UNIVERSITY RECOMMENDATION SYSTEM FOR GRADUATE STUDIES WITH STATEMENT OF PURPOSE ANALYZER

Ujjwal Kulkarni

Student

Department of Software engineering,

SRM Institute of Science and Technology, Kattankulathur, Chennai, 603203, Tamilnadu, India

Karan N. Davey

Student

Department of Software engineering,

SRM Institute of Science and Technology, Kattankulathur, Chennai, 603203, Tamilnadu, India

Dr. M. Ferni Ukrit

Faculty and Guide

Department of Software engineering,

SRM Institute of Science and Technology, Kattankulathur, Chennai, 603203, Tamilnadu, India

Abstract

Shortlisting of universities according to their application profile, is a major problem faced by students pursuing their master's degrees. The master's application packet includes quantifiable scores like the GRE score, English proficiency test score, and the College grades. Apart from these scores, another important document is the Statement of Purpose. A Statement of Purpose can be the difference between an admit and a reject. There are many university recommendation applications but none of them takes the Statement of Purpose into account. This project aims to help undergraduate students overcome this ambiguity and provide a list of colleges that are safe to apply to. The Statement of Purpose is analyzed using Natural Language Processing. Gradient Boost regression technique is used in order to train and test the data set. Along with it, various functions of NLTK are used to extract features from the essay. The recommender provides a list of 8 colleges which the user can most likely get admission from. This project will prove to be extremely useful for the student community as it provides proper guidance which is not available for free. The recommendation system is a small step in the right direction to ensure that students apply for the right universities.

Keywords: *University Recommendation, Natural Language Processing, Statement of Purpose, Gradient Boost Regression*

I. INTRODUCTION

When applying for post graduation courses abroad, students often face various issues when it comes to shortlisting colleges. It is difficult to ascertain for a n applicant whether a university would be ambitious or can be considered safe without proper guidance. The entire process of shortlisting colleges and paying the application fees is tedious and can be rather expensive. When a college reviews an application, they consider several factors and each has a different level of importance. A recommendation system to aid with this process can be beneficial from a financial point of view and also make the entire process simpler.

The main idea behind the University Recommendation System (URS) is to help students who aim to pursue their master's degree. The user will be asked to enter details that are commonly required during the admission process. The program will evaluate the students' profiles and help them choose the correct program. It mainly aims to help junior and senior year students who will be applying for their Masters or Ph.D. program in the future. The system is the first of its kind and unique in every aspect. It recommends undergraduate students colleges where they can most likely get an acceptance letter from. The recommender system judges the profile based on quantifiable parameters such as exam scores and Grade Point as well as by rating the Statement of Purpose, which makes this system distinctive. Although several systems online help with the shortlisting process, almost none of them consider the Statement of purpose (SOP) as an important parameter. The SOP helps to distinguish promising candidates from the average ones as it cannot be plagiarised and is original. Thus, the list of colleges is prepared not only using the scores and also by evaluating the essay.

II LITERATURE SURVEY

In [1], Mahamudul Hasan , Shibbir Ahmed , Deen Md. Abdullah , and Md. Shamimur Rahman, the researchers have tried designing and developing a recommender system for graduate admission seekers which can help them to choose graduate school matching their entire academic profile. K-nearest Neighbor algorithm for calculating top N similar users for the test users and recommend Top K universities to users from N similar users. The dataset consists only of Bangladeshi students who study abroad making the system quite specific and unreliable in many cases. This highlights the need for a more robust and reliable system which is exactly what Alisha Baskota and Yim-Kai Ng discuss in [2]. Their source of data comes from portals, such as Edulix.com and Yocket.com, which archive data of universities and these websites. The system selects appropriate features from the data, runs the Support Vector Machine (SVM) classiſcation algorithm and K-Nearest Neighbors (KNN) machine learning algorithm on them, and suggests appropriate universities to the applicants accordingly. SVM is a good algorithm for classifying non linear data and provides high accuracy. The factual data of schools includes the acceptance rate of graduate schools, CGPA, GRE scores, location, number of students admitted to the schools, private/public universities, ranking of the schools,safety (likeliness of admission). This makes it one of the most complete recommendation systems proposed. However, although several parameters like the GRE scores and other exams have been considered, the SOP is not taken into consideration which is vital to secure admission into top universities.

