Implementing a Recommender System for Bangladeshi Faculty Search using Machine Learning

by

Md.Khalid Hasan
12101102

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
December 2019

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. I have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Khalid Hasan
12101102

Approval

The thesis titled “Implementing a Recommender System for Bangladeshi Faculty Search using Machine Learning” submitted by

1. Md.Khalid Hasan (12101102)

Of Fall, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on December 26, 2019.

Examining Committee:

Supervisor:
(Member)

Hossain Arif
Assistant professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Mahbubul Alam Majumdar, PhD
Professor and Chairman
Department of Computer Science and Engineering
Brac University

Abstract

Machine learning is a one of the popular fields in Computer Science. In my thesis research the focus is to implement a recommender system for Bangladeshi faculty search. Selecting a appropriate faculty or thesis supervisor is a very important part in a student's life. Even choosing right academy is also an important part in their study life. This research paper presents a faculty recommender system to assist students in making these choices. Here the main focus is to cover our own country, Bangladesh, to help the students of our country to pursue their own interest. I proposed this recommender system by using collaborative filtering algorithm. I used a very popular machine learning algorithm, K-Nearest Neighbor algorithm with cosine similarity to predict faculty members. It works on a vast database and being analyzed by different criteria. It applies multiple filtering conditions to retrieve relevant supervisor or faculty member based on the research interest or preferences. The preference field of the faculties based on preferred research area, making part of the decision specific. This system helps a user finding faculty or supervisor according to own individual interests. It contains information about faculties around Bangladesh from different institutions. A classification accuracy of 76.0 % for the predicted results ac hived by the proposed model.

Keywords: Faculty Search, Machine Learning, Collaborative Filtering algorithm, Prediction, Recommender System, K-Nearest Neighbor algorithm, cosine similarity.

Dedication

To my supporting faculty body and well wishers from the department and beyond. Love goes out to my friends and family for giving me the latent energy I always needed to get through this.

Acknowledgement

Firstly, all praise to the Great Allah for whom my thesis have been completed.
Throughout my long journey, I have received uncountable assistance and contribution from many well-wishers. The journey would have been incomplete without their constant support and priceless contribution.
I want to express my wholehearted gratitude to my thesis supervisor Hossain Arif, Assistant Professor and Dilruba Showkat, Senior Lecturer – both from Department of Computer Science and Engineering of BRAC University. Their persistent motivation, guidance and expertise fueled my research progress.
Additionally, I am grateful to those students of CSE department, BRAC University who have helped me to collect the data which I used for my research work.
Nonetheless, I would like to show appreciation to the Department of Computer Science and Engineering, Brac University for providing me with all the fundamental help.
And finally to my parents without their throughout support it may not be possible. With their kind support and prayer I am now on the verge of my graduation.

Table of Contents

Declaration **i**

Approval **ii**

Abstract **iii**

Dedication **iv**

Acknowledgment **v**

Table of Contents **vi**

List of Figures **viii**

List of Tables **ix**

Nomenclature **x**

1 Introduction **1**

1.1 Thoughts behind the Prediction Model 1

1.2 Aims 2

1.3 Necessity of the system 2

1.4 Thesis Orientation 2

2 Literature Review **4**

3 Methodology **7**

3.1 Data Collection 7

3.1.1 Data-set for Algorithm 8

3.2 Algorithm 9

3.2.1 Collaborative Filtering Algorithm 9

3.2.2 Content Based Approach 9

3.2.3 Cosine Similarity 10

3.2.4 KNN Algorithm 11

4 Implementations **12**

4.1 Cosine Similarity as distance metrics 14

4.2 KNN Algorithm Pseudocode 14

4.3 Language and libraries used 14

5 Result Analysis 16
5.1 Algorithm Evaluation Process 16
5.1.1 Confusion Matrix 16
5.1.2 Accuracy 17
5.1.3 Precision 17
5.1.4 Recall(Sensitivity) 17
5.1.5 F1 Score 17
5.2 Algorithm Performance Evaluation 18
5.3 Comparison 18

6 Conclusion and Future Plan 20
6.1 Conclusion 20
6.2 Future Plan 20

Bibliography 24

List of Figures

3.1 Data collection Sample[28] 7
3.2 Partial Data-set for Filtering algorithm(part-1) 8
3.3 Partial Data-set for User Based Filtering(part-2) 8
3.4 Cosine similarity between item A and item B [13] 10

4.1 Flowchart of the algorithm 13
4.2 Pseudocode for KNN Algorithm [16] 14

List of Tables

3.1 Collaborative Filtering data-set example 9

5.1 Confusion Matrix 17
5.2 Confusion Matrix for used algorithm 18
5.3 performance measurement results 18

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

CF	Collaborative Filtering
KNN	K Nearest Neighbors
ML	Machine Learning
TEL	Technology Enhanced Learning

Chapter 1

Introduction

1.1 Thoughts behind the Prediction Model

Selection of thesis supervisor, Masters Degree admission or selection of PhD supervisor is an significant chapter that a student must go through at an crucial stage in their student life. Advice of the supervisor or instructor is a important determiner of quality in a thesis work. This may make an important impact in the student's ultimate goal. When it is time to decide whether a faculty or professor is the right one to be his or her supervisor, the student should consider the person based on a set of fields that are important in the faculty or supervisor selection process. Things might go in different way while doing this part. Manderson indicates in his paper [18] that when selecting a supervisor, students should estimate their needs and the pros and cons of the supervisors in details. So that students could make themselves very clear about choosing their right faculty. In this regard, they should take help from such system that would guide them to pursue their goal by recommending them faculties according to their own interests.

