Journal of Applied Research and Multidisciplinary Studies (JARMS)
Volume No. 1, Issue No. 1, Spring 2020
ISSN$_{(P)}$: 2707-5087 ISSN$_{(E)}$: 2707-5095
Homepage: https://journals.umt.edu.pk/index.php/JARMS

Article:

Automatic Generation of Teachers' Course Preferences Using Document Clustering

Author(s):

Amna Shoukat
Malik Tahir Hassan
Hira Asim

Online Published: Spring 2020

Article DOI: https://doi.org/10.32350/jarms.11.01

Article QR Code:

Amna Shoukat

To Cite Article:

Shoukat, A., Hassan, M. T., & Asim, H. (2020). Automatic generation of teachers' course preferences using document clustering. *Journal of Applied Research and Multidisciplinary Studies, 1*(1), 01–16. Crossref

A publication of the
School of Professional Advancement
University of Management and Technology, Lahore, Pakistan.

Automatic Generation of Teachers' Course Preferences using Document Clustering

Amna Shoukat*, Malik Tahir Hassan and Hira Asim
University of Management and Technology, Lahore, Pakistan

Abstract

The current study examined the automated course preferences of teachers using document clustering. Data regarding teachers' course preferences and course outlines were collected and preprocessed for further analysis. Two separate clustering solutions were generated for teachers and courses datasets. The clustering solution for teachers contained clusters of similar faculty members grouped together on the basis of their course preferences and courses taught by them in previous years. The clustering solution generated for courses contained the list of course outlines of assigned courses. Good quality clusters for both teachers and courses were generated using K-means clustering method in CLUTO software package. The generated clustering solutions were mapped for automated generation of course preferences for each teacher in the dataset. Precision, Recall and F-measure values were also reported and they indicated promising results.

Keywords: course allocation, data mining, document clustering, higher education, teaching quality, teaching management

Introduction

Data mining is an actively growing field of computer science that deals with facts and statistics to generate knowledge and to solve complex real-life problems. In this research work, automated generation of teachers' course preferences was performed using data mining techniques with an aim to assist the higher management of universities in effective course allocation. Course allocation to teachers is a complex problem that every university's higher authorities such as deans and CODs face at the start of every semester. It is a very challenging situation for them to allocate courses in such a way that all teachers are satisfied and also possess sufficient expertise in their assigned courses. Teachers' expertise and preference for courses have a strong impact on infusing quality knowledge into students. This research work will help authorities in Higher Education Institutes (HEIs) in assigning courses to faculty members in a better way, keeping in view their preferences and their respective department's needs.

*Corresponding author: hira.asim@umt.edu.pk

We collected 45 different course outlines of the subjects related to Computer Science (CS) and Software Engineering (SE) degree programs. These course outlines were downloaded from the official website of the Higher Education Commission (HEC) of Pakistan[1]. Moreover, the data regarding teachers' course preference and course expertise taught by them were also gathered and analyzed.

The proposed solution is based on document clustering which is a data mining technique used to organize a large collection of unlabeled documents into homogenous groups, automatically. These groups are referred to as clusters. Each cluster comprises objects that are similar in nature. Document clustering is performed based on descriptors (groups of words that represent the content of a cluster). Due to the ongoing transition of the world towards a paperless environment and the dominance of web in our lives, the importance of textual document clustering has increased. Document clustering helps in improving the classification of documents that directly or indirectly influences information retrieval and storage. Moreover, it also helps in handling an enormous collection of documents in order to produce semantically and readily understandable knowledge patterns (Liu, Wang, Xu & Guan, 2006). In this research work, we created two separate datasets in the form of textual documents for teachers and courses. Teachers' dataset contained information about faculty members regarding courses taught by them in previous years and their preferred list of courses. Similarly, courses' dataset included the list of standard course outlines. We used the software package CLUTO: A Clustering Toolkit to discover teachers and courses' clusters.

CLUTO provides two standalone programs for document clustering known as *vcluster* and *scluster* (Karypis, 2003), *vcluster* supports matrix format dataset, while *scluster* supports graph format dataset as input. In this research work, *vcluster* program was applied on both courses' and teachers' preprocessed datasets to generate clustering solutions along with the information about discriminant and descriptive terms. CLUTO provides different parameters to control a clustering solution such as cluster method, similarity function, criteria function and number of clusters (k).

