills. On the other hand, employee is also looking for suitable jobs which require their existing skills. Furthermore, there is a need to improve and enhance the technical knowledge to gain excellence which requires choosing a correct stream such as post-graduation in business line or in technical line [1].

With the advancement in the technology sector, today the industrialists are seeking right candidates for their firms. When a job is posted by a company in market, hundreds or even thousands of applications are received by the human resource (HR) teams. These applications have résumé of the candidate. Moreover, the time needed to read individual résumé and choose a candidate for further selection rounds by the HR or the hiring team can vary from 2 to 4 days. Employees are seeking best job from all available jobs in the market. Graduates are seeking right career path and best institutions for enhance their technical or business skills. It takes a lot of research work by an individual to find the results of above-mentioned problems [2].

Hence, we came with a solution which provides results of all these queries at single platform by a very simple process. As software is created which takes candidate's résumé as input and searches for the skills, a candidate has mentioned in his/her résumé using big data mining and key matching technique to provide the results for a query. An online platform is created by which an employee can find the list of all job available in the market (city/state/country) which matches with his existing technical and soft skills. An employer can find a list of all candidates having skill suitable for the job. A student can find a suitable stream (MBA, MS, M.Com, etc.) which is predicted by our application and also names of colleges where he can apply. Thus, this application reduces the searching time of individuals, provides results of all queries at a single platform, and simplifies the lives of many students.

RELATED WORKS

To provide recommendations based on given input by user, various kinds of models are used, out of which recommender system is one:

Recommender system

It is a way of information filtering which presents a desirable set of information. This has been became very common in our day-to-day life as it is used in the wide range of areas such as music, movie, research papers, articles, books, social tags, and many general things. There are recommendation systems which are available for restaurants (such as zomato app), online dating applications [3], social media application (such as Facebook, LinkedIn, and Twitter) [4] experts, [5] jokes, etc. The recommendation list generated by the recommender system works in two ways through collaborative filtering and [6] content-based filtering.

Content-based filtering

This approach uses series of discrete characteristics of an item, and based on these characteristics, new item set are predicted having similar properties. This model is what used by various marketing companies over a century which required opinions of others to predict something [7].

Collaborative filtering

This approach uses a model which takes users past behavior as well as a similar decision taken to predict the item sets which user finds useful. This model is used by all the online shopping services such as Flipkart, Amazon, eBay, and Snapdeal, website such as YouTube to suggest video based on its past viewing history, and by the job searching websites such as naukari.com and internshala.com [8].

Information extractions (IEs)

The recommender system requires efficient IE technology. As the amount of data collected by these systems is so voluminous that large-scale data processing approach is required for IE and pre-processing. Today, various job finder websites are facing a similar problem of IE due to the high influx of résumés and cover letters. The HR and hiring team perform manual screening to shortlist the candidates which takes good amount of working hours as well as workforce. IBM and HP came up with individual system to solve the problem of IE to help companies to select a good candidate and reduce the efforts of HR and hiring staff, but they failed to provide appropriate job recommendation to the job seekers. HP used layered data extracting technique to process résumés, whereas on the other hand, IBM uses conditional random field technique to extract meaningful data from the résumés [9,10].

Natural language processing (NLP)

This field is related with machine learning approach which is concerned with the interaction between computers and human beings.

NLP started back in 1950s by the publication of Alan turning's article "computing machinery and intelligence." Modern NLP is based on statistical machine learning which is helpful in performing various tasks as follows [11]:

- IE: IE along with named entity recognition, relationship extraction, coreference resolution, etc.,
- Information retrieval: This is related with data storing, data searching, and data retrieval.
- Word segmentation: This is a process of separating a large chunk of word into smaller words.
- Parts of speech tagging: This is a process to determine the part of speech of a given word.

CAREER RECOMMENDER OUTLINE

Overall application outline

The career recommender system crawls from job postings on sites such as naukri.com, etc., through a jobs crawler implemented in Python using BeautifulSoup. The jobs are filtered and stored in a comma-separated value (CSV) file. A graduate can create his account and upload his details directly by uploading his/her résumé. Automatically, the system mines the useful information from the student's résumé and by text mining techniques and compares this with requirements part in the job postings (Fig. 1). Résumé model is also used to compare domain space corpora of the different post-grad streams and calculate a similarity measure. Then, the system suggests the stream with highest, a similarity measure along with details of colleges (Fig. 2).