At present times, recommender system is being used in many sectors. Though recommender system is not a new idea, it's development considering particular purpose is always a concern. In this modern era of information, it's become more difficult to get required information which someone is intended to get. In general, recommender systems are usually algorithms which aim at proposing relevant things to different users. Now a days, recommender systems are being involved in economic and profit section of different fields. They become more crucial for that reason. [23] Almost most of the online platforms are using their recommendation systems to offer their users new and delightful services. In this way, they are making themselves significant to their competitors. Recommender systems are also becoming very useful in different educational purpose.

Many renowned companies are using their own recommendation systems. Facebook, Amazon, YouTube are some of the big names. They make classification of their fascinating services and make recommendations for the similar group of users. Now the fact need to be in mind is how they are making these similar groups. These similarities can be found in many way. Finding similarity in documents based on major words in that document is one of the criteria. Activities assessment of the users can be another criterion. In terms of different purpose, similar approaches can be applied. Evaluating user ratings can be also a way for recommending such things. We can consider the recommendation system by Netflix as an example. They uses their user ratings to give their users movie recommendations that the other users might like to watch. [18]

The main goal of this research is to make a common platform for searching Bangladeshi faculties and make such recommendation system to help the students to get recommended faculty members according to their needs. This process will make the decision-making process much more easier by recommending the right person to the users. The system will help students in different slots in their proceedings. Lastly, I can say that, this whole process will be beneficial to the users in many fragments of their life.

1.2 Aims

The aim of this thesis is to prepare a model which will recommend faculty members for users. My main objective is to combine preference rating data and preference fields in order to implement a sufficient predictive model. Analyzing the different fields is also important so that the model becomes more efficient in better recommendation and increase it's performance.

1.3 Necessity of the system

For students, it is important to select thesis or project supervisor wisely. Even after graduation it is important to select master's degree institution or even PH.D. supervisor becomes very crucial for students. In our country it is more difficult for the students. So all these necessary information might be helpful if they are in a common place. We may get information in online or university websites, but choosing between them and comparing them is very important. Before starting the project, I made a survey among around 50 final year students about the necessity or importance of such project. Maximum of them responded positively. Also related works in this field is not significant enough to help students properly. In contrast, all are biased in foreign country faculties. So, students in our country are having trouble regarding getting their own county faculty information. In recent time, research field in our own country in also getting enrich and helpful for students. So, more students are being encouraged to do their research works in their own country. Thus, they are creating more scope for future generations.

1.4 Thesis Orientation

The rest of my thesis report have been organized in the following way -

Chapter 2 includes the literature review related to recommender systems. It has a wide description part about past related works on recommender systems.

Chapter 3 contains the details of methodology of the thesis. This chapter has four sections with several subsections. Section 3.1 contains the details of data-sets that

I used for my research work.Section 3.2 have a brief description of my used algorithm.

Chapter 4 contains the implementation process where I have discussed how my used algorithm is implemented and how I processed the data-sets.

Chapter 5 shows result analysis. There are sections with subsections. I have used necessary metrics for result analysis which I described within the subsections under the first section. Then, I have placed the results of the performance metrics for the algorithm. Under next section I placed a table showing the results of overall performance metrics for used algorithm.
Later, next section contains the conclusion. At the end of this paper, the references are given for citations in necessary format.

Chapter 6 contains the conclusion and future plan. The first section contains the conclusion. Later, next has information about my future work plan. Here, I have described the system of structure of my faculty recommender web platform which can be developed later using my research. At the end, the references are given for citations in necessary format.

Chapter 2

Literature Review

Several types of research have been done on collaborator recommendations, research paper recommender system, academic search etc. Tewaria and Barman in their research [13], analyzed user's profile according to their navigation history. They calculated similarities and dissimilarities among user resources and likings. At first, they analyzed learner profiles and content based profiles using relevant techniques. Then, they used different recommendation strategies to prepare relevant contents to the learners.

Salehi and Kamalabadi in their research paper [3] , proposed a framework based on it's design. Where they used vector space model along with good learner's ratings and suggest relevant contents according to the learner's similarity. In this way the learner's learning process has been improved by considering good learner's ratings. They gave priority to the attributes for their recommendation process.

Three different techniques used in the research paper to give more priority in recommendation based on similarities [2]. Momeni and the team apply user similarities to another set of users with their preferences. In order to find these similarities the used approaches are- memory-based, model based and hybrid. Here, they gave priority to the knowledge of particular users. The knowledge is calculated with their ratings considered as top selection criteria. The research paper of Manderson [1], which is very much in same basis with another one, also give priority to preferences.

Another research paper [14] presented by Hasan and Schwartz, a recommender system was built to give assistance in recommending Ph.D. advisors to the students. Different types of criteria has been considered here. They also gave hints in selecting supervisors based on their related research fields. They showed in which extend that study could make results. This approach was a statistical one which take student's performance in academic fields in consideration. They introduced fuzzy process approach at their work.

Gipp and his team in their research [9] narrated by citing Tan et al, using ETL (Extract, Transform and Load) and focusing on user based collaborative filtering model, the recommendation process becomes more compulsive. In addition, Gomley [5] showed how on-screen learning can be affected by it's recommendation system.

The research paper [11] by Vivek and Khadse was a recommender systems for movies. But their used techniques which is ontology based and discussed the future trends of this recommendation approach in the relevant fields. They made caomparisons of their used methods and focused on the contents.