Hence, we obtained two types of clusters. The first were the clusters of teachers containing the groups of faculty members who taught similar courses and provided similar course preferences. The second were the clusters of courses containing groups of similar course outlines of the degree programs of CS and SE. In order to establish a link between the two sets of clusters, they were mapped onto each other on the basis of the discriminant terms of each cluster from both teachers' and courses' clustering solutions. Once the one-to-one mapping of teachers' and courses' clusters was finalized, each member of teachers' clusters could be recommended courses from the relevant courses' clusters. For the evaluation of the proposed system, the previous year's course preferences comprising the actual data provided by teachers of the Software Engineering (SE) department of the University of Management and Technology (UMT) were used. Precision, Recall and F-measure values were also reported.

Literature Review

Data mining techniques have been used in the field of education to solve different problems such as timetable scheduling, faculty performance evaluation, students' feedback analysis etc. Karimpour & Mavizi (2016) reviewed course timetable scheduling problems with the aim to provide a solution in such a way that it meets each lecturer's satisfaction level and also minimizes the loss of resources in the department. In their proposed methodology, each department initially allocates courses to teachers and then clustering algorithms are applied to group all common lecturers. Traversing techniques are applied to find unused resources within the department. After the execution of clustering and traversing processes, mapping action is performed based on the principles of common constraints on redundant resources.

Singh, (2017) tried to solve the course timetabling problem using genetic algorithm. He explored different selection methods such as roulette wheel, ranking selection and tournament selection to perform crossover and mutation on selected fields. Ganguli & Roy (2017) used a heuristic approach named graph coloring to solve the course timetabling problem in their study. Graph coloring approach was applied on multiple datasets to obtain optimal solutions. Among all optimal solutions, only those solutions were shortlisted which satisfied all hard and soft constraints.

The research study of Kaur & Kaur (2014) introduced a new approach for solving course timetable problem. The authors applied modified k-means clustering algorithm and rule based classifier technique to get an optimal solution of the course-timetabling problem.

Faculty performance is evaluated on the basis of factors such as student feedback, management feedback, institutional support in terms of finance, research activities and managerial support (Pal & Pal, 2013). Hidden patterns are identified with the help of classification algorithms and 'Naïve Bayes' yields the highest accuracy. Results have shown that potential productivity of the faculty can help higher authorities to evaluate the details about teachers' motivation, growth and decline.

Dhanalakshmi, Bino & Saravanan (2016) performed opinion mining on students' feedback data collected via survey. Supervised machine learning algorithms were applied using the RapidMiner tool. The findings revealed that in terms of recall and accuracy, Naïve Bayes yielded the best results, whereas K-nearest neighbor showed the highest precision. From the data comprising students' feedback key features affecting teaching and learning were classified as positive and negative, keeping in account the features which needed improvement.

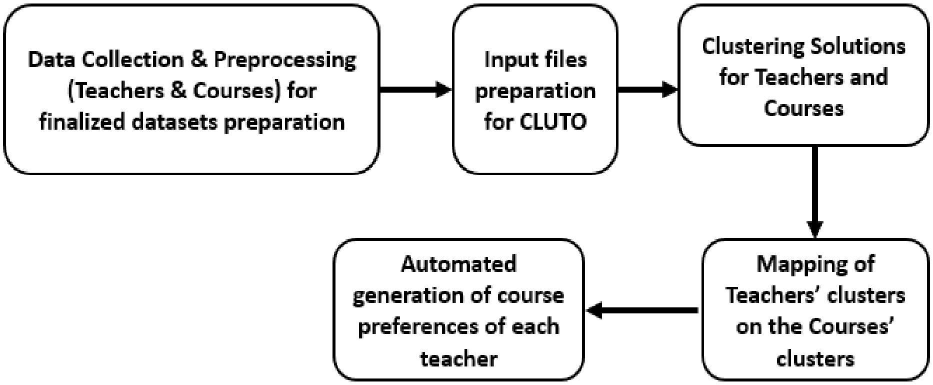
There is a surplus of information available in electronic form. This is why text document clustering is considered as a rapidly growing area of research these days. Jensi & Jiji (2014) surveyed different optimization techniques of text document clustering. There are algorithms available which effectively navigate and organize information and also provide localized search results. By applying high speed and high quality optimization algorithms, a global optimal clustering solution can be obtained. Abualigah, Khader & Al-Betar (2016) introduced multi-objective based technique which combined similarity measures and refined text document clustering methods. They measured the performance of the proposed technique using k-means text clustering approach. Experimental research was carried out on seven text document datasets. They demonstrated that text document clustering is the right mining tool for unsupervised text clustering as it categorizes different documents in the same cluster. Hassan, Karim, Kim & Jeon (2015) stated that the existing techniques related to document clustering rely on generic measures of similarity and mostly ignore the semantics of the terms present in the documents. According to the authors, document clustering algorithms should generate clusters that are semantically relevant. So, in their paper they proposed an algorithmic framework labelled Clustering by Discrimination Information Maximization (CDIM).

