Analysing Recommendation of Colleges for Students Using Data Mining Techniques

Varsha Powar
Department of Information Technology
Maharashtra Institute of Technology
Pune, India
varsha.powar@mitpune.edu.in

Sheetal Girase
Department of Information Technology
Maharashtra Institute of Technology
Pune, India
sheetal.girase@mitpune.edu.in

Debajyoti Mukhopadhyay
Department of Information Technology
Maharashtra Institute of Technology
Pune, India
debajyoti.mukhopadhyay@mitpune.edu.in

Anuja Jadhav
Department of Information Technology
Maharashtra Institute of Technology
Pune, India
anuja.jadhav9955@gmail.com

Shweta Khude
Department of Information Technology
Maharashtra Institute of Technology
Pune, India
shwetakhude1@gmail.com

Shital Mandlik
Department of Information Technology
Maharashtra Institute of Technology
Pune, India
shitalmandlik327@gmail.com

***Abstract*—We proposed a system to provide a recommendation system which will generate the user interested colleges list. This will be done by asking few questions related to the college like- college infrastructure, campus life, placement, sports and cultural activities. Information seeker will give ratings to the questions according to his interest rate. The list of college will be stored in the dataset for calculating the accuracy and precision of the system. Also the students will be giving feedback to the system for their performance.**

***Keywords*—Recommendation System, Accuracy and Precision.**

I. Introduction

Education is the right of every citizen in India. People are also encouraged to enhance their skills like sports, singing, dancing, etc. In Engineering colleges, various activities are arranged for enhancing their technical as well as non technical skills. We can get reviews of the colleges from social networking sites but they might not be valid. The real reviews should be collected from the students who actually studied in that particular college. This will be helpful for giving real time recommendations to information seeker.

In rural area people are unaware of the colleges and their facilities. There are some systems which are having colleges information including their facilities also there is no such system which will provide recommendations of Engineering Colleges at one place based on the information seekers/ students interest. There are some attributes which a student usually want in an Engineering college like- Placement, Extra curricular, Sports, Cultural activities, Library also Discipline and Security etc. After generation of recommendation analysis should be done to strengthen the system or to increase the effectiveness of the system. Thus we have proposed a system which will generate recommendation by considering feature-set of interest based on real time reviews of alumni. Also the system should analyze the effectiveness of the generated recommendations by taking feedback from students or information seekers and taking feedback about the performance of the system using NLP(Natural Language Processing).
The primary objective of the topic is to-

- Organizing the task of user profiling.
- Extracting profile of each college.
- Doing user interest discovery from the available information.
- Collecting reviews from alumni of respective colleges.
- Creating a system which will provide recommendation through simple interface.
- Analyzing the recommendation generated for students using precision and recall.
- Analyzing the performance of the system using sentiment analysis.

II. Literature Survey

Recommendation Systems are the type of information filtering systems designed to help users to find their way through todays large information spaces.Till date Recommendation Systems are the best examples of personalization system. Recommendation Systems uses a number of different technologies. The goal of a Recommendation System is to generate meaningful recommendations to a collection of users for items or products that might interest them. Suggestions for books on Amazon, or movies on Netflix, are real world examples of the operation of industry-strength recommendation systems. The design of such Recommendation engines depends on the domain and the particular characteristics of the data available. For example, movie watchers on Netflix frequently provide ratings on a scale of 1 (Disliked) to 5 (liked). Such a data source records the quality of interactions between users and items. Additionally, the system may have access to user-specific and item-specific profile attributes such