Verma and Virk in their paper [7], introduced a multi criteria based recommendation system. This hybrid recommender system is built by using genre-based method. This proposed approach has been implemented by conetent based method. Most of the recommender system usually does not consider the user's potential and need in particular. Here, they mentioned about their methodology by proposing a collaborative voting approach which will help calculating difficulty parameter for recommendations.

Verma and Virk also proposed in his research work [6] by a developing genre-based system which can by genetic algorithm. They targeted the user's interst and previous knowledge about the content. They also made differences between the content based and hybrid approaches also the effectiveness of them in particular. They showed how individual approach could make impact on the performance of the results obtained by different approaches.

Ganarath and Sreenath [22] suggested us to look for different ways. They preferred to use mining techniques. They also used the popularity concepts in their work. In the research [4] ,Adhatrao and the team did different work by proposing a framework which is based on ID3 and C4.5 algorithm. Serrano [12] in his paper have addressed this issue and made a proposal for a recommendation model. The model is based on Random Neural Network. According to them, their work is capable of a intelligent recommending.

Badarneh and Alsakran in their work [8], have introduced an automated recommender system and what could be the further challenges of the system. They focused on data mining and pattern recognition in their work. In their research work [10], Jazayeriy and the team have developed system for cold user by using fast approach for categorized items. Survey on students found that their system was highly appreciated.

In the research by Pazzani and Billsus [19] , where they have developed a recommender system which is content based, has similar goals as the system built by Verma and Virk. Also their work has covered a vast domain which make their work more special. Accuracy of their work is quite satisfactory which encourages others to work on the similar process to get more accurate results in their way of prediction.

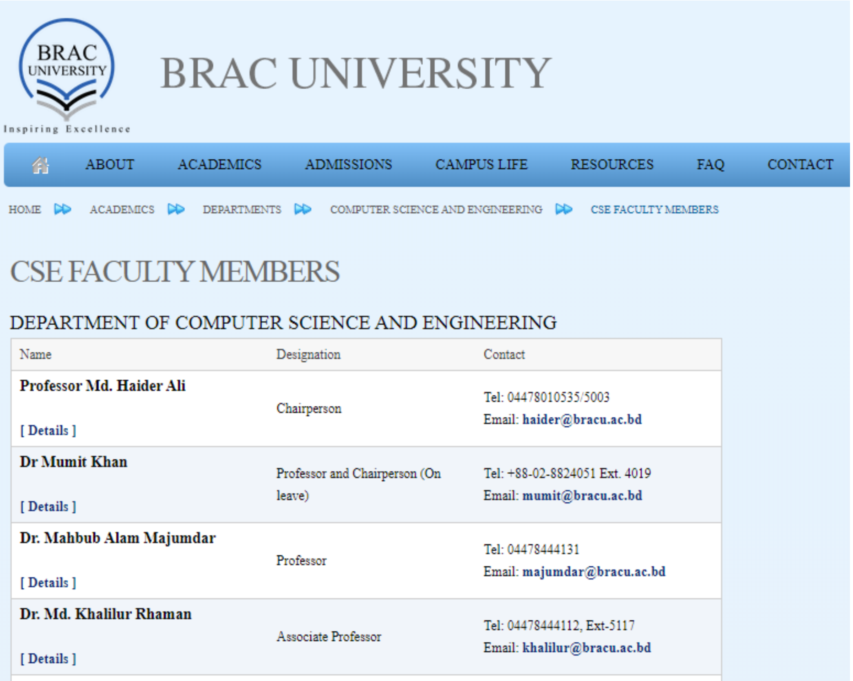
All these related works might have pretty good efficiencies in their fields but the initial purpose of the thesis work to recommend faculties to users will not be served. All these works thought that the students might have previous knowledge about faculties or supervisors to compare between them and choose the right person. But the scenario might be different at several times. So, students could fail to achieve their dream for not having enough details at one place.That's why having those things in one platform is very important. This thesis report presents a faculty recommender system where users will be able to find a list of faculties based on their research interest and other criteria. Users can contact the faculties or meet the faculties in person if there's a probability before making ultimate decision. Moreover, such common platform is not present in Bangladesh.

Chapter 3

Methodology

3.1 Data Collection

I have conducted a survey among the students of CSE department in BRAC University, Bangladesh to collect data. The survey was conducted through forms with necessary questions regarding the thesis. 50 students were taken part in the process and their feedback were taken for consideration. Further, the ratings were collected after the name of individual faculty. The faculty information were manually maintained in database. These data sets were then used to implement collaborative filtering algorithm.



CSE FACULTY MEMBERS

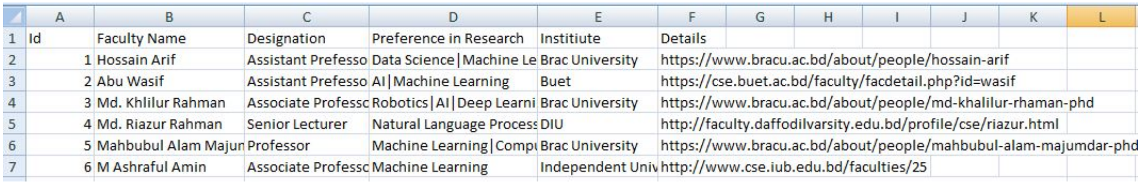
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Name	Designation	Contact
Professor Md. Haider Ali [Details]	Chairperson	Tel: 04478010535/5003 Email: **haider@bracu.ac.bd**
Dr Mumit Khan [Details]	Professor and Chairperson (On leave)	Tel: +88-02-8824051 Ext. 4019 Email: **mumit@bracu.ac.bd**
Dr. Mahbub Alam Majumdar [Details]	Professor	Tel: 04478444131 Email: **majumdar@bracu.ac.bd**
Dr. Md. Khalilur Rhaman [Details]	Associate Professor	Tel: 04478444112, Ext-5117 Email: **khalilur@bracu.ac.bd**