Methodology

This section addresses the process of data collection and it's preprocessing, data file preparation for CLUTO and document clustering. The mapping of teachers' clusters onto courses' clusters for the automated generation of course preferences is also discussed further. Figure summarizes the entire process of document clustering.

Figure 1

Process Flow of Automatic Generation of Teachers' Course Preferences Using Document Clustering Algorithms in CLUTO


Data Collection & Preprocessing (Teachers & Courses) for finalized datasets preparation
Input files preparation for CLUTO
Clustering Solutions for Teachers and Courses
Mapping of Teachers' clusters on the Courses' clusters
Automated generation of course preferences of each teacher

Dataset Collection and Preprocessing

For automated generation of course preferences, we gathered data related to both courses and teachers to prepare our datasets. To gather the data of courses, we collected course outlines of core as well as technical / elective courses related to Software Engineering (SE) and Computer Science (CS) degree programs from the official website of the Higher Education Commission (HEC) of Pakistan for the year 2016-2017[2]. The collected course outlines were presented in the form of unstructured textual documents.

After the collection of course outlines, we gathered the data of teachers related to courses they taught in the previous years and course preferences given by them in the previous year (2018) from the Software Engineering (SE) department of the University of Management and Technology (UMT), Lahore, Pakistan. Table 1 shows a sample of the dataset related to faculty members.

After collecting the relevant information regarding courses and teachers, we performed the following preprocessing steps to get a complete dataset:

a) Individual course outlines were integrated into a single text file referred to as courses' dataset.
b) The gathered information about each faculty member, as shown in Table 1, was integrated into a single text file referred to as teachers' dataset.

c) Stop word00s and special characters were removed.
d) Stemming and lemmatization steps were performed.

Table 1

Examples of Courses Taught in Previous Years and Preferences Given by Teachers in the Previous Year (2018)

TID	Courses Taught in Previous Years	Preferred Courses of 2018
2	Formal Methods in Software Engineering, Software Engineering, Data Structures and Algorithms, Discrete Structures, Operating Systems	Artificial Intelligence, Data Structures and Algorithms, Introduction to Computing, Object Oriented Programming, Programming Fundamentals, Software Engineering, Compiler Construction, Computer Networks, Formal Methods in Software Engineering, Operating Systems
4	Distributed Systems, Secure Software Development, Web Technologies, Software Project Management	Distributed Systems, Secure Software Development, Object Oriented Programming, Programming Fundamentals, Web Technologies
6	Networking, Operating Systems	Computer Networks, Operating Systems
8	Big Data Programming, Information Retrieval, Data Mining, Machine Learning	Machine Learning, Database Systems, Data Warehousing, Natural Language Programming, Big Data Programming, Information Retrieval
10	Compiler Construction, Design Patterns and Refactoring, Programming Fundamentals	Data Structures and Algorithms, Object Oriented Programming, Software Engineering, Compiler Construction, Design Patterns and Refactoring, Database Systems

Preparation of Input Files for CLUTO

In order to perform document clustering using CLUTO, we needed to generate three separate input files against courses' and teachers' datasets as described below.

a) *.mat file:* A sparse matrix format file that contained information about the frequency of each unique term present in the document. The header line of

.mat file contained additional information about the number of rows, columns and non-zero entries. In our case, the number of rows showed the total number of course outlines / faculty members, while the number of columns represented the total number of unique terms present in courses' / teachers' dataset. The number of non-zero entries showed the total number of all the terms with a frequency greater than or equal to one from both datasets.

b) *.rclass* file*:* This file contained actual labels assigned manually to each record for the purpose of evaluation and better understanding of results. In total, seven class labels were assigned to course outlines as well as faculty members. These labels included *SE* (Software Engineering), *MWAD* (Mobile and Web Application Development), *CS* (Computer Science (Programming)), *ML* (Machine Learning / Artificial Intelligence), *Netw* (Networking), *Secu* (Security), and *Data* (Database).

c) *.clabel* file*:* This file contained the list of all the unique terms present in the dataset. We created two *.clabel* files containing all the unique terms from courses' as well as teachers' dataset files.

Discovery and Mapping of Clusters

After creating the input files *(.mat, .rclass, .clabel*) as mentioned above for both courses' and teachers' datasets, we performed document clustering to discover different clusters for courses and teachers, separately. To generate clustering solution in CLUTO, we needed to provide input parameters such as the number of clusters (k), cluster method (clustering algorithm), criteria function and similarity function along with the three input files (*.mat, .rclass, .clabel*) (Karypis, 2003). Different settings of these parameters were implemented to reach high purity (close to 1) and low entropy (close to 0) clustering solutions. Clustering solution of CLUTO yielded additional information about the descriptive and discriminating terms of each cluster. Discriminating terms are keywords on the basis of which documents in a cluster can be distinguished from the documents included in other clusters. Descriptive terms are the list of common / repeatedly occurring words in the documents included within a cluster.