978-1-5386-3852-1/17/$31.00 ©2017 IEEE

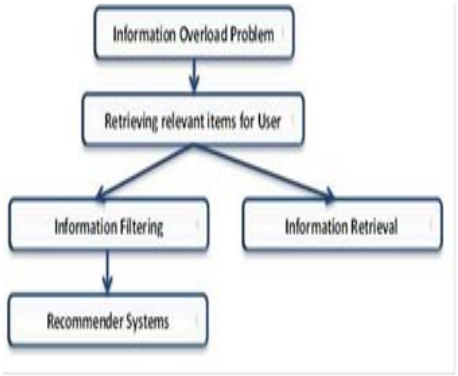

Fig. 1: Recommendation System

as Demographics and product descriptions respectively. Recommendation systems differ in the way they analyze these data sources to develop notions of affinity between users and items which can be used to identify well-matched pairs. Collaborative Filtering systems analyze historical interactions alone, while Content-based Filtering systems are based on profile attributes; and Hybrid techniques attempt to combine both of these designs. The architecture of recommendation systems and their evaluation on real-world problems is an active area of research. There is an extensive class of Web applications that involve predicting user Responses to options. Such a facility is called a recommendation system.

However, to bring the problem into focus, few good examples of Recommendation systems are:

- Offering news articles to online newspaper readers, based on a prediction of reader interests.
- Offering customers some online retailer suggestions about what they might like to buy, based on their past history of purchases and/or product searches.
- Offering research articles to online researchers based on their current profile and search behavior

We can classify these systems into two broad groups. Sentiment analysis (sometimes known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. Information comes in many shapes and sizes. Information extraction is used to automatically extracting structured information from unstructured or semi-structured data.

III. Proposed System

In our proposed system student will be able to get the most suitable college of their interest just by giving ratings out of five to the questions

1. College infrastructure
2. Industry Exposure
3. College campus life
4. College placement activity
5. College teaching Staff and faculty
6. Hostel facility
7. Sports activity
8. Extracurricular activity
9. location of college
10. Cultural Activities
11. College library
12. Security and discipline

Fig. 2: Architecture of Proposed System

The reviews of the alumni of respective colleges are stored in the dataset with respect to the above 12 attributes. The recommendation will be generated based on the alumni reviews. After a year or 6 months to admission they will give the feedback to the system and the system will analyze effectiveness by processing their feedback.
The Users feedback to the system will be analyzed using natural language processing. The feedback of user will be split into the tokens and feedback will be implicitly calculated by the Stanford library.The system will have two types of users one is student or information seeker and other is administrator.

1. Recommendation System flow

   - For getting access to the system services user should register first to the system. After registration, they can access the services offered by the system.
   - Student/ information seeker plays a vital role to the system for generating their most likely to have colleges list.
   - So firstly the attributes mentioned earlier will be asked to the user. User are supposed to give ratings to all twelve attributes in 0 to 5.
   - After taking the ratings from student/ information seeker, the system will give this as a information seeker feature set to our recommendation algorithm.

- Our recommendation algorithm will find the similarity between alumni reviews stored in the dataset and the information seeker feature set. If similarity found in a certain threshold value then the system will add the college code into the cluster of positive class and other college code into negative class.
- The positive class will again processed to remove redundancy of the college code. And the list will be displayed.
- Students/ Information seekers are supposed to give feedback to the system and the system should analyze the feedback of the user and make sentiment analysis of the user comment about system using NLP (Natural Language Processing).
- Student/ Information seeker will provide the current college code in the feedback form which will be given as an input to our analysis algorithm. Our analysis algorithm will process the input and cluster of positive class in two subsection:precision and accuracy [5].
- This will generate a graph at analysis system which will show the performance of the recommendation system.
- We will also take feature set against the college code which they are currently pursued/ pursuing for generating real time recommendation to the new user.

2. Analysis System flow

- In this system admin is responsible for uploading alumni reviews dataset.
- Also admin is responsible for extraction of college information from social site.
- Admin will perform the task of controlling the information in the system and will view user's information and also sentiment generated to the user.
- admin will perform aggregation of the college rating given by alumni to calculate overall rating of each college.

IV. Methodology

1. **College profile extraction:**
Information extraction of various colleges is done with using python script. DTE (Directorate of Technical Education) site is having information about the colleges in the form of unstructured or semi-structured data. The system collect this unstructured or semi-structured data into the database. MongoDB provides an ease of storing this unstructured or semi-structured data efficiently. After storing this data system will simply retrieve the data present in the database.