Figure 3.1: Data collection Sample[28]

3.1.1 Data-set for Algorithm

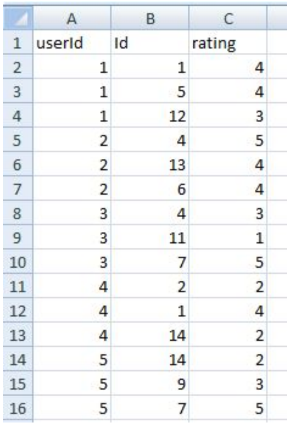
The following data-set (first five columns are attached) was used to implement filtering on particular user.



Id	Faculty Name	Designation	Preference in Research	Instiitute	Details
1	Hossain Arif	Assistant Prefesso	Data Science|Machine Le	Brac University	https://www.bracu.ac.bd/about/people/hossain-arif
2	Abu Wasif	Assistant Prefesso	AI|Machine Learning	Buet	https://cse.buet.ac.bd/faculty/facdetail.php?id=wasif
3	Md. Khlilur Rahman	Associate Professc	Robotics|AI|Deep Learni	Brac University	https://www.bracu.ac.bd/about/people/md-khalilur-rhaman-phd
4	Md. Riazur Rahman	Senior Lecturer	Natural Language Process	DIU	http://faculty.daffodilvarsity.edu.bd/profile/cse/riazur.html
5	Mahbubul Alam Majun	Professor	Machine Learning|Compu	Brac University	https://www.bracu.ac.bd/about/people/mahbubul-alam-majumdar-phd
6	M Ashraful Amin	Associate Professc	Machine Learning	Independent Univ	http://www.cse.iub.edu.bd/faculties/25

Figure 3.2: Partial Data-set for Filtering algorithm(part-1)

The following data-set (first few columns are attached) was used to implement filtering on particular user. After implementation of the algorithm, user is recommended with related faculties. I have implemented this algorithm on a data-set of 180 user profile which are random imaginary data with surveyed students data and 150 faculty members data form top renowned universities.
The following figure shows the given user database of individual liked data. These database is being used in the data analysis process of the algorithm.



userId	Id	rating
1	1	4
1	5	4
1	12	3
2	4	5
2	13	4
2	6	4
3	4	3
3	11	1
3	7	5
4	2	2
4	1	4
4	14	2
5	14	2
5	9	3
5	7	5

Figure 3.3: Partial Data-set for User Based Filtering(part-2)

The data-sets are used in the implementation process. At First to implement content based recommendation based on preference in research field and then to make predictions using both data-sets among preference ratings.

Survey on the Students:
I have collected data from the survey which was put into more useful form by sorting them in proper way in a .csv file. The data has been used in the analysis part. The questions for the survey were whether the system will be helpful or not and their preferred faculty member for research work and their preference ratings for them. Preference ratings are used only for calculation, not will be shown in results.

3.2 Algorithm

Luo described in his paper [17], a recommender system usually makes recommendations depending on a particular user's previous history and liking. This is also common in other machine learning techniques. The model recommends user liking for different conditions based on previous liking. In general, two types of algorithms are being used in recommender systems [15]. One is content based filtering and other one is by using collaborative filtering algorithm. collaborative filtering is categorized in two types also. First one is user based collaborative filtering and second one is item based collaborative filtering. Common recommendation for general terms also can be done by finding item-item similarity.

3.2.1 Collaborative Filtering Algorithm

The collaborative filtering (CF) algorithm is mainly on the basis of user-user interactions [20]. Collaborative filtering can be used in many different situations which is also a common advantage of using this algorithm. In order to make recommendation using this algorithm, users' past interactions has been considered for different items. In order to make recommendation to a particular user, that user is being grouped in a group with its most related neighbours with the help of user-user interactions. This process is being done by taking the preferences or ratings of those users in consideration. The next part comes in consideration is to find the similarities among users. We can find Similarity between users by using Cosine similarity or Jaccard similarity [27]. For KNN algorithm it is more preferable to use Cosine Similarity as distance metric. So, I used this technique as distance metric.

	Batman	Star Wars	Titanic
Bill	3	3	
Jane		2	4
cell1		5	

Table 3.1: Collaborative Filtering data-set example

In the table(3.1), a sample collaborative system has been shown where different users gave ratings to their liked different movies. From the table we can relate the similarities between the users ac according to their given preference for different items.

3.2.2 Content Based Approach

Content based approach is more to the point and less complex than collaborative approach. This method does not take the user user interactions in computation process which gives more focused results for the target purpose. So, to get results for interested research fields and get recommendations accordingly, this approach helps the users more precisely which makes this approach more robust. This helps to consider the recommendation process when it does not require user profile information or need to give recommendation in general. This approach also helps to reduce cold start problem and give better recommendations.

3.2.3 Cosine Similarity

Cosine similarity is the cosine of the angle between two vector in a n-dimensional space. I have used cosine similarity as distance metric to compute how much similar two users are with respect to each other.
We can compute between items with the help of cosine similarity. The angle between the items are the factor here. The two items (item A and item B) are more similar if the angle between them is more smaller. Similar items usually have more common attributes between them.

