Eng. & Tech. Journal, Vol.30, No.10, 2012

Mining Tutors' Interesting Areas to Develop Researched Papers Using a Proposed Educational Data Mining System

Dr. Reem Jafar Ismail
Computer Science Department, University of Technology/ Baghdad
Email: reemaljanabi@yahoo.com

Received on: 24/10/2011 & Accepted on: 1/3/2012

ABSTRACT

Educational Data Mining (EDM) is the process of converting raw data from educational systems to useful information that can be used by educational software developers, students, teachers, parents, and other educational researchers. One of the difficulties in the educational institutes that face the tutor is how to write a paper. This work aims to help the tutor to write a researched paper on specific subject by finding another tutor who is also inter

ested in the same subject. This is done by exploring the tutor database by using the proposed educational data mining system, the tutor database is arranged in multidimensional form will include: tutor's teaching subjects, tutor's interesting areas, tutor's published researches, tutor's Msc. and Ph.D research subjects. The proposed system implements SMC and Cosine similarity measures with new proposed representation of tutor's database. A clustering K-Means techniques and associated rule generation is implemented by using WEKA data mining tool. The results obtained from that work are very useful for tutor and they give a rich analysis for developing researched papers for different tutors.

Keywords: Educational data mining, SMC and Cosine similarity measures, K- Means Clustering, WEKA data mining tool.

تنقيب اهتمامات التدريسيين لكتابة بحث باستخدام نظام مقترح في تنقيب البيانات التعليمي

الخـــلاصـــة

يقصد بتنقيب البيانات التعليمي هو عملية تحويل كم البيانات في انظمة التعليم الى معلومات مفيدة يمكن استخدامها من قبل مبرمجي البرامجيات التعليمية او الطلاب او التدريسيين او الوالدين او اي شخص تعليمي باحث. احدى الصعوبات التي تواجه التدريسي في المؤسسات التعليمية هو ايجاد موضوع لكتابة بحث من اجل الترقية العلمية في مجال التدريس لذا فان البحث المقترح يهدف الى مساعدة التدريسي لكتابة بحث في موضوع ما عن طريق ايجاد تدريسي اخر له اهتمام في نفس الموضوع وسيتم ذلك من خلال التنقيب في قاعدة البيانات التابعة الى التدريسيين والتي تحتوي على معلومات تشمل: موضوعات الدرس التي تم تدريسها من قبل التدريسي ،مجالات الاهتمامات البحثية للتدريسي ،البحوث التي تم نشرها من قبل التدريسي وكذلك موضوعات الماجستير والدكتوراه التابعة له. ان البحث المقترح استخدم صيغة جديدة لتمثيل البيانات في قاعدة البيانات وكذلك استخدم مقياسين للتشابه. لقد تم ترتيب بيانات التدريسيين ضمن K-Means

https://doi.org/10.30684/etj.30.10.5
2412-0758/University of Technology-Iraq, Baghdad, Iraq

Clustering و Associated rule generation باستخدام WEKA data mining.ان النتائج المستخلصة من هذا البحث مفيدة جدا للتدريسيين وتعطي تحليل مستفيض من اجل تطوير البحوث العلمية للتدريسيين.

INTRODUCTION

Data mining is data analysis methodology used to identify hidden pattern in a large data set. It has been successfully used in different areas including the educational environment. Educational data mining (EDM) is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding and improved educational performance [1].

Concern with EDM many studies are applied for analyzing the teaching performance of the tutors, these studies will improve teaching quality and help teachers improve their teaching effectiveness [2, 3]. This paper proposes EDM system to help the teacher to write a researched paper on specific subject by finding another teacher who is also interested in the same subject. EDM differs from knowledge discovery in other domains in several ways. One of them is the fact that it is difficult, to compare different methods or measures a posteriori and decide which is the best. It is therefore essential to use techniques and measurements that are fairly intuitive and easy to interpret [4].

One of the important measures in data mining is similarity measures; similarity between two objects is a numerical measure of the degree to which the two objects are alike. Consequently, similarities are higher for pair of objects that are more alike. Similarities are usually non-negative and are often between 0 (no similarity) and 1 (complete similarity) [5]. In the following subsection a two similarity measures are explained.

- **Simple Matching Coefficient (SMC) for binary data [5]**

Similarity measures between objects that contain only binary attributes are called **similarity coefficients**, and typically have values between 0 and 1. A value of 1 indicates that the two objects are completely similar, while a value of 0 indicates that the objects are not at all similar.