2. **Collecting reviews from alumnis:**
We have collected the reviews of the alumni based on the 12 attributes mentioned in previous chapter.

{"_id":"ObjectId(56f618796b1f29db504876c9)","Code":3208,"Name":"Don Bosco Institute of Technology, Mumbai","Address":"Premier Automobiles Road,Opp. Fiat Company, Kurla (West), Mumbai
","Website":"www.dbit.in","Email":"dbit@dbit.in","Year":2001,"Nearest_Railway_Station":"Vidyavihar (Central Railway)","Railway_Distance":1,"Nearest_Bus_Station":"Vidyavihar (BEST)","Bus_Distance":1,"Nearest_Airport":"Santacruz","Airport_Diatance":"","Hostel_Boys":"No","Hostel_Boys_Intake":0,"Hostel_Girls":"No","Hostel_Girls_Intake":0,"Status1":"Un-Aided","Status2":"Non-Autonomous","Status3":"Religious Minority - Roman Catholics","Girls hostel":0,"cafeteria":1,"Auditorium":1,"Gym":0,"Boys Hostel":0,"Total Faculty":47,"Faculty Student Ratio":"23:01","CET":1,"AIEEE":0,"Mechanical":1,"IT":1,"Computer":1,"Electronics":0,"UG Intake":150,"PG Intake":30,"Location":"Kurla (West), Mumbai "}
{"_id":"ObjectId(56f618796b1f29db504876ca)","Code":3135,"Name":"Manjara Charitable Trust's Rajiv Gandhi Institute of Technology, Mumbai","Address":"Off. Juhu Versova Link Road, Versova, Andheri(W) Mumbai
","Website":"www.mctrgit.ac.in","Email":"mctrgit02@gmail.com","Year":1992,"Nearest_Railway_Station":"Andheri","Railway_Distance":3,"Nearest_Bus_Station":"Juhu","Bus_Distance":1,"Nearest_Airport":"Santacruz","Airport_Diatance":"","Hostel_Boys":"No","Hostel_Boys_Intake":0,"Hostel_Girls":"No","Hostel_Girls_Intake":0,"Status1":"Un-Aided","Status2":"Non-Autonomous","Status3":"Non-Minority","Girls hostel":0,"cafeteria":1,"Auditorium":1,"Gym":1,"Boys Hostel":0,"Total Faculty":92,"Faculty Student Ratio":"22:01","CET":1,"AIEEE":1,"Mechanical":1,"IT":1,"Computer":0,"Electronics":1,"UG Intake":60,"PG Intake":15,"Location":"Andheri,Mumbai"}

Fig. 3: College Database

And the reviews are stored in the dataset. As given

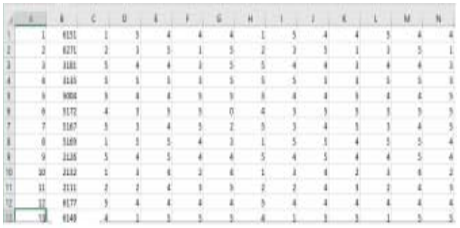

Fig. 4: Alumni Reviews

in the figure the first field is the user id, the second shows the college code of the alumni and the next 12 fields shows the rating given by the alumni on those 12 attribute.

3. **Collecting Feature set of student/information seeker:**
Student/Information seeker will give their feature set by rating to the 12 attributes. Recommendation system will take this as a feature set and give it to our recommendation algorithm.

4. **Generating Recommendations:**
Now the Recommendation system will take the feature set and process with the dataset of alumni review for generating list of college.