This system can also be used by employers to search candidates according to their requirements, and the results will be the matching candidates which have uploaded their résumé in our website (Fig. 3).

Recommendation system architecture

Information processing components

The web crawler downloads the information from websites such as Naukri.com, Monstor.com and Shiksha.com. It has following components:

1. Web crawling application: The web crawler uses a Python BeautifulSoup implementation to download XML data of the websites and extracts only useful information from that and converts it into text format.
2. Search job: The job postings extracted by web crawler are filtered to only useful information to be stored in a CSV file four columns: Company name, location, job title, and job requirements.

Algorithmic steps:

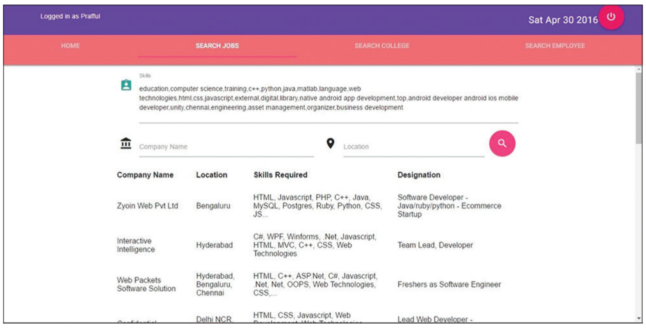
1) Link= http://www.naukri.com/jobs-in-India.
This link is for extracting jobs posted in Naukri.com

2) Four columns: Company name, location, job title, and job requirements were only extracted from the HTML page using following code:

```
comp_list=soup.find_all("span",{"class":"org"})
loc_list=soup.find_all("span",{"class":"loc"})
skill_list=soup.find_all("span",{"class":"skill"})
des_list=soup.find_all("span",{"class":"desig"})
```

3) This data are saved in a CSV file using the following code:

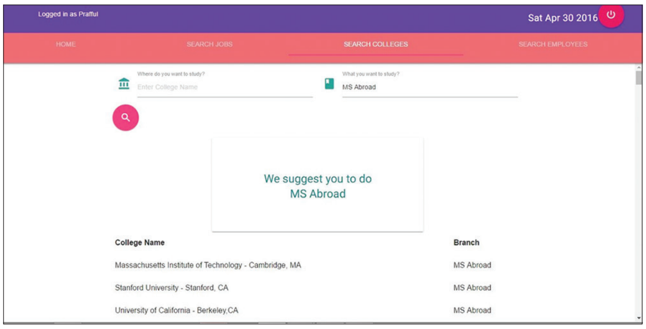
```
rb=xlrd.open_workbook("jobs.csv")
r_sheet=rb.sheet_by_index(0)
j=r_sheet.nrows
wb=copy(rb)
sheet1=wb.get_sheet(0)
for i in range(len(comp_list)):

        sheet1.write(j, 0, comp_list[i].text)
        sheet1.write(j, 1, des_list[i].text)
        sheet1.write(j, 2, loc_list[i].text)
        sheet1.write(j, 3, skill_list[i].text)
        j=j+1
        wb.save("jobs.csv")
```



education,computer science,training,c++,python,java,matlab,language,web technologies,html,css,javascript,external,digital,library,native android app development,top,android developer android ios mobile developer,unity,chennai,engineering,asset management,organizer,business development

Company Name	Location	Skills Required	Designation
Zyoin Web Pvt Ltd	Bengaluru	HTML, Javascript, PHP, C++, Java, MySQL, Postgres, Ruby, Python, CSS, JS...	Software Developer - Java/ruby/python - Ecommerce Startup
Interactive Intelligence	Hyderabad	C#, WPF, Winforms, .Net, Javascript, HTML, MVC, C++, CSS, Web Technologies	Team Lead, Developer
Web Packets Software Solution	Hyderabad, Bengaluru, Chennai	HTML, C++, ASP.Net, C#, Javascript, .Net, Net, OOPS, Web Technologies, CSS,...	Freshers as Software Engineer
	Delhi NCR,	HTML, CSS, Javascript, Web	Lead Web Developer -

