An Intelligent Student Advising System Using Collaborative Filtering

Kathiravelu Ganeshan
Department of Computing
Unitec Institute of Technology
Auckland, New Zealand
kghaneshan@unitec.ac.nz

Xiaosong Li
Department of Computing
Unitec Institute of Technology
Auckland, New Zealand
xli@unitec.ac.nz

Abstract—We propose a web based intelligent student advising system using collaborative filtering, a technique commonly used in recommendation systems assuming that users with similar characteristics and behaviors will have similar preferences. With our advising system, students are sorted into groups and given advice based on their similarities to the groups. If a student is determined to be similar to a group students, a course preferred by that group might be recommended to the student.

K-means algorithm has been used to determine the similarity of the students. This is an extremely efficient and simple algorithm for clustering analysis and widely used in data mining. Given a value of K, the algorithm partitions a data set into K clusters.

Seven experiments on the whole data set and ten experiments on the training data set and testing data set were conducted. A descriptive analysis was performed on the experiment results. Based on these results, K=7 was identified as the most informative and effective value for the K-means algorithm used in this system. The high performance, merit performance and low performance student groups were identified with the help of the clusters generated by the K-means algorithm. Future work will make use of a two-phase approach using Cobweb to produce a balanced tree with sub-clusters at the leaves as in [11], and then applying K-means to the resulting sub-clusters. Possible improvements for the student model were identified. Limitation of this research is discussed.

Keywords—K-means; clustering; collaborative filtering; rules; intelligent academic advising system; course.

I. INTRODUCTION

In tertiary institutions, online academic advising systems can provide prompt advice as and when required, and thus enhance student experience and save staff time and other institutional resources. Therefore such systems are gaining popularity. Research into such systems and development of such systems are in progress. In [1], the authors describe a Web-Based Decision Support Tool for Academic Advising. In their study, 90% of users, consisting of 20 undergraduate students and 5 faculty members, found their system effective and efficient. They consider four models of advising: prescriptive, developmental, integrated and engagement. They also discuss the importance of making systems that are more than data repositories and including more intelligence so the systems are able to provide reliable advice that students do not have to check back with human advisors. In other words, systems that provide reliable advice.

The advice given by academic advising systems will vary, from student to student, depending on a number of factors such as the academic performance of the student concerned and the major area of study and also may be impacted by other factors such as the nationality, age and gender of the students. There are also academic regulations and rules, which can be applied to provide simple answers to basic questions with the assistance of normal SQL queries.

Goals for academic advising include development of suitable educational plans, selection of appropriate courses, interpretation of institutional requirements, enhancement of student awareness about available educational resources, evaluation of student progress toward established goals and development of decision making skills with reinforcement of student self-direction [1]. Others have used such systems to verify their tracks towards their degree program.

We propose a web based intelligent student advising system using collaborative filtering, a technique commonly used in recommendation systems. This technique assumes that users with similar characteristics and behaviors will have similar preferences [3, 4]. With our advising system, students are sorted into groups and given advice taking into account the relevant factors and also considering their similarities to specific groups. A major use of the online advising system we have proposed and prototyped is to help students choose courses from over fifty courses and five interlinked pathways in the Bachelor of Computing (BCS) program. If a student belongs to a particular group, a course that other students in that group have preferred or performed well in, may be recommended to the student.

The system is developed to be integrated into our current student management system, PeopleSoft. Therefore, our students don’t need to create a profile to use this system. Real student data with complete records for the last four years (2011 to 2014) of all 743 students enrolled in over 50 courses in the Bachelor of Computing Systems (BCS) was anonymized and used in training and testing the prototype. Data included academic transcripts as well as biographic data.
K-means algorithm is used to cluster the 743 students into a number of clusters. This is an extremely efficient and simple algorithm for cluster analysis, widely used in data mining. Given a value of K, the algorithm partitions a data set into K clusters [5, 6 and 7].