Recommendation will be generated by following steps:

1. Set the threshold value to 6.
2. Pointer should points to the first column in the alumni dataset and the other pointer should point to the first question rating given by the student/information seeker.
3. Set count=0;
4. Check the question rating with the alumni first question rating.
5. If match found increase count
6. Repeat step 4 and 5 12 times
7. If threshold¡count Store the college code into temporary array. Given in the alumni dataset.
8. Repeat step 2 to 7 for all alumni reviews.
9. Now check whether there is any repetition exists. If yes then keep single copy of college code.
10. Set the threshold value to 6.

5. **Analysis of Recommendation System:**
System will store recommendations generated in previous step. After taking admission to engineering college student will give feedback to the system after a year. Admin will send request to the student through mail to give feedback to the system. Now system will ask student to enter his current college code and the comment about system in text. Here we will apply the NLP to the comment entered by the student. We will check the current college code given by student against the recommendation stored in the system. After processing them admin will be able to see the feedback in the form of- positive, negative, very positive, very negative and neutral. We will make ranking of the student out of 10 by simply calculating:

Rank=Rank of college code in the recommendations*10/10 If the college code is not known to the system it will store NULL as rank for particular record. Storing this rank into temporary file, Precision and accuracy will be calculated by the following formulas:

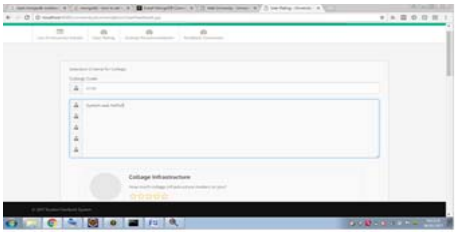

Fig. 5: Student Feedback

- **Accuracy**
Accuracy gives hundred percent effectiveness of the system through following formula-

TP/TP+FP+TN+FN ———(1)

where,

TP-True Positive
FP-False Positive
TN-True Negative
FN-False Negative

Steps to calculate Accuracy

1. accurate=0, `no_of_stud=0`;
2. Open temporary file and check the rank
3. If Rank==1 then accurate++;
4. `no_of_stud++;`
5. Repeat the 2 and 3 steps till the last record of the file
6. accuracy=accurate/`no_of_stud`*100;

- **Precision**
Precision gives the system performance estimation

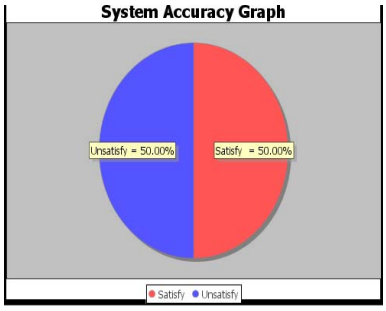

Fig. 6: System Accuracy Graph

through following formula-

TP-TN/TP+FP+TN+FN ———(2)

where,

TP-True Positive
TN-True Negative
FP-False Positive
FN-False Negative

Steps to calculate Precision

1. precision=0, `no_of_stud`=0, Fail=0;
2. Open temporary file and check the rank
3. If Rank==NULL then Fail++;
4. `no_of_stud`++;
5. Repeat the 2 and 3 steps till the last record of the file
6. precision=`no_of_stud`-Fail/`no_of_stud`*100;
7. precision=`no_of_stud`-Fail/`no_of_stud`*100;

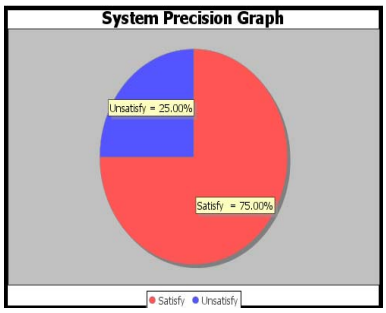

Fig. 7: System Precision Graph

Users will give the input in the textual format about the systems performance like- speed, user interface etc. Stanford library of NLP will implicitly analyze the feedback. And the algorithm will represent users feedback in the form of positive, negative, neutral, very positive, very negative as given in the figure. For example: the word good comes in positive feedback and bad comes in negative feedback but the word not bad comes in neutral feedback. These kind of

ambiguity will be handled by the Stanford library.