For the mapping of teachers' clusters onto courses' clusters, we reviewed the discriminating terms of each cluster (teachers' and courses'). In simple words, the clusters of courses and teachers were mapped onto each other on the basis of similar discriminating terms for automatic generation of teachers' course preferences. Section 5 discusses the mapping of courses' clusters onto teachers' clusters in detail.

Results

This section will discuss the clustering solutions obtained for both courses' and teachers' datasets in CLUTO.

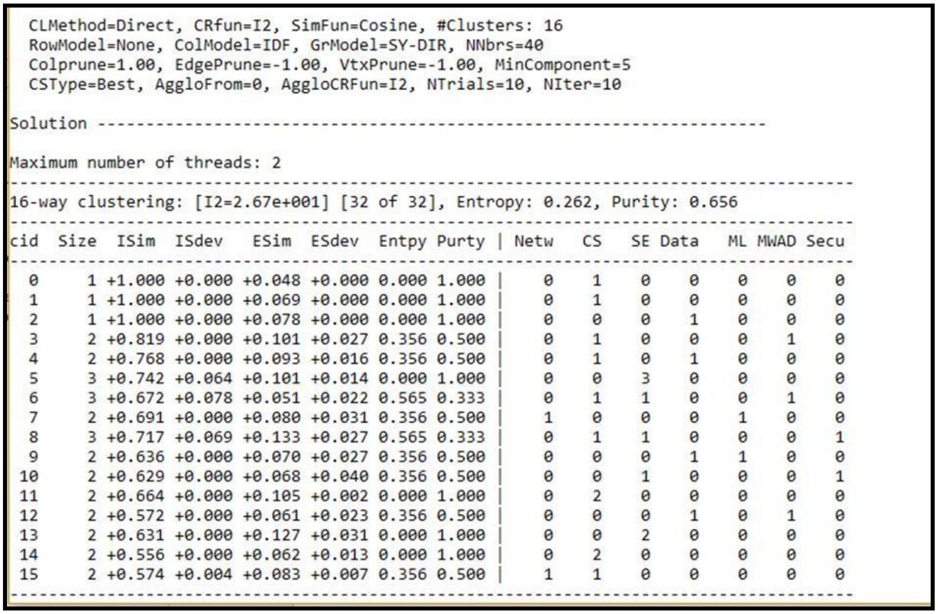
Clustering Teachers and Courses in CLUTO

To cluster both teachers and courses, we generated many clustering solutions by taking (k = 15, 16, 17, 18) clustering methods such as *repeated bisection*, *direct*, *agglomerative* criteria functions including (*G1, H1 E1, I1, I2*) and similarity functions such as *cosine*, *Euclidean distance*, *correlation coefficient*. We also referred to the official manual of CLUTO for more details regarding input parameters (Karypis, 2003). After obtaining different clustering solutions, it was observed that for k = 16 (k is the number of clusters), cluster method = *direct*, criteria function = *I2* and similarity function = *cosine.* We achieved good clustering solutions for both teachers' and courses' datasets by obtaining maximum purity and minimum entropy values.

Figure 2 and Figure 3 depict the clustering solutions we obtained for both courses' and teachers' datasets by setting the input parameters mentioned earlier.

Figure 2

Clustering Solution for Courses' Dataset for k = 16, Cluster Method = direct, Criteria Function = I2 and Similarity Function = Cosine