Let **x** and **y** be two objects that consist of n binary attributes. The comparison of two such objects, i.e., two binary vectors, leads to the following frequencies:

f_{00} = the number of attributes where **x** is 0 and **y** is 0
f_{01} = the number of attributes where **x** is 0 and **y** is 1
f_{10} = the number of attributes where **x** is 1 and **y** is 0
f_{01} = the number of attributes where **x** is 1 and **y** is 1

$$SMC = \frac{number\ of\ matching\ attribue\ values}{number\ of\ attributes} = \frac{f_{11}+f_{00}}{f_{01}+f_{10}+f_{11}+f_{00}} \quad \ldots (1)$$

COSINE SIMILARITY [4]

It is one of the most common measures of document similarity. Consider two vectors x and y and the angle they form when they are placed so that their tails coincide. When this angle near 0^0 ,then cosine near 1, i.e. the two vectors are very similar. When this angle is 90^0, the two vectors are perpendicular, the most dissimilar, and cosine is 0.

If x and y are two document vectors, then:

$$cos(x, y) = \frac{x.y}{\|x\| \|y\|} \quad \ldots (2)$$

Where. indicates the vector dot product, $x.y = \sum_{k=1}^{n} x_k \; y_k$, and $\|x\|$ is the length of vector x, $\|x\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{x.x}$

K-MEANS CLUSTERING AND WEKA DATA MINING TOOL

Clustering analysis aims to identify homogeneous objects into a set of groups, named clusters, by given criteria. Clustering is a very important technique of knowledge discovery for human beings. The grouped objects are called clusters, where the similarity of objects is high within clusters and low between clusters. To achieve different application purposes, a large number of clustering algorithms have been developed [6].

K-Means is one of the algorithms that solve the well known clustering problem. The algorithm classifies objects to a pre-defined number of clusters, which is given by the user (assume k clusters). The idea is to choose random cluster centers, one for each cluster. These centers are preferred to be as far as possible from each other. Starting points affect the clustering process and results. After that, each point will be taken into consideration to calculate similarity with all cluster centers through a distance measure, and it will be assigned to the most similar cluster, the nearest cluster center. When this assignment process is over, a new center will be calculated for each cluster using the points in it. For each cluster, the mean value will be calculated for the coordinates of all the points in that cluster and set as the coordinates of the new center. Once we have these knew centroids or center points, the assignment process must start over. As a result of this loop we may notice that the k centroids change their locations step by step until no more changes are made. When the centroids do not move any more or no more errors exist in the clusters, we call the clustering has reached a minima [7].

The Waikato Environment for Knowledge Analysis (WEKA) came about through the perceived need for a unified workbench that would allow researchers easy access to techniques in machine learning. Nowadays, WEKA is recognized as a landmark system in data mining and machine learning. It has achieved widespread acceptance within academia and business circles, and has become a widely used tool for data mining research [8].

WEKA has several graphical user interfaces that enable easy access to the underlying functionality, which includes algorithms for regression, classification, clustering, association rule mining and attribute selection [9].

THE PROPOSED SYSTEM

Data Collection and Preparation Stage

Table (1) summarizes the most important attributes for each teacher that supplies our objective. About 100 records are saved in the database of table (1) where each teacher has a name, subject teaching, interesting areas, published researched papers, M. Sc. and Ph. D fields.

Table (1) Real Information for each Teacher Collected from Teacher's C.V.

Teacher name	Tutor's subjects teaching	Tutor's interesting areas	Tutor's published researches	Tutor's M.Sc subject	Tutor's Ph.D subject
John	NLP, fuzzy logic, OS., computer networks	NLP, Internet, SSD	Information hiding	Information hiding	Computer network
Pop	computer networks, security, information hiding	NLP, computer network, SSD	Security	E-commerce	E-commerce
Lysa	Neural networks, compiler, image processing	Image processing	-	Web design	Web design
Adem	compiler, image processing, NLP	OS, compiler, Internet	Data mining, security	Networks	Security
. . .	. . .	. . .	. . .	. . .	. . .

We have preprocessed the data in order to transform them into a suitable format to be used by the proposed education data mining algorithms as explained in the following section.

Proposed Binary Representation

The system proposed that the field of interesting areas in the teachers' table are converted to binary vectors, where each teacher will have a binary vector with a stream of 1's and 0's (1=indicated that the subject is included in the interesting field and 0= otherwise) , here instead of applying sequential search techniques to

find specific subject on the interesting area field by matching all words that stand for computer science subjects a simple (logic binary **And** operation) is done between two binary vectors in order to find matching subjects. This binary representation will reduce the data size that are stored in the tables and also reduce the time that is required to find the results, so the results will be found faster since the compares will be reduced by using a binary logic **And** operation. This binary representation is also important in our analysis since the SMC measure is done between two binary vectors. Example (1) will explain how is the binary logic **And** operation done to find the interesting areas between teachers.