Sentiment analysis code:

```
String[] sentimentText = {
”Very Negative”,”Negative”,
”Neutral”, ”Positive”, ”Very Positive”
}
for(CoreMap sentence:annotation.get(Core-
Annotations.SentencesAnnotation.class))
{
Tree tree=sentence.get(SentimentCoreAnnotations.Annotated-
Tree.class);
int score=RNNCoreAnnotations.getPredictedClass(tree);
System.out.println(sentimentText[score]);
}
```

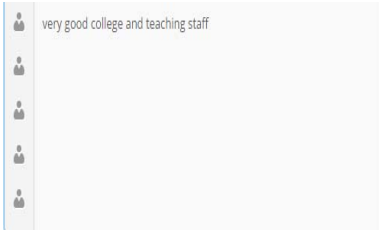

Fig. 8: Textual Feedback From User

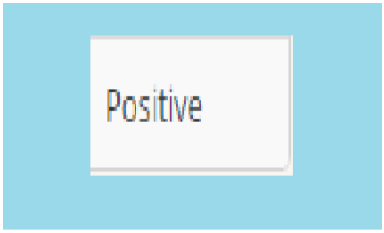

Fig. 9: Analysis of Feedback

V. Acknowledgment

This paper is based on a project grant 60,000 Proposal No : 14ENG002138 received from BCUD of the Savitribai Phule.

VI. Conclusion

There is very less precise and exact data available about Universities/Colleges as well as users. This has brought the need of a one stop portal where this information could be placed in a systematic manner and can be accessed by the users for better decision making. So we have used our Profile extraction model to extract data from various web sources. Also we have profiled users implicitly based on their social networking website. After getting this data we have converted this unstructured data into structured keyword based Profile. While there was no provision for recommending Universities to users, we have built a User Profiling System for Universities and Users. Once the Profile is generated, extracting Knowledge out of this Data is a big deal. We have managed to find out change in the ranks of the institutes with respect to user interest, which we have analyzed by calculating the ranks of each institute with changing criteria weights.

References

[1] Hasan Omar, Benjamin Habegger, Lionel Brunie, Nadia Bennani, and Ernesto Damiani. A Discussion of Privacy Challenges in User Proling with Big Data Techniques: The EEXCESS Use Case. In Big Data (BigData Congress), 2013 IEEE International Congress on, pp. 25-30. IEEE, Santa Clara Marriott, CA, USA, 2013.

[2] Machine Learning: Hands-On for Developers and Technical Professionals by Jason Bell. Published by John Wiley and Sons, Inc. 10475 Crosspoint Boulevard Indianapolis, IN 46256, www.wiley.com.

[3] Goldberg, D. Nichols, D. Oki, B. M. and Terry D. (1992). Using Collaborative Filtering to Weave an Information Tapestry. Communications of the ACM, 35(12), 6170.

[4] Sumitkumar Kanoje, Debajyoti Mukhopadhyay, Sheetal Girase; User Profiling University Recommender System using Automatic Information Retrieval, 1st International Conference on Information Security and Privacy, ICISP 2015 Proceedings; Nagpur, India; Elsevier Proceedia Computer Science, USA; December 11-12, 2015; pp.5-12;

[5] Sumitkumar Kanoje, Varsha Powar, Debajyoti Mukhopadhyay. Using MongoDB for Social Networking Website - Deciphering the Pros and Cons; IEEE sponsore International Conference on Innovation in Information Embedded and Communication Systems, ICIIECS 2015 Proceedings; Coimbatore, India; IEEE Computer Society Press, California, USA; March 19-20, 2015; ISBN 978-1-4799-6818-3.

[6] Rodrigo Miranda Feitosa, Sofiane Labidi, Andr Lus Silva dos Santos, Nilson Santos, Social Recommendation in Location-Based Social Network using Text Mining, 4th International Conference on Intelligent Systems, Modelling and Simulation, 2013