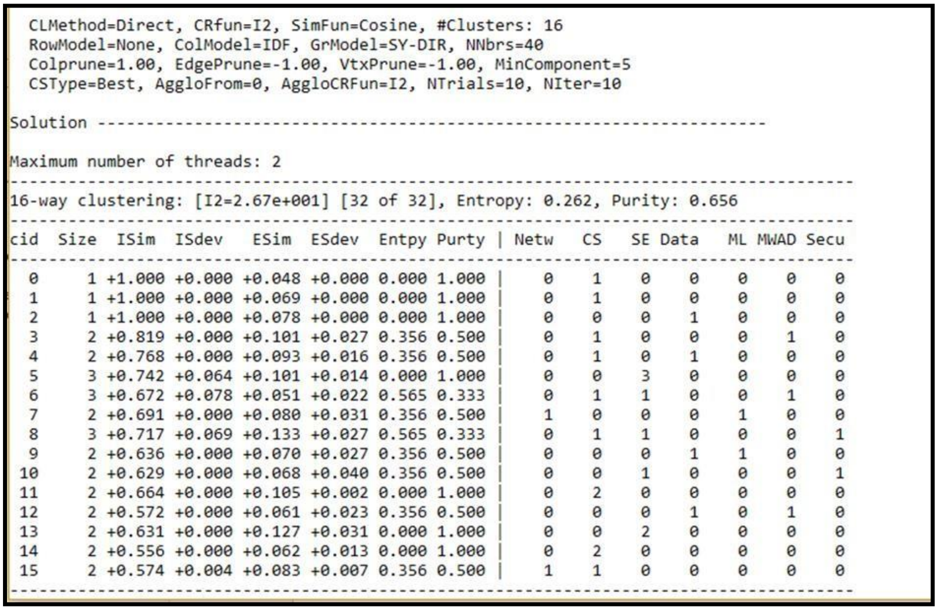
```
  CLMethod=Direct, CRfun=I2, SimFun=Cosine, #Clusters: 16
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution -------------------------------------------------------------------

Maximum number of threads: 2
-------------------------------------------------------------------------------
16-way clustering: [I2=2.67e+001] [32 of 32], Entropy: 0.262, Purity: 0.656
-------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | Netw   CS   SE Data   ML MWAD Secu
-------------------------------------------------------------------------------
  0     1 +1.000 +0.000 +0.048 +0.000 0.000 1.000 |    0    1    0    0    0    0    0
  1     1 +1.000 +0.000 +0.069 +0.000 0.000 1.000 |    0    1    0    0    0    0    0
  2     1 +1.000 +0.000 +0.078 +0.000 0.000 1.000 |    0    0    0    1    0    0    0
  3     2 +0.819 +0.000 +0.101 +0.027 0.356 0.500 |    0    1    0    0    0    1    0
  4     2 +0.768 +0.000 +0.093 +0.016 0.356 0.500 |    0    1    0    1    0    0    0
  5     3 +0.742 +0.064 +0.101 +0.014 0.000 1.000 |    0    0    3    0    0    0    0
  6     3 +0.672 +0.078 +0.051 +0.022 0.565 0.333 |    0    1    1    0    0    1    0
  7     2 +0.691 +0.000 +0.080 +0.031 0.356 0.500 |    1    0    0    0    1    0    0
  8     3 +0.717 +0.069 +0.133 +0.027 0.565 0.333 |    0    1    1    0    0    0    1
  9     2 +0.636 +0.000 +0.070 +0.027 0.356 0.500 |    0    0    0    1    1    0    0
 10     2 +0.629 +0.000 +0.068 +0.040 0.356 0.500 |    0    0    1    0    0    0    1
 11     2 +0.664 +0.000 +0.105 +0.002 0.000 1.000 |    0    2    0    0    0    0    0
 12     2 +0.572 +0.000 +0.061 +0.023 0.356 0.500 |    0    0    0    1    0    1    0
 13     2 +0.631 +0.000 +0.127 +0.031 0.000 1.000 |    0    0    2    0    0    0    0
 14     2 +0.556 +0.000 +0.062 +0.013 0.000 1.000 |    0    2    0    0    0    0    0
 15     2 +0.574 +0.004 +0.083 +0.007 0.356 0.500 |    1    1    0    0    0    0    0
-------------------------------------------------------------------------------
```