Fig. 1: Screenshot of job search interface



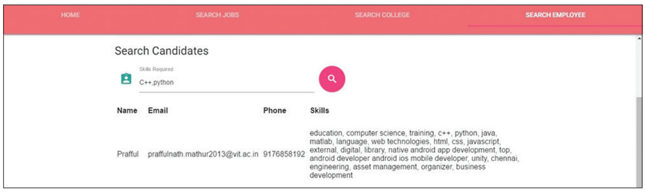
Where do you want to study? Enter College Name

What you want to study? MS Abroad

We suggest you to do
MS Abroad

College Name	Branch
Massachusetts Institute of Technology - Cambridge, MA	MS Abroad
Stanford University - Stanford, CA	MS Abroad
University of California - Berkeley,CA	MS Abroad

Fig. 2: Screenshot of post-grad search interface



Search Candidates

Skills Required: C++,python

Name	Email	Phone	Skills
Prafful	praffulnath.mathur2013@vit.ac.in	9176858192	education, computer science, training, c++, python, java, matlab, language, web technologies, html, css, javascript, external, digital, library, native android app development, top, android developer android ios mobile developer, unity, chennai, engineering, asset management, organizer, business development

Fig. 3: Screenshot of candidate search interface

3. Search college: The college lists are stored as CSV files containing 2 columns: College name and stream offered sorted in order of their rankings.

Algorithmic steps:

1) link= http://www.shiksha.com
This link is for extracting college lists posted in Shiksha.com

2) Two columns: College name and stream offered were only extracted from the HTML page using following code:

```
college_list=soup.find_all("a",{"class":"ranking-inst-title"})
branch_list=soup.find_all("span",{"class":"college-clr"})
```

3) This data is saved in a CSV file.

User interface

1. Résumé upload module: This is the by-default home page and the student can upload his/her résumé in one click. This interface accepts résumés in portable document format (PDF). It also has options to add some extra details such as hobbies and strengths.
2. Job search: The features extracted from the users are automatically submitted as a query for job search. The query can be updated by the user and the jobs are displayed sorted by similarity score.
3. College search: The features extracted from the users are

automatically submitted as a query for college search. This query cannot be updated by the user, but he/she can search for colleges for other streams. Colleges are displayed according to a ranking given various education websites such as Shiksha.com.

Query processor

This component runs in the background and is implemented using Python to mine skills from the résumé and match them with the job requirements and various streams corpora. This is explained more in Section 5. The overall architecture is illustrated in Fig. 4.

TEXT MINING

Pre-processing

The pre-processing converts the résumés which are available in PDF format into a format which can be further used for NLP. To do this, we first segment the document into sentences using sentences tokenizer in NLTK. Now, these sentences are further tokenized into words and English language stop words are removed along with meaningless words and symbols.

NLP labeling

The array of tokens is given parts of speech tags. A chunking operation selects only desired words using a well-trained regular expression [13].

SKILLS MATCHING FROM RÉSUMÉS FOR JOBS AND COLLEGES

Skills corpus

This corpus contains all the skill keywords in root word format collected by parsing millions of job postings on web. This can be used match with keywords from résumé to identify technical skill and extracurricular skills of a person.

Streams corpora

Every post-grad stream requires some specific set of skills in individuals, for example, communication skills and leadership qualities for MBA. The streams corpora contain specific skill sets for MBA, Master of Science, Doctor of Medicine, Master of Surgery, etc.

Skills matching

Keywords from the résumés undergo partial matching with corpus words, and match score is set by averaging the match sores of skills using a train set on a list of 1000 skills. The partial match score can be updated continuously by adding more skills to until the model over-fits the data (Fig. 5).

EFFICIENCY OF OUR APPROACH

We collected 150 résumés from graduates with at least 10 résumés from each stream, namely, B.Tech, BE, MBBS, CA, B.Com, BBA, and BA.

Recall of jobs extracted

$$\text{Recall} = \frac{\text{Releveant items extracted}}{\text{Total relevant items}}$$

We extracted the jobs by taking input as resumes of students belonging to each degree of graduation and calculated the recall of jobs on the 20 output job predictions using the above formula. The results are illustrated in Fig. 6. The average overall recall is 95% [14].

Precision versus recall of jobs extracted

$$\text{Precision} = \frac{\text{Releveant items extracted}}{\text{Total items extracted}}$$

We extracted the job prediction by taking input 150 resumes of students one by one calculated the recall and precision of jobs on the interval of 10 using the 20 output job predictions per résumé using the above formulae. The results are illustrated in Fig. 7 [14].