[3] discusses a system used to present a new college admission portal using data mining techniques for tackling college admission prediction problems. This System uses content-based filtering to provide aspiring students (Master of Science) with the most appropriate choices of colleges based on different parameters. The system analyses the student academics, merit, background, student records and the college admission criteria. The system to be developed is to recommend the universities to students applying GRE which moreover includes all aspects like, GRE score, TOEFL score, wish list analysis, SOP analysis, etc. This system only considered two or three parameters to analyse the SOP which is not a comprehensive approach.

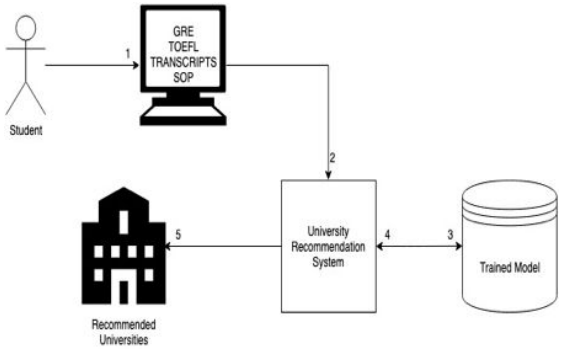
Essay analysis is an important part of evaluating the SOP. [4] proposes a regression based approach for automatically scoring essays that are written in English. Standard NLP techniques are used to extract various features from the essays and based on their importance scoring is done. This paper also takes a look at the pre-existing essay raters that are available on the internet such as the ETS (Electronic testing service) and Electronic Essay rater (E-Rater). The rater aims to consider various parameters which human graders might consider to make the scoring similar. The Cohen's kappa score which basically finds the percentage of agreement between the two scores turns out to be 0.72.

III PROPOSED WORK

A. Proposed Methodology

The implementation of the system is done in a python console with functions for extracting the various essay features using NLTK library. Upon creating their account, the student is asked to enter their GRE and TOEFL scores, CGPA, number of publications if any. Another document which is also considered is the Statement of Purpose. Using the pandas library, this document is converted to a much more usable .csv format. The essay is rated based on several factors. Features such as number of adverbs, grammatical errors, number of verbs and many more are extracted from the essay using nltk library functions. A set of essays is used for training and model creation by using gradient boosting regression in order to obtain the best results and a score is given for the essay. Each parameter holds importance and needs to be carefully considered while determining the chance of admittance. After reading the scores, based on the weight given for each parameter, scores are assigned and added. We also factor in the score given to the statement of purpose and a final score is obtained. On the basis of this final score, a list of colleges is displayed to the candidate. This list consists of colleges which are safe to apply to based on the student's profile. It takes into account the previous admits into the universities and the scoring is decided according to that.

B. Proposed Architecture



C. Abbreviations and acronyms

SOP- Statement of Purpose
URS- University recommendation system
ETS- Electronic testing service
E-Rater - Electronic essay rater
SVM - Support Vector Machine
KNN- K nearest neighbour

IV. IMPLEMENTATION

The proposed system is implemented using the NLTK framework in Anaconda IDE.

A. Pre-Processing

It is a vital part of the training process. The statement of purpose to be scored must be in appropriate format and in english language in order to train the model. The stop words are first removed from the essay and then tokenized in order to extract features and evaluate them.

There are 11 features using which a score is graded, which are as follows:

1) Bag of Words (BOW) counts (10000 words with maximum frequency)
2) Number of characters in an essay
3) Number of words in an essay
4) Number of sentences in an essay
5) Average word length of an essay
6) Number of lemmas in an essay
7) Number of spelling errors in an essay
8) Number of nouns in an essay
9) Number of adjectives in an essay
10) Number of verbs in an essay
11) Number of adverbs in an essay

The dataset is divided into two parts. 70% of the data is training dataset used to train the model. The remaining 30% is used to test the dataset in order to check the accuracy of the model.

B. GBM Essay Rater

Gradient Boosting Regression is used to rate the essays. Gradient Boosting trains many models in a gradual, additive and sequential manner.(*y=ax+b+e , e needs a special mention as it is the error term)*. Gradient boosting is better than AdaBoost as it identifies the shortcomings of the weak learners. The objective is to minimize the loss of the model by adding weak learners using a gradient descent like procedure. This class of algorithms was described as a stage-wise additive model.