Figure 3

Clustering Solution for Teachers' Dataset for k = 16, Cluster Method = direct, Criteria Function = I2 and Similarity Function = Cosine



```
  CLMethod=Direct, CRfun=I2, SimFun=Cosine, #Clusters: 16
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution ---------------------------------------------------------------------

Maximum number of threads: 2
-------------------------------------------------------------------------------
16-way clustering: [I2=2.67e+001] [32 of 32], Entropy: 0.262, Purity: 0.656
-------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | Netw   CS   SE Data   ML MWAD Secu
-------------------------------------------------------------------------------
  0     1 +1.000 +0.000 +0.048 +0.000 0.000 1.000 |    0    1    0    0    0    0    0
  1     1 +1.000 +0.000 +0.069 +0.000 0.000 1.000 |    0    1    0    0    0    0    0
  2     1 +1.000 +0.000 +0.078 +0.000 0.000 1.000 |    0    0    0    1    0    0    0
  3     2 +0.819 +0.000 +0.101 +0.027 0.356 0.500 |    0    1    0    0    0    1    0
  4     2 +0.768 +0.000 +0.093 +0.016 0.356 0.500 |    0    1    0    1    0    0    0
  5     3 +0.742 +0.064 +0.101 +0.014 0.000 1.000 |    0    0    3    0    0    0    0
  6     3 +0.672 +0.078 +0.051 +0.022 0.565 0.333 |    0    1    1    0    0    1    0
  7     2 +0.691 +0.000 +0.080 +0.031 0.356 0.500 |    1    0    0    0    1    0    0
  8     3 +0.717 +0.069 +0.133 +0.027 0.565 0.333 |    0    1    1    0    0    0    1
  9     2 +0.636 +0.000 +0.070 +0.027 0.356 0.500 |    0    0    0    1    1    0    0
 10     2 +0.629 +0.000 +0.068 +0.040 0.356 0.500 |    0    0    1    0    0    0    1
 11     2 +0.664 +0.000 +0.105 +0.002 0.000 1.000 |    0    2    0    0    0    0    0
 12     2 +0.572 +0.000 +0.061 +0.023 0.356 0.500 |    0    0    0    1    0    1    0
 13     2 +0.631 +0.000 +0.127 +0.031 0.000 1.000 |    0    0    2    0    0    0    0
 14     2 +0.556 +0.000 +0.062 +0.013 0.000 1.000 |    0    2    0    0    0    0    0
 15     2 +0.574 +0.004 +0.083 +0.007 0.356 0.500 |    1    1    0    0    0    0    0
-------------------------------------------------------------------------------
```

As shown in Figure 2 and Figure 3, the average value of *Entropy* = 0.297 and *Purity* = 0.667 for the courses' clusters and the average value of *Entropy* = 0.262 and *Purity* = 0.656 for teachers' clusters which shows that reasonable quality clusters for k = 16 were generated. *Entrpy* and *Purty* show the computed values of *Entropy* and *Purity* respectively of each cluster for courses' (as shown in Figure 2) and teachers' (as shown in Figure 3) datasets. *Purty* close to 1 and *Entrpy* close to 0 shows that the labels assigned manually and those assigned by the *vcluster* program of CLUTO are the same, which implies that the actual and the expected clustering solutions are similar. Similarly, the value of *Purty* close to 0 and *Entrpy* close to 1 indicates that the actual and the expected clustering solutions are somehow different from each other (Karypis, 2003). *SE* (Software Engineering), *MWAD* (Mobile and Web Application Development), *CS* (Computer Science (Programming)), *ML* (Machine Learning / Artificial Intelligence), *Netw* (Networking), *Secu* (Security), and *Data* (Database) were the labels assigned originally to course outlines and faculty members. The value against each label shows the number of course outlines (Figure 2) and faculty members (Figure 3) present in a cluster assigned with the respective label. For example, in Figure 2, cluster 0 has two objects (course outlines) assigned with the label *Data*. The values of *Entrpy* and *Purty* clearly depict that the actual as well as the expected clustering solutions are the same. Similarly, as shown in Figure 3, cluster 14 has two objects (faculty members) assigned with the label *CS*. The values of *Entrpy* and *Purty* clearly depict that the actual as well as the expected clustering solutions are the same.

Table 2

Top Discriminating Terms of Sample Teachers' Clusters and the Teachers Included in Those Clusters

T-cid	Top Discriminating Terms (Teachers)	Teachers' IDs
3	Object, Oriented, Fundamentals, Algorithm, Analysis	3,17,22
5	Security, Algorithm, Requirement, Quality, Mining	8,15,21
6	Mobile, Design, System, Information, Security, Requirement	16,18,24
8	Fundamentals, Advance, Designs, Administration, Spatial	12,10,29
15	Human, Image, Vision, Web, Compiler	31,20

Table 3

Top Discriminating Terms of Sample Clusters of Courses and the Courses Included in Those Clusters

C-cid	Top Discriminating Terms (Courses)	Courses' IDs
3	Intelligence, Language, Chomsky, Refactoring, Database	31,38,40
6	Consuming, Mobile, Language, Security, System	20,32,41
8	Configuration, Testing, Evolution, Identification, Geography	17,28,30
12	Processing, VLSI, Automata, Electronics, Context	15,25,6
14	Heuristics, Quality, Recurrence, Evaluation, Requirement	18,19,27,29

As discussed in Section 3, CLUTO also reported a list of discriminating and descriptive terms of each cluster. Based on the clustering solutions obtained for courses and teachers, the next step was mapping the clusters of teachers onto the clusters of courses for the generation of teachers' course preferences. This mapping was done on the basis of similar discriminating terms present in each cluster of courses and teachers. Table 2 shows some sample teachers' clusters IDs (T-cid) and the top discriminating terms present in the respective clusters of teachers along with their ID information.