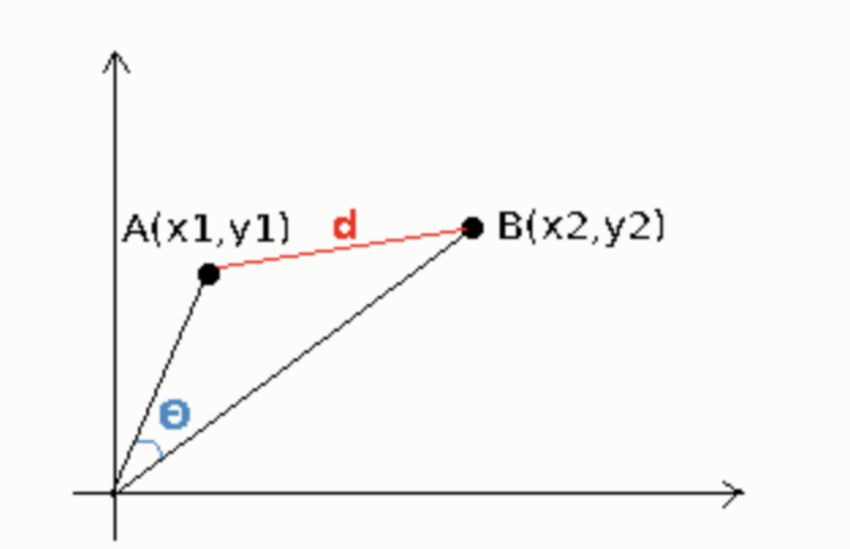


Figure 3.4: Cosine similarity between item A and item B [13]

The cosine similarity distance can be expressed by below formula. Here, values range between 0 and 1, where 0 is the most dissimilar and 1 is the most similar. In figure (3.1) the angle between two items (item A and item B) is denoted by theta. $X1$ and $X2$ represents the X and Y axis respectively.

$$\cos\theta = \frac{A.B}{\mid A \mid . \mid B \mid} \tag{3.1}$$

3.2.4 KNN Algorithm

K-Nearest Neighbor (KNN) is a straightforward algorithm that stores every access able cases an classifies the new cases dependant on a similarity measure [25]. This algorithm is a type of supervised ML algorithm. It is usually used for classification issues. However, K is the number of closest 'neighbors' that are taken into consideration. Selecting a suitable value for K is crucial step to get desired results from KNN. If we select the value of K as too low, then we will get random results mostly. So, it is better to set the value of K as a standard value for the calculation.

K-nearest neighbors (KNN) algorithm is a simple approach which works based on the distance of the sample from training set of the k nearest neighbors. After selecting k nearest neighbors, we get the classified results to predict for the target sample. Several distance metrics can be used in this algorithm. Common names of them are- Euclidean distance and Cosine similarity. The value of K is set according to the data size.

Chapter 4

Implementations

This algorithm is implemented based on particular user's data-set which is named as user data-set and the data-set of all faculties in the system which is named as faculty data-set. This user Data-set contains information of ratings about all faculties those were liked by the particular users. The faculty data-set contains all the information about faculties.

As the faculty data-set will contain all the information about all faculty members, using tf-idf vectorizer, the faculty members are matrix factorized with the corresponding values for their research area . After then, cosine similarity between those items are measured. This measurement will detect how similar are the items with each other. Items who have similar features – are similar to each other. The items which have higher cosine similarity or smaller angle between them are more similar to each other. Thus, by searching a faculty members name, the user can get faculty recommendation with similar faculties.

The user Data-set will contain all the information of all the users. This information contains their preferences given by them while registering or signing up for the system. Here is only a sample value ratings are considered. All ratings of the users are considered and cosine similarity between them are measured. This measurement will detect how similar are the users with each other.Users who have liked similar faculties – are similar to each other. The users which have higher cosine similarity score or smaller angle between them are more similar to each other.

After that, based on the similarities, groups are formed with nearest k neighbors of users. A particular user is grouped with its nearest k neighbors in such a way that all the other users in that group has similar interest and preferences.

Then, the names of the faculties are checked which are liked by the k-nearest neighbors. This names are retrieved from each neighbor's individual database. From all these faculties, the faculties whom the particular user has not yet liked are recommended to the user. If that particular user has already liked it, then it will not be recommended to him again.