Example 1: Suppose we have the following vectors:

Names id vector:(1=john, 2=pop, 3=lysa, 4=adem, etc...)

Subjects vector: ("NLP", "fuzzy", "OS", "network", "security", "information hiding", "neural networks", "compiler", "internet", "web programming", "data structure", "programming", "image processing", "AI", "SSD", , etc…)

Binary vector of the interesting area for each teacher is
ID=1, int_area vector = ("100000001000001")
ID=2, int_area vector = ("100100000000001")
ID=3, int_area vector = ("000000000000100")
ID=4, int_area vector = ("001000011000000")

When binary **And** operation is done between ID=1 and ID=2 the results will be the binary vector ("100000000000001") this means that John and Pop can make a researched paper on "NLP" and "SSD".

The Algorithms of the Proposed System
Algorithm Description

The system is done by implementing six steps, which are:

Step 1: Enter the database for all teachers which include the fields that are explained in table (1).

Step 2: Convert the interesting areas filed to binary vectors as explained in the **proposed Algorithm (1)**.

Step 3:Proposed algorithms to find relationships in data mining by finding interesting subjects

Phase I: All interesting relationship subjects

Algorithm 2: Proposed algorithm to find the relationship between ID teacher and all the interesting areas in all other teachers

Phase II: Specific interesting relationship subjects

Algorithm 3: Proposed algorithm to find the relationship between id teacher and all the interesting areas in all other teachers for specific subject

Step 4: Proposed algorithms to find similarity measures in educational data mining

Phase I: is done by using a **SMC similarity** measure where each subject is considered as a single entity in the vector, so the vector will contain the subjects names.
Algorithm 4: SMC similarity measure of interesting area subjects in CS between specific teacher and all other teachers.

Example 2: SMC similarity measure.

Suppose we have the following two binary vectors of the interesting area for ID=1 teacher and ID=2 teacher:

ID=1, int_area vector1 = ("100000001000001")
ID=2, int_area vector2 = ("100100000000001")

All the frequencies: f_{00}, f_{01}, f_{10} and f_{11} are calculated between *int_area vector1* and *int_area vector2*, then after applying equation (1) the following results are obtained:

$$f_{00}=11, f_{01}=1, f_{10}=1, f_{11}=2$$

$$SMC = \frac{2 + 11}{1 + 1 + 2 + 11} = 0.8666$$

Therefore ID=1 teacher and ID=2 teacher have a SMC similarity equal to 0.8666

Phase II: is done by using **Cos similarity** measure where the related subjects are grouped with each other and consider each subject as an entity in that group, so the vector will contains the group name of main subject that the other subjects are within.
Algorithm 5: Cos similarity measure of interesting area subjects in CS between specific teacher and all other teachers.
Example 3: Cos similarity measure.

Suppose we have the following vectors:

Names id vector:(1=john, 2=pop, 3=lysa, 4=adem, etc...)

Group1: Adaptive :(GA, neural networks, fuzzy logic)
Group2: Programming: (C, Prolog, VB, HTML, Java, C++)
Group3: Internet: (networks, e-commerce, web programming, wireless net)

Group4: Security: (cipher, advanced security, information hiding, secure software design)

Subjects Group vector: (" Adaptive ", " Programming ", " Internet ", " Security ")
Id=1, int_area group vector = ("2423")
Id=2, int_area group vector = ("2622")
Id=3, int_area group vector = ("3423")
Id=4, int_area group vector = ("3634")

For id=1 the int_area group vector = ("2423") means:

- The int_area group vector = ("**2**423") for id=1 means that only two subjects out of 3 from the adaptive group is known by id=1.
- The int_area group vector = ("2**4**23") for id=1 means that four subjects out of 6 from the programming group is known by id=1.
- The int_area group vector = ("24**2**3") for id=1 means that only two subjects out of 4 from the internet group is known by id=1.
- The int_area group vector = ("242**3**") for id=1 means that three subjects out of 4 from the security group is known by id=1.

Suppose we have the following two vectors of the interesting area for ID=1 teacher and ID=2 teacher:

ID=1, int_area group vector1 = ("2423")
ID=2, int_area group vecto2r = ("2622")

Then after applying equation (2) the result that is obtained equal to 0.9547, Therefore ID=1 teacher and ID=2 teacher have Cos similarity equal to 0.9547

Step 5: Implementing K-Means clustering analysis: This is done by using WEKA data mining tool.

Step 6: Implementing Apriori association rule generation: This is done by