Similarly, Table 3 shows some sample courses' clusters IDs (C-cid) and the top discriminating terms present in the respective clusters of courses along with the courses' ID information.

Mapping of Teachers' and Courses' Clusters to Generate Preferences

Automatic generation of course preferences of teachers was done on the basis of discriminant features identified in the clusters of teachers as well as courses. Mapping was done based on at least two or more discriminant features from both types of clusters (teachers and courses). This ensured the relevance of teachers with their respective courses and vice versa. Table 4 shows the common discriminating terms from both types of clusters along with the T-cid, C-cid, teachers' ID and courses' ID information.

Table 4

Mapping of Teachers Courses on the Basis of Common Discriminating Terms from Teachers' and Courses' Clusters

T-cid	C-cid	Teachers' ID	Courses' ID	Common Discriminating Terms from T-cid and C-cid
0	12	1	15,25,6	Processing, VLSI
1	7	26	26,10,13	Object, Analysis
2	13	5	3,33,36	Operations, Mining
3	7	3,17,	26,10,13	Analysis, Algorithm, Object, Oriented
4	7	2,19	26,10,13	Stacks, Algorithm
5	14	8,15,21	18,19,27,29	Quality, Requirement
6	6	16,18,24	20,32,41	Security, System, Mobile
7	12	11,13	15,25,6	Automata, Electronics
8	4	12,10,29	12,11	Administration, Spatial
9	13	23,28	3,33,36	Natural, Assembly
10	5	4,32	8,34,35	Testing, Software
11	0	6,9	24,42	Refactoring, Security
12	6	25,30	20,32,41	Mobile, Security
13	1	14,27	1,16	Network, Management
14	12	7	15,25,6	Processing, VLSI
15	9	31,20	5,9,14	Image, Vision

The mapping of courses onto teachers shown in Table 4 depicts the automatic generation of course preferences of teachers. As an example from Table 4, T-cid 3 is mapped on C-cid 7 on the basis of common discriminating terms including *analysis, algorithm, object,* and *oriented.* Thus, the courses having course IDs 26, 10 and 13 are the automated course preferences of teachers having teacher IDs 3 and 17.

Evaluation and Discussion

Based on the generated course preferences for teachers, we made a comparison of the actual course preferences given by teachers in the previous year (2018) with the preferences generated automatically (shown in Table 4). Table 5 shows the overall research findings and results. For validation, we computed the Precision (P) and Recall (R) for each TID as well as Average Precision, Average Recall and F-measure of the results with the help of the following equations (**(1) – (5)**).

P = Right automated course preferences of TID / Total automated course preferences of TID **(1)**

Average Precision = ∑P / Frequency of teachers **(2)**

R = Right automated course preferences of TID / Total course preferences of 2018 of TID **(3)**

Average Recall = ∑R / Frequency of teachers **(4)**

FM = 2 * ((P * R) / (P + R**))** **(5)**

On the basis of the computed values of Average Precision (0.72), Average Recall (0.67), and F-measure (0.69), we can conclude that automated preferences are in line with the preferences originally provided by teachers to a reasonable extent.

Applying the proposed methodology, we successfully generated automated course preferences of each teacher present in the dataset using document clustering. In addition to the generated course preferences, the mapping of teachers onto courses shown in Table 4 and the statistics presented in Table 5 can be used to recommend new but relevant courses to each faculty member as well. For example, as shown in Table 5, TID 1 provided a list of four preferred courses in 2018. Applying the proposed methodology, we generated four course preferences, out of which three were in accordance with old preferences, whereas the remaining one was actually a new but relevant course that could be recommended to TID 1. The respective teacher can prepare this course for future offerings. This will increase the number of courses a teacher can teach and it helps the department to augment its resources.

TID shows teachers' ID, AP shows the number of actual course preferences of 2018, PP shows the predicted course preferences, MP shows the number of matching course preferences, P stands for the Precision value and R stands for the Recall value.

Table 5

Comparison of the Predicted Course Preferences with Old Course Preferences Given by Teachers in 2018

TID	AP	PP	MP	P	R	TID	AP	PP	MP	P	R
1	4	4	3	0.75	1	17	6	4	4	1	0.66
2	3	3	2	0.66	1	18	8	4	4	1	0.375
3	5	5	3	0.66	0.8	19	7	5	5	1	0.714
4	3	3	2	0.66	1	20	8	5	4	0.8	0.625
5	6	3	3	1	0.5	21	6	3	2	0.66	0.5
6	4	4	3	0.75	1	22	5	3	1	0.33	0.6
7	3	3	2	0.75	1	23	4	3	1	0.33	0.75
8	3	4	3	0.75	1.33	24	6	4	3	0.75	0.66
9	5	4	3	0.75	0.8	25	8	4	3	0.75	0.5
10	5	4	2	0.5	0.8	26	8	5	4	0.83	0.625
11	7	4	2	0.66	0.57	27	7	2	1	0.66	0.285
12	3	3	3	1	1	28	6	3	3	1	0.5
13	4	3	2	0.83	0.75	29	5	3	3	1	0.6
14	8	3	2	0.66	0.37	30	4	4	3	0.75	1
15	4	4	3	0.66	1	31	3	3	2	0.66	1
16	6	4	4	1	0.5	32	5	4	4	1	0.8