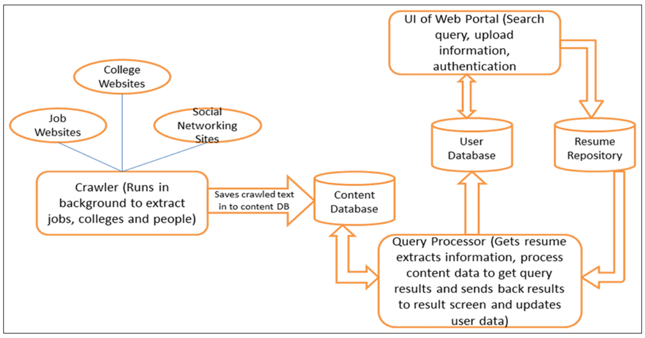

UI of Web Portal (Search query, upload information, authentication
College Websites
Job Websites
Social Networking Sites
User Database
Resume Repository
Crawler (Runs in background to extract jobs, colleges and people)
Saves crawled text in to content DB
Content Database
Query Processor (Gets resume extracts information, process content data to get query results and sends back results to result screen and updates user data)

Fig. 4: System architecture

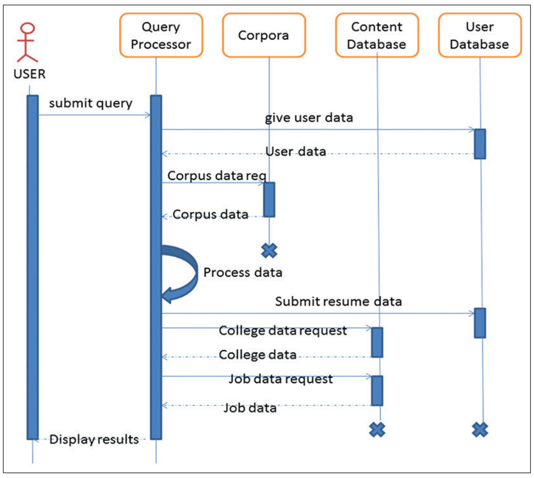


Fig. 5: Sequence diagram of information extraction

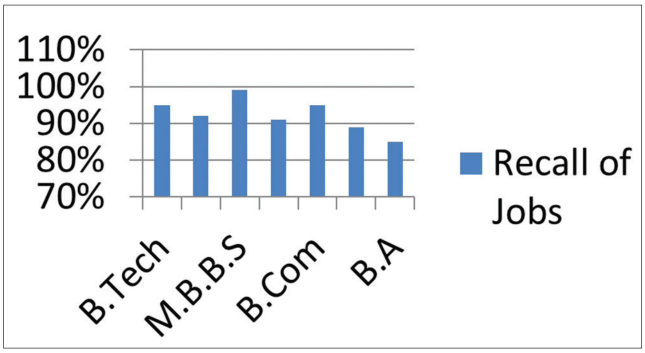


Fig. 6: Bar graph showing recall value of job search against graduate degree

Accuracy of prediction of MBA versus M.Tech/MS for B.Tech students

We extracted extracurricular skills of a student from his/her résumé. We matched these with our corpus of interpersonal skills required for MBA in MBA-skills corpus. As the number of résumé's increase, the MBA-skill corpus improves, hence the accuracy the prediction of MBA for a B.Tech student increases. Similarly, the accuracy of prediction of M.Tech/MS for B.Tech student's increases with the improving skills corpus with increasing number of résumé tested. The results are illustrated in Fig. 8.

Overall accuracy of college prediction

Accuracy was measured using a training set of 150 résumés. We compared student's original choice of stream by our prediction. In 77.5% cases, our prediction was correct. This may not be a high accuracy rate as original choice may be influenced by parental or peer pressure and not by skills.