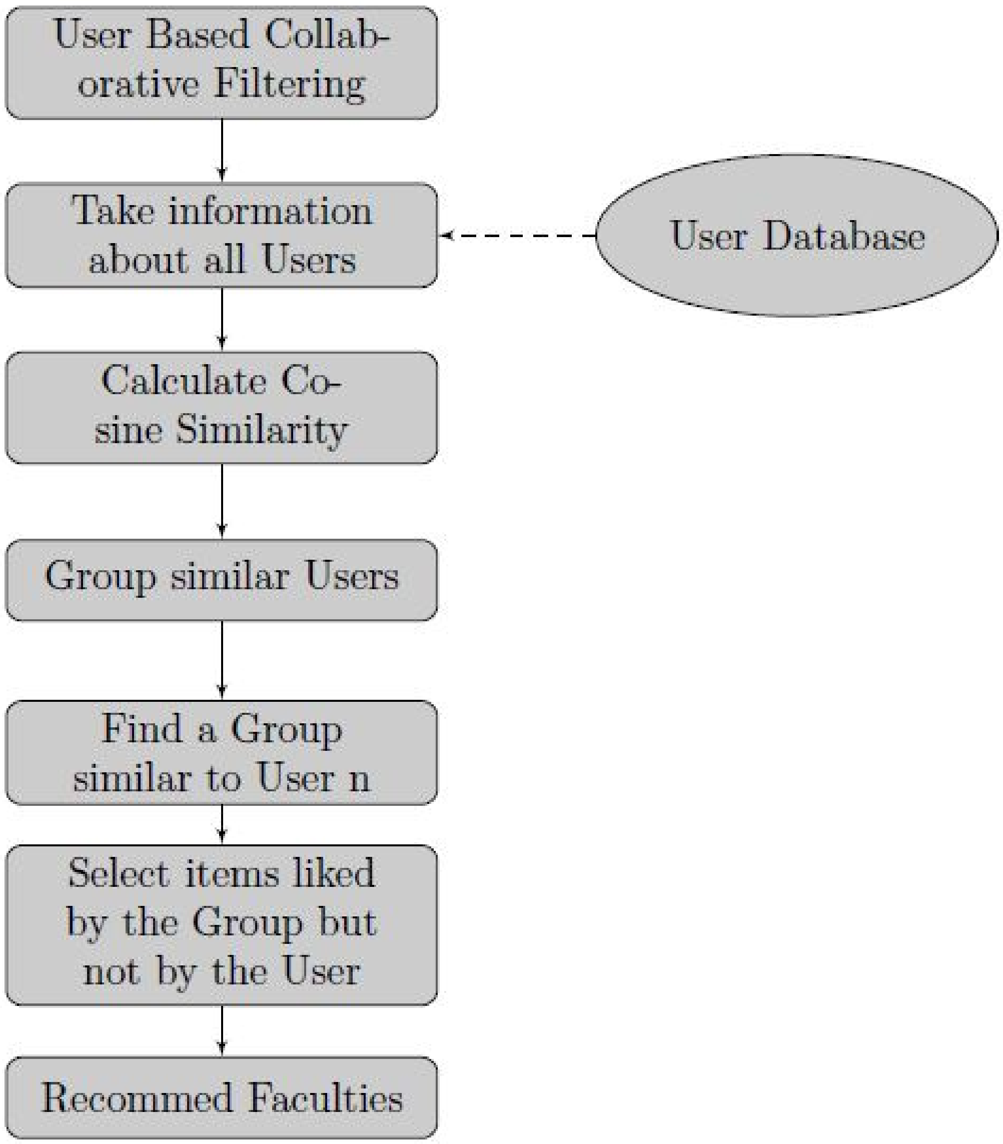

User Based Collaborative Filtering
Take information about all Users
User Database
Calculate Cosine Similarity
Group similar Users
Find a Group similar to User n
Select items liked by the Group but not by the User
Recommed Faculties

Figure 4.1: Flowchart of the algorithm

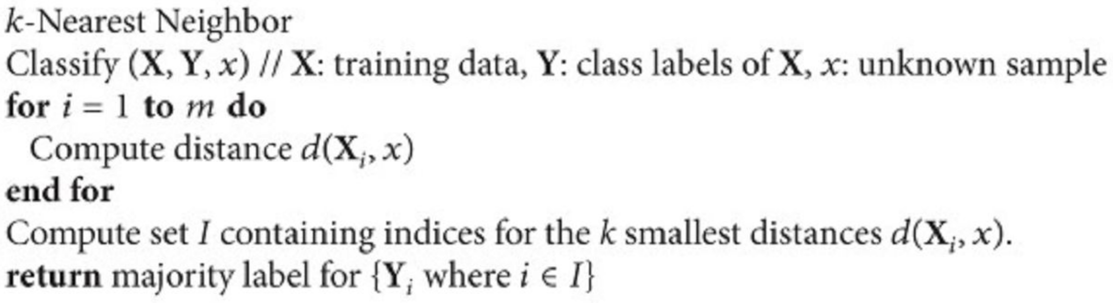
4.1 Cosine Similarity as distance metrics

A distance function provides distance between the elements of a set. If the distance is zero then elements are equivalent else they are different from each other. There are several techniques which can be used as distance metrics. Some of them are Euclidean distance, Jaccard similarity, Manhattan distance, mahalanobis distance, Cosine Similarity etc. Although similarity measures are often expressed using a distance metric, it is in fact a more flexible measure as it is not required to be symmetric or fulfill the triangle inequality. [27]

- If the matrix values only needed, binary calculations and no need to worry about the inner values, then we can use Jaccard Similarity.
- If the data sets of the formed matrix are not sparse and have almost values inside all the blanks, then it is wiser to use Euclidean distance.
- And if the data sets of the formed matrix are sparse and need to fulfill for further calculation, then it is better to use Cosine similarity. Hence I am using cosine similarity as distance vector along with KNN algorithm.

4.2 KNN Algorithm Pseudocode

The following figure (4.2) illustrates the K-nearest neighbor (KNN) algorithm pseudo code:



```
$k$-Nearest Neighbor
Classify $(\mathbf{X}, \mathbf{Y}, x)$ // $\mathbf{X}$: training data, $\mathbf{Y}$: class labels of $\mathbf{X}$, $x$: unknown sample
for $i = 1$ to $m$ do
  Compute distance $d(\mathbf{X}_i, x)$
end for
Compute set $I$ containing indices for the $k$ smallest distances $d(\mathbf{X}_i, x)$.
return majority label for $\{\mathbf{Y}_i$ where $i \in I\}$
```

Figure 4.2: Pseudocode for KNN Algorithm [16]

4.3 Language and libraries used

Here, I used Python programming language and it's built-in libraries for data analysis and showing the results. Among the libraries, pandas, numpy, scikit-learn, matplotlib etc. are used in the implementation process. Python language provides a better environment for data analysis and visualization.