Algorithm 1: Gradient Boost

- Initialize $F_0(x) = \arg\min_\rho \sum_{i-1}^{N} L(y_i, \rho)$
- For $m = 1$ to M do:
  - Step 1. Compute the negative gradient
    $$\tilde{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F_{x_i}}\right]$$
  - Step 2. Fit a model
    $$\alpha_m = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} [\tilde{y} - \beta h(x_i; \alpha_m)]^2$$
  - Step 3. Choose a gradient descent step size as
    $$\rho_m = \arg\min_{\rho} \sum_{i-1}^{N} L(y_i, Fm - 1(x_i) + \rho h(x_i; \alpha))$$
  - Step 4. Update the estimation of $F(x)$
    $$F_m(x) = F_{m-1}(x) + \rho_m h(x; \alpha_m)$$
- end for
- Output the final regression function $F_m(x)$

Fig. 1. Gradient boosting algorithm

C. Algorithm

1. The student logs into the system and enters the details.
2. The Statement of Purpose is uploaded in .csv format.
3. Essay Rater scores the Statement of Purpose based on the trained model.
4. The score is appended to the tuple of the student.
5. Every feature in the tuple is assigned a weight according to the score.
6. The weights are summed up in order to find the category of universities suitable for the profile.
7. The list of universities are displayed to the user.

V. RESULTS

The essay rater model fared well and had a mean squared error of 0.62. An accurate essay rater would enable the recommender system an appropriate list to the user. After analyzing the performance using Linear, Lasso, Ridge, and Gradient Boosting Regression, it was concluded that GBR had the most accurate score. Hence, the essay rater model applies a light GBM to score the essays based on 11 different features. The essay score appended to the other quantifiable features. All the features are scored on a scale based on their importance in the application packet. The final score obtained is then used to provide a list of universities to the user. The criteria on which the admissions are rolled out is ambiguous as the admission committee does not explicitly mention it. But, based on the admission trend, we were able to infer the trend of admissions for different universities. The challenging part of this recommendation system is scoring the Statement of Purpose as it is the only unquantifiable feature of the application packet. As the Statement of Purpose is a very confidential document, the dataset available is limited. The essay rater scores the essay based on the features which we found we most important in an essay.

VI. CONCLUSIONS

In summary, the University Recommender scores the Statement of Purpose and appends the score with the other quantifiable features. These features are scored and assigned a category based on which a list of appropriate universities is provided to the user. The accuracy of the essay rater can be increased by taking in more features, such as punctuation counts and essay sentiment. The University recommendation is a small step towards solving the ambiguous problem of college shortlisting.

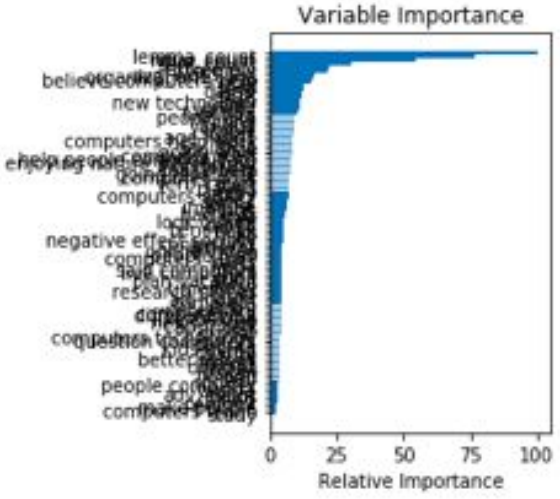


Fig. Importance of variables involved in scoring an essay

REFERENCES

1. Mahamudul Hasan , Shibbir Ahmed , Deen Md. Abdullah , and Md. Shamimur Rahman, “Graduate School Recommender System: Assisting Admission Seekers to Apply for Graduate Studies in Appropriate Graduate Schools”, 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)
2. Alisha Baskota,Yiu-Kai Ng, “A Graduate School Recommendation System Using the Multi-Class Support Vector Machine and KNN Approaches, 2018 IEEE International Conference on Information Reuse and Integration for Data Science.
3. Kapase Vidya, Musale Pooja, Shinde Nikita, Paryani Kanchan, “University Recommendation Engine for MS’, IJIRST –International Journal for Innovative Research in Science & Technology| Volume 2 | Issue 11 | April 2016 ISSN (online): 2349-6010
4. Sristi Drolia, Shrey Rupani, Pooja Agarwal, Abheejeet Singh, “ Automated Essay Rater using Natural Language Processing”, International Journal of Computer Applications (0975 – 8887) Volume 163 – No 10, April 2017