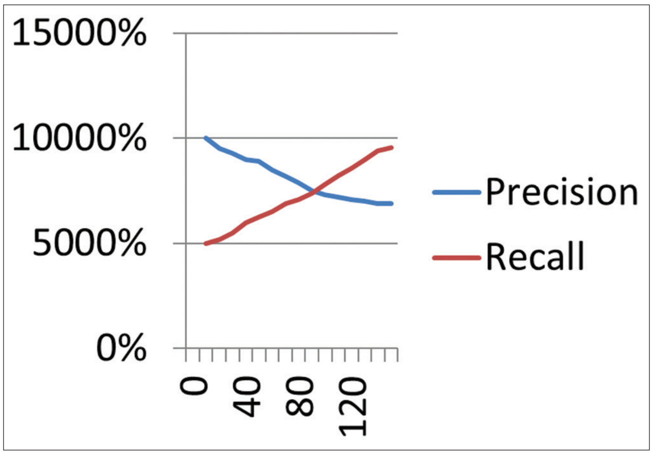


Fig. 7: Line graph showing decreasing precision and increasing recall of jobs prediction against number of résumés in the database

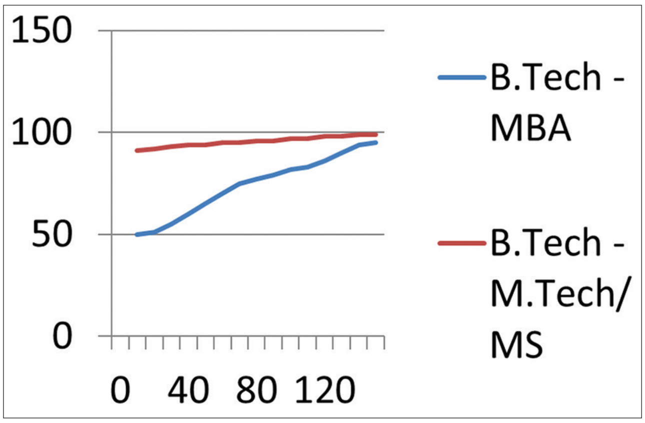


Fig. 8: Accuracy of prediction of post-graduation as MBA versus M.Tech/MS for B.Tech students

CONCLUSION AND FUTURE WORK

In this paper, we have proposed an approach to successfully implement a career recommendation system which suggests jobs or colleges for post-graduation using his/her résumé on one platform. This approach significantly improves the mechanism to choose a stream for higher education by taking into consideration their previous knowledge in the form of technical skills developed during graduation and their overall personality by considering their extracurricular skills, hobbies, and strengths. This system can be improved by refining the chunking regular expression which extracts the skill words from the résumé and features used for evaluating match score with streams corpora. The streams corpora can also be further refined.

REFERENCES

1. Biswas S, Srikanth G. Technology Vision - 2020: IT in Services. New Delhi: TIFAC; 2000. p. 61-71.
2. Taylor NF. Hiring in the Digital Age: What's Next for Recruiting? Available from: http://www.businessnewsdaily.com/6975-future-of-recruiting.html. [Last accessed on 2016 Jan 11].
3. Gupta P, Goel A, Lin J, Sharma A, Wang D, Zadeh R. Wtf: The who to follow service at twitter. In: Proceedings of the 22nd International Conference on World Wide Web. ACM; May, 2013. p. 505-14.
4. Chen HH, Ororbia II, Alexander G, Giles CL. Expert Seer: A keyphrase based expert recommender for digital libraries. 2015
5. Chen H, Gou L, Zhang X, Giles C. Collabseer: A search engine for collaboration discovery. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL); 2011.
6. Melville P, Sindhwani V. Recommender systems. In: Encyclopedia of Machine Learning. US: Springer; 2011. p. 829-38.
7. Ricci F, Rokach L, Shapira B. Introduction to Recommender Systems Handbook. US: Springer; 2011. p. 1-35.
8. Collaborative Filtering. March; 2005. Available from: https://www.en.wikipedia.org/wiki/Collaborative_filtering.
9. Case Study Resume Search: Supplying IT Talent in Contract Engagement. November; 2014. Available from: https://www.hpe.com/h20195/v2/GetPDF.aspx/4AA5-5979ENW.pdf.
10. Awasthi P, Gagrani A, Ravindran B. Image modeling using tree structured conditional random fields. In: IJCAI; 2007. p. 2060-5.
11. Natural Language Processing. May; 2004. Available from: https://www.en.wikipedia.org/wiki/Natural_language_processing.
12. Bird S, Klein E, Loper E. Processing Raw Text; 2015. Available from: http://www.nltk.org/book/ch03.html.
13. Bird S, Klein E, Loper E. Categorising and Tagging Words; 2015. Available from: http://www.nltk.org/book/ch05.html.
14. Kowalski GJ, Maybury MT. Information Storage and Retrieval Systems: Theory and Implementation. Vol. 8. Norwell: Springer, Science & Business Media; 2006