Numpy stands for numerical python. Numpy array is a N-dimensional array object which is in the form of rows and columns. It can also be used as an structured multi-dimensional container for collective data. [21]

Pandas library is used for data manipulation, analysis and cleaning. Python pandas library is well suited for different kinds of data. Using this library, we can calculate a lot of operations. Operations can be with missing data, data frames, series, group by etc. [21]

Scikit-learn is also a popular library for Python. It includes various algorithms. It also supports Python libraries like NumPy and others. [21]

Chapter 5

Result Analysis

This chapter is to evaluate the results given by the algorithm. Here I have discussed them in two sections.

5.1 Algorithm Evaluation Process

The algorithm is evaluated by using various metrics measurement and calculations. All the measurement metrics are described here in details.

5.1.1 Confusion Matrix

A confusion matrix is widely used for evaluating the accuracy and performance of any algorithm. Confusion matrix is usually a table having different combinations of actual and predicted values [24]. The measuring metrics give us a general overview. But the confusion matrix make us to visualize the actual results of the prediction. From table(5.1), we can see each categorical representations of every segments. From a confusion matrix, further we can calculate accuracy, precision, recall and F1 score of an algorithm.

Terms related to the confusion matrix:

- True Positives (TP) - True positives are the cases when both the actual class and predicted class is true.
- True Negatives (TN) - True negatives are the cases when both the actual class and predicted class is false.
- False Positives (FP) - False positives are the cases when the actual class is false but predicted class is true.
- False Negatives (FN) - False negatives are the cases when the actual class is true but predicted class is false.

		Actual Value		
		Positives	Negatives	Total
Predicted Value	Positives	m	n	$m+n$
	Negatives	p	q	$p+q$
	Total	$m+p$	$n+q$	N

Table 5.1: Confusion Matrix

5.1.2 Accuracy

Accuracy is the fraction of all the assessments which are correct between all the assessments. [24]

$$accuracy = \frac{TruePositives + TrueNegatives}{Totalnumberofpredictions} \tag{5.1}$$

5.1.3 Precision

Precision is a performance measure which tells us the ratio of the prediction was actually correct out of all the classes. [24]

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{5.2}$$

5.1.4 Recall(Sensitivity)

Recall is the ratio of assessments that are true positives between all positive assessments in actual class. [24]

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{5.3}$$

5.1.5 F1 Score

F1 score uses Harmonic Mean in place of Arithmetic Mean. It is the weighted average of Precision and Recall [24]

$$F1 = \frac{2 * (sensitivity * precision)}{sensitivity + precision} \tag{5.4}$$

5.2 Algorithm Performance Evaluation

I have described all the necessary measure and the related formulas in the previous section. This section represents the results of the algorithm after applying those measurement metrics.
The table 5.2 is showing the confusion matrix for collaborative filtering and table 5.3 is showing measured performance metrics results of the algorithm.

		Actual Value		
		Positive	Negative	Total
Predicted Value	Positive	40	16	
	Negative	32	112	
	Total			N

Table 5.2: Confusion Matrix for used algorithm

performance metrics	Results Obtained
Accuracy	76 %
Precision	79 %
Recall	76 %
F1 Score	77 %

Table 5.3: performance measurement results

5.3 Comparison

I have designed a faculty recommender system for which recommends faculty names to users considering their past preferences and field of interest. Unfortunately, I have found very few related research [26] papers for my recommendation system which is closely connected to my research. So that I could compare my work with them. Here I compared my results with traditional performance measurement metrics. From my result, the values of all performance measures for my filtering algorithm is quite similar to those whose fields are relevant.

Accuracy score is used in wide range to measure and differentiate the performance of any algorithm. The accuracy score of my used algorithm is 76 %. Based on the accuracy score, we can say that my used algorithm performs quite well to give results.

The paper I cited here, they have used different techniques for individual algorithms but here i used one particular algorithm which is KNN algorithm. I also used sparse data where the data-sets are filled, not sparse. Still, the accuracy in both cases are nearly same. But there precision results is showing 100% which means they predict no error for negative predictions which is really difficult to do for an algorithm.

At the end, considering all the conditions, I made this recommender system to work for users who will give necessary information needed and make recommendations for those users.

Chapter 6

Conclusion and Future Plan

6.1 Conclusion

K Nearest Neighbors is one of the most popular machine learning algorithms because of its predictive power. In this thesis report, I proposed a faculty recommendation system, which uses data sets to give the users vast information about faculty or supervisor selection difficulties. It gives specific direction on variety of options having the user's want in consideration. The proposed system recommends faculties or supervisors on the basis of selecting research fields by the user. My system can compute larger data-sets and recommend the faculties or professors according to the required criteria. The research may help students widely in numerous ways. Evaluation results of the proposed algorithm shows that allowing users to select the research fields or preferred institutions provides more perfection in the recommendation process.

6.2 Future Plan

In my research, I have developed a model to make faculty members recommendations to the users. I have used machine learning classifier algorithm to do the prediction job. For my future work, I want to improve the data-sets first. I also want to work on data mining techniques so that the data-set could be more categorized and organized.

I want to work with a team or try myself and make a Web Platform using the works done for this research paper. The platform will be a website based recommender system. The users will be mainly the students who will try to seek information about faculties. They can register on the website using their necessary information. The registration page will show the users to enter required information about them. In general, top recommendations will be displayed when using without user information.

The users can sign in with their necessary information. The required information will be their email and password. After logging into the profile, the users will be displayed their profile pages. Here, users will be able to search faculties, see their details, like them and can view a list of recommended faculties for him.

When a user will search for faculty names, he will be needed to type the preferred research field or title of the faculty he will be looking for. The search process will be performed on the database and the detail information about that faculty will be gathered and displayed to the website by using necessary software.