Average Precision = 0.72
Average Recall = 0.67
F - Measure = 0.69

The application of our research is not limited to CS / SE / IT fields. The designed methodology can be applied to any field of education in Higher Education Institutes (HEIs). For the sake of experimentation only, we took into consideration the data of SE faculty members.

Conclusion and Future Work

In this research work, automatic generation of course preferences was done for teachers based on courses taught by them in previous years and course preferences given by them. As a sample for testing, the data of teachers from the Software Engineering department of UMT was collected to prepare the teachers' dataset. Likewise, the data of 45 distinct HEC (Pakistan) approved course outlines was taken into account for the preparation of the courses' dataset. Document clustering using CLUTO was applied to the preprocessed datasets of courses and teachers. For k = 16, we obtained reasonable quality clusters against both teachers' and courses' datasets. For courses' clustering, we achieved an average *Entropy* of 0.297 and an average *Purity* of 0.667, respectively. Similarly, for teachers' clustering we achieved an average *Entropy* of 0.262 and an average *Purity* of 0.656, respectively. We also compared the old preferences of courses with the automated ones and obtained the average Recall value of 0.72, the average Precision value of 0.67 and F-measure of 0.69, respectively. In addition to the generation of automated course preferences, we also recommended relevant new courses to faculty members, which they may prepare for future.

It can be concluded that automated assignment is reasonably accurate and is in accordance with teachers' skills and course preferences. A teacher's expertise and preference of courses have a strong impact on delivering quality education to students. Hence, the proposed methodology can help the higher authorities of HEIs in course assignment to faculty members.

Future researchers can perform course allocation by adding the feedback given by students to each teacher regarding the previously assigned courses. We are planning to design and develop a working system that performs automatic allocation of courses to respective teachers based on their preferences of time and subjects keeping in view their skills and feedback given to them by students.

References

Abualigah, L. M., Khader, A. T., & Al-Betar, M. A. (2016). Multi-objectives-based text clustering technique using K-mean algorithm. Paper presented at *7th International Conference on Computer Science and Information Technology* (CSIT) (p. 1–6). IEEE.

Dařena, F., & Přichystal, J. (2018). Analysis of the association between topics in online documents and stock price movements. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 66*(6), 1431–1439.

Dhanalakshmi, V., Bino, D., & Saravanan, A. M. (2016). Opinion mining from student feedback data using supervised learning algorithms. Paper presented at *3rd MEC International Conference on Big Data and Smart City* (ICBDSC) (p. 1–5). IEEE.

Ganguli, R., & Roy, S. (2017). A study on course timetable scheduling using graph coloring approach. *International Journal of Computational and Applied Mathematics, 12*(2), 469–485.

Hassan, M. T., Karim, A., Kim, J.-B., & Jeon, M. (2015). Cdim: Document clustering by discrimination information maximization. *Information Sciences, 316,* 87–106.

Jensi, R., & Jiji, D. G. W. (2014). A survey on optimization approaches to text document clustering. *ArXiv Preprint, 1401*, 2229.

Karimpour, J., & Mavizi, S. (2016). Using k-means clustering algorithm for common lecturers timetabling among departments. *Advances in Computer Science: An International Journal, 5*(1), 86–102.

Karypis, G. (2003). *CLUTO: A clustering toolkit*. Minneapolis, MN: University of Minnesota.

Kaur, E. J., & Kaur, A. (2014). Timetable scheduling using modified clustering. *International Journal of Research in Information Technology (IJRIT), 2*(7), 1–8.

Liu, Y., WANG, X., XU, Z., & Guan, Y. (2006). A survey of document clustering [J]. *Journal of Chinese Information Processing, 20*(3), 55–62.

Pal, A. K., & Pal, S. (2013). Evaluation of teacher's performance: A data mining approach. *International Journal of Computer Science and Mobile Computing, 2*(12), 359–369.

Singh, S. (2017). Timetable generation using value encoding and different selection methods in genetic algorithm. *SRMS Journal of Mathematical Sciences, 1*(1), 90–102.