K-means was chosen at this stage, instead of other clustering techniques, due to its ease of use and fitness for purpose. In the next stage of this project, we will have more knowledge on the data. Cobweb will be used to produce a balanced tree with sub-clusters at the leaves as in [11], before applying K-means to the resulting sub-clusters.

A prototype of the proposed system has been developed for concept approval. This prototype is implemented in ASP.NET and MS SQL database server. The system consists of two main components: a simple question and answer (QA) component and an intelligent advising component (IA) [8].

The QA component maintains a list of frequently asked questions (FAQs) of undergraduate students. A number of academic rules have been implemented, so it can answer routine questions like “am I eligible to take course X next semester?” This component is aimed at helping students in the development of suitable educational plans, selection of appropriate courses, interpretation of institutional requirements and increasing students’ awareness of the multitude of educational and other (such as counselling) resources available to students.

The IA component uses records from our current student management system. This allows the two systems to be integrated easily. There is no need for a student to create a new profile; instead, advice will be given taking into relevant, current student data. Academic and biographic information relating to the student concerned, is extracted from the student management system and used to identify which predefined group the student belongs to, and advice given taking into consideration, among other things, the specific characteristics of that group. It attempts to answer questions like “What courses should I take next semester?” This component is aimed at helping students in the selection of appropriate courses and development of decision making skills with reinforcement of student self-direction. K-means algorithm has been used to generate student groups.

A testing and experiment (TE) component is added for data gathering and experiments. The TE component helps to understand the characteristics of the different student groups and to select a suitable K value for K-means algorithm.

743 Bachelor of Computing student records from 2011 to 2014 were used as the data sample. From these records, ethnicity, age, gender, GPA and courses the student had taken were extracted. Based on these data, we gained a general understanding about the characteristics, performance and the interests for different student groups. The system is trained accordingly to provide useful and helpful advice.

In the rest of this paper, the overview of the system is provided first, the experiments related to K means are then discussed, the testing results and future improvements are discussed after that, and a summary is given at last.
B. The intelligent advising (IA) component

As the records in our current student management system are used, the information can be extracted from the data is limited, thus the attributes can be used to model a student are limited. The students are modelled by using the existing data as closely as possible. The advising system is expected to help improve the students' performance. Much of existing research around the world use GPA to measure student performance [8]. Research also shows that students’ performance depends on many factors such as gender, age, student's competence in English and even nationality and ethnicity. Some other factors that have a significant impact on student performance have been identified as students’ communication skills, learning facilities, proper guidance and family commitments related stress [9].

According to [9], students’ academic accomplishments and activities, perceptions of their coping strategies and positive attributions, and background characteristics, including family income, parents’ level of education, guidance from parents and number of negative situations in the home, were indirectly related to their composite scores, through academic achievement.

According to [9], the student performance should be improved if the administration of the college provides proper learning facilities; the student performance should be improved if the students have good and effective communication skills and have good competence in English; the student should perform well if they are properly guided by their parents and also by their teacher.

Based on the available data, previous research findings and our experience teaching into as well as leading (one of the authors) the program over the past 13 plus years, we loosely related and defined, as a first approximation the following attributes.

- GPA: Relevant to performance.
- Age: Relevant to family stress, e.g. mature students are more likely to have family commitment.
- Ethnicity: Relevant to English competency and family background.
- Gender: Relevant to learning style and how they can cope with the provided learning facilities.

Based on the above attributes, the samples were partitioned using K-means algorithm. Seven experiments were conducted on the 743 records for k=2, 3, 4, 5, 6, 7 and 8. A descriptive analysis on the clusters was done to select a suitable K value and to find the major features for different student groups. The top twelve courses taken by the students in each cluster were also identified. The data was then split into two data sets: training (372 records) and testing data set (371 records), to test the performance unsupervised learning [6].

To determine the K value, TE component provides an environment (Fig. 4) which allows the researchers to experiment with seven K values (2, 3, 4, 5, 6, 7 and 8) on the whole sample data set (743 records). The results are displayed in a combination of text and graphic format. For a given K value, the graph visualize the attribute (GPA, age, ethnicity and gender) distribution and the course selection distribution for four majors (Software Development, Networking and Security, Business Intelligence and Other) for each cluster. The text results provide more detailed information. The top 12 most popular courses for each cluster have also been identified (Fig. 5).

Similar experiment environments are provided for both of the training data set (372 records) and testing data set (371 records), which allow the researchers to experiment with ten K values (2, 3, 4, 5, 6, 7, 8, 9, 10 and 15).

III. THE K-MEANS EXPERIMENT

Let

\[ X = \{x_i\}, i = 1, ..., n \]  

be the set of n d-dimensional points to be clustered into a set of K clusters,

\[ C = \{c_k\}, i = 1, ..., k \]  

Where d=4, representing GPA, age, ethnicity and gender.

This project is supported by Unitec Foci research fund.

This project is supported by Unitec Foci research fund.
Let $GPA(c_k)$ represent the average GPA of cluster $c_k$; $Age(c_k)$ represent the average of cluster $c_k$. Let $m$ represent the $K$ value for a particular experiment. Let $SDTEV(GPA_m)$ represent standard deviation of \{GPA($c_1$), ..., GPA($c_m$)}; $SDTEV(AGE_m)$ represent standard deviation of \{Age($c_1$), ..., Age($c_m$)}). For $L$ experiments, $L$ standard deviation \{$SDTEV(GPA_1)$, ..., $SDTEV(GPA_L)$\} and \{$SDTEV(AGE_1)$, ..., $SDTEV(AGE_L)$\} will be obtained. The larger standard deviation should reflect the larger amount information provided by the clusters in the experiment.

A. The Selection of the $K$ Value

To determine the $K$ value, seven ($L=7$) experiments have been conducted on the whole sample data set, $n=743$, for $K=2$, 3, 4, 5, 6, 7 and 8. Seven standard deviation \{$SDTEV(GPA_1)$, ..., $SDTEV(GPA_7)$\} and \{$SDTEV(AGE_1)$, ..., $SDTEV(AGE_7)$\} have been obtained. These values are depicted in Fig. 6 to identify a suitable $K$ value.
It can be clearly observed that when $K=7$, both standard deviation for GPA average and age average are much larger than the other $K$ values. This suggests that the clusters obtained from $K=7$ provide more information than the clusters obtained from other $K$ values. A close look at the individual clusters produced in this experiment, clear characteristics can be identified for each cluster. Cluster 1 consists of all the young female students; cluster 2 consists all the young male students with very high GPA and most of them are of Asian background; cluster 3 consists of all the young male students with very low GPA and most of them are of European background; cluster 4 consists of all the young male students with high GPA and most of them are of European background; cluster 5 consists of all the male students with low GPA and mixed ethnicities; cluster 6 consists of all the middle aged students with high GPA and of mixed ethnicities; cluster 7 consists all the male students with merit GPA and of mixed ethnicities. Therefore, the clusters generated by the K means algorithm on the whole data set are informative and effective.

B. The Training Data

To verify the value $K=7$ obtained from the above seven ($L=7$) experiments on the whole sample data set, the data was split into two data sets: training data (372 records) and testing data (371 records). Ten ($L=10$) experiments on the training data set ($n=372$) were conducted for $K=2$, $3$, $4$, $5$, $6$, $7$, $8$, $9$, $10$ and $15$. Ten standard deviation {SDTEV(GPA1), ..., SDTEV(GPA10)} and {SDTEV(AGE1), ..., SDTEV(AGE10)} have been obtained. These values are depicted in Fig. 7 to verify the $K$ value.

It can be clearly observed that when $K=5$ and $K=7$, both standard deviation for GPA average and age average are much larger than the other $K$ values. The standard deviation for ethnicity does not make too much difference. This suggests that the clusters obtained from both $K=5$ and $K=7$ provide more information than the clusters obtained from other $K$ values. Which value to take?

C. The Testing Data

As K means is unsupervised learning, there are no predefined labels for the clusters. We would like to check how meaningful it is if we use the course related information obtained from the training data to provide advice for our students.

The training data and the testing data are two data sets randomly and equally split from the whole data set. It is fair to assume that these two data sets should have similar distribution. We classify each testing data record by using the training data clusters, if the resulting testing data course distribution is close to the training data course distribution, then this approach is promising.

Fig. 8 shows the standard deviations of GPA, age, ethnicity and their sum for the clusters generated from the testing result. The standard deviation distribution of the testing result is quite similar to the training data except that there is a slight difference when $K=5$. However, it can be clearly observed that the standard deviation distribution of the testing result is very similar to the whole data set, in particular when $K=7$, both standard deviation for GPA average and age average are much larger than the other $K$ values. The whole data set size is larger than the training data set size; it should be closer to the actual data. So this approach is promising and $K=7$ is the most suitable $K$ value for this system.

D. The comparison of the Training Data and Testing Result

Fig. 9 shows the training data course distribution (percentage) against the clusters and the testing result course distribution against the clusters when $K=7$ for Software Development major, Fig. 10 shows the same data for Network and Security major, Fig. 11 shows the same data for Business Intelligence major, Fig. 12 shows the same data for the other majors.
It can be observed that the course distributions in the two cases are quite close for all the majors in general, except for cluster 3; there is a big difference (around 10%) for all the majors. Further investigation is required to find the reason for this and the way to improve this. It is still meaningful to use the training data to predict the preference of the testing data.

**E. The procedure to provide recommendations**

Given a student record $x_i$, we could provide the study pathway the student could take or the courses the student could take for next semester. The following is the procedure to follow in providing advices in the IA component.

1) Generate clusters $C = \{c_k\}, i = 1, \ldots, k$ by using the $K$ means algorithm on the whole data set, where $k=7$.
2) Identify which cluster $x_i$ belongs to, say $c_m$, where $7 \geq m \geq 1$.
3) Find out the top 12 most popular courses in $c_m$, eliminate those $x_i$ has taken, recommend the rest to $x_i$.
4) Calculate all the average marks for all the courses taken by the students in cluster $c_m$, eliminate those $x_i$ has taken, recommend five courses with the highest average marks.
5) Recommend the most popular pathway (major) in cluster $c_m$ to $x_i$. This is particular useful to new student or for a student who wish to change major.

**IV. DISCUSSION**

We have a discussion on all the clusters when $K=7$ for the whole data set (Fig. 13).

For $K=7$, if we classify the clusters according to the average GPA, we can have three levels. High performance level, GPA > 5, including cluster 2, 4 and 6 (male, European + Maori + Asian, could be mature students); merit performance level, 4 > GPA > 2, including cluster 1 and 7 (female or male, aged around 25.5, mixed ethnicities); and low performance level, 2 > GPA, including cluster 3 and 5 (male, aged around 24.2, mixed ethnicities).

An interesting fact is that cluster 3 and 4 have very similar characteristics except that the average GPA for cluster 3 is 1.88 and the average GPA for cluster 4 is 5.81. What could make this difference? By looking at the clustering statistics by major, it is noted that 43.27% courses taken by the students in cluster 3 are network and security courses, only 30.51% courses taken by these students are software development courses. On the other hand, only 34.53% courses taken by the students in cluster 4 are network and security courses, and 39.97% courses taken by these students are software development courses.
The following is a comparison of the top 12 most popular courses for cluster 3 and cluster 4.

Excluding some basic courses such as “Operating System Fundamentals”, “Professional Skills” “Information Systems in Business” and “Project Planning and Control”, we have the following two lists:

Top most popular courses for cluster 3 are:
- Hardware Fundamentals
- Introduction to Databases
- Programming Fundamentals
- Networking Fundamentals
- Hardware Technology
- Multimedia and Web Development
- Information Gathering

Top most popular courses for cluster 4 are:
- Programming Fundamentals
- Hardware Fundamentals
- Introduction to Databases
- Multimedia and Web Development
- Networking Fundamentals

From the above two lists we can see that cluster 3 has four network & security courses and three software development courses; cluster 4 has three network & security courses and four software development courses. And also, in cluster 3, the network & security courses have higher priority and in cluster 4, the software courses have higher priority. This suggests that the low performance students prefer network & security courses, while the high performance students prefer software development courses.

Examining all the clusters in the high performance level (cluster 2, 4 and 6), software development courses are much more popular than network and security courses in cluster 2 and 4, cluster 6 is an exception, however the student average age is 46.8 which is also an exception.

Examining all the clusters in the low performance level (cluster 3 and 5), network and security courses are consistently (43.27% and 45.86%) popular while software development courses are consistently less popular (30.51% and 30.83%).

Cluster 5 and 7 are in the similar situation, however, the GPA difference between the two clusters is not that larger, 1.18 and 3.3. Still network courses are more popular in cluster 5 than cluster 7. On the other hand, business intelligence courses are more popular in cluster 7 than cluster 5.

The above also suggests that the low performance students are more likely in network major and high performance stu-
dent s are more likely in software major. So in addition to the
two attributes used modelling a student in this system, major is
another possible attribute.

V. SUMMARY

Based on our experiment data, it was identified that when
K=7, the clusters generated from the K means algorithm are
more informative and effective.

An outline of the procedure to provide recommendations in
the IA component was given. These recommendations can only
provide a rough guideline to the students for course selection
and major selection.

It is meaningful to use the training data to predict the prefe-
rances of the testing data; however, there is a big difference
(around 10%) in the course distribution for cluster 3 for all the
majors. Further investigation is required to find the reason for
this and the way to improve this.

The K-means experiment results on the whole data set
(K=7) suggested that male students aged between 24 and 27 in
the software major, mostly of European, Maori and Asian
backgrounds, as well as middle aged (around 45) students are
more likely to be high performance students. On the other
hand, young male students with network major are more likely
to be low performance students. Strategies on how to help low
performing students to improve their performance and the high
performing students to maintain their level of high performance
should be developed and integrated into the advice given by
this system.

The experiment results also suggested that major is a possi-
ble factor that is related to the students’ performance. In the
next version of the system, in addition to attributes (GPA, age,
etnicity and gender) in the current student model, major
should be added as another attribute in the student model.

Existing records in our current student management system
are used. While this facilitates the seamless integration between
the advising system and the current student management sys-
tem, the information that can be extracted from the data is lim-
ited, and therefore the attributes that can be used to model
students are also limited.

Another area to explore is how closely these results based
on the 743 student records represent the entire student popula-
tion.

Further work is planned using a two phase approach similar
to the work done by [11] combining the well-known Cobweb
algorithm with K-means. Cobweb will be used to produce a
balanced tree with sub-clusters at the leaves as in [11], and then
K-means applied to the resulting sub-clusters.

Future work will also involve taking learning styles into ac-
count as well as how best to incorporate advice regarding new
elective courses in to the system.

ACKNOWLEDGMENT

We would like to thank the “Student Answer” BCS project
team for the implementation of the initial prototype.

We would like to thank Thang Nguyen for revising the da-
tabase, entering all the testing data into the database and im-
plementing the K-means experiment component.

REFERENCES

information system for academic advising. Information Studies Journal
Retrieved from search.shamaa.org.

The 2011 World Congress in Computer Science, Computer Engineering,
on 28/1/2015.

and algorithms. IBM DeveloperWorks: 12 December 2013.

from: http://webwhompers.com/collaborative-filtering.html

http://www.ommyphd.com/?p=k-means.clustering#h3 BadExample

recognition letters, 31(8), 651-666.

(2013) Retrieved 23rd March 2014 from:
http://visualstudiomagazine.com/articles/2013/12/01/k-means-data-
clustering-using-c.aspx.


performance. Global Journal of Management and Business Research,
12(9).

behavior, perceptions and other factors affecting ACT scores. Iowa City,
Iowa: ACT, Inc.

Yi(Eds.), Proceedings of the 17th Australian Joint Conference on
Artificial Intelligence, Cairns, Australia, December 4-6, 2004. (pp. 368-
379). Berlin: Springer.