Necessary information about the faculty which will be required for the algorithm to give future predictions of the contents to the users will be stored in the particular database. This will help the system to make more efficient recommendations in future.

Bibliography

[1] E. M. PHILLIPS and D. S. PUGH, *How to get a PhD: A handbook for students and their supervisors.* McGraw-Hill EducationUK, 1996.

[2] M. Momeni, B. Samimi, M. A. Afshari, M. H. Maleki, and J. Mohammadi, “Selection process of supervisor for doctoral dissertation using analytical network process (anp): An iranian study.”, *Journal of Management and Strategy*, 2011.

[3] M. Salehi and I. N. Kamalabadi, “A hybrid attribute-based recommender system for e-learning material recommendation.”, *International Conference on Future Computer Supported Education*, 2012.

[4] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha, and V. Honrao, “Predicting student’s performance using id3 and c4.5 classification algorithms.”, *International Journal of Data Mining Knowledge Management Process*, vol. 3, 2013.

[5] C. Gormley and Z. Tong, *Elasticsearch: The Definitive Giude: A Distributive Real-Time Search and Analytics Engine.* O’Reilly Media, INc, 2015.

[6] D. A. Verma and H. K. Virk, “A hybrid genre-based recommender system for movies using genetic algorithm and knn approach.”, *IJIET*, 2015.

[7] ——, “A hybrid online genre-based recommender system.”, *IJCST*, vol. 6, 2015.

[8] A. Al-Badarneh and J. Alsakran, “An automated recommender system for course selection.”, *International Journal of Advanced Computer Science and Applications*, 2016.

[9] J. Beel, B. G. S. Langer, and C. Breitinger, “Research-paper recommender systems: A literature survey.”, *International Journal on Digital Libraries*, vol. 17(4), pp. 305–388, 2016.

[10] H. Jazayeriy, S. Mohammadi, and S. Shamshirband, “A fast recommender system for cold user using categorized items.”, *Applied Modern Mathematics in Complex Networks*, 2018.

[11] V. P. Khadse, A. P, S. M. Basha, N. Iyengar, and R. D. Caytiles, “Recommendation engine for predicting best rated movies.”, *International Conference of Advance Science and Technology*, vol. 110, 2018.

[12] W. Serrano, “Intelligent recommender system for big data applications based on the random neural network.”, *International Neural Network Society Conference on Big Data*, 2018.

[13] A. S. Tewari, J. P. Singh, and A. G. Barman, “Generating top-n items recommendation set using collaborative, content based filtering and rating variance”, *International Conference on Computational Intelligence and Data Science.*, vol. 132, pp. 1678–1684, 2018.

[14] M. A. Hasan and D. Schwartz, “A multi-criteria decision support system for ph.d. supervisor selection: A hybrid approach.”, *52th Annual Hawaii International Conference on System Sciences*, 2019.

[15] M. Howe. (). Pandora’s music recommender, [Online]. Available: https://courses.cs.washington.edu/courses/csep521/07wi/prj/michael.pdf. (accessed: 2015).

[16] J. K. Hyun. (). A machine learning approach for specification of spinal cord injuries using fractional anisotropy values obtained from diffusion tensor images, [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24575150. (accessed: 2014).

[17] S. Luo. (). Introduction to recommender system: Approaches of collaborative filtering: Nearest neighborhood and matrix factorization, [Online]. Available: https://towardsdatascience.com/intro-to-recommender-system-collaborative-filtering-64a238194a26. (accessed: 2018).

[18] D. Manderson. (). ASKING BETTER QUESTIONS : APPROACHING THE PROCESS OF THESIS SUPERVISION, [Online]. Available: http://docs.manupatra.in/newsline/articles/Upload/497B07D5-90B6-411E-9A9E-C5409070F00E.pdf.

[19] M. J. Pazzani and D. Billsus. (). Content-based recommendation systems, [Online]. Available: https://www.researchgate.net/publication/280113634_Content-Based_Recommendation_Systems. (accessed: 2017).

[20] C. Pinela. (). Recommender Systems — User-Based and Item-Based Collaborative Filtering, [Online]. Available: https://medium.com/@cfpinela/recommender-systems-user-based-and-item-based-collaborative-filtering-5d5f375a127f.

[21] T. N. Prabhu. (). Top Python Libraries Used In Data Science, [Online]. Available: https://towardsdatascience.com/top-python-libraries-used-in-data-science-a58e90f1b4ba.

[22] G. R and S. R. (). Calculating popularity using a simple algorithm, [Online]. Available: http://vixra.org/abs/1312.0052. (accessed: 08.12.2013).

[23] B. Roccaa. (). Introduction to recommender systems, [Online]. Available: https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e.

[24] M. Sansura. (). Performance Metrics for Classification problems in Machine Learning, [Online]. Available: https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b.

[25] D. Subramanian. (). A Simple Introduction to K-Nearest Neighbors Algorithm, [Online]. Available: https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b.

[26] U. S. Tasnuva, A. A. Aumi, and S. I. Shishir. (). Implementing a recommender system for cs undergraduate students using machine learning, [Online]. Available: http://hdl.handle.net/10361/12294. (accessed: 2013).

[27] T. N. J. D. Ullman. (). MMDS and Automata MOOC's, [Online]. Available: http://infolab.stanford.edu/~ullman/mmds/ch9.pdf.

[28] B. University. (). Cse faculty members, [Online]. Available: http://www-cs-faculty.stanford.edu/~uno/abcde.html. (accessed: 01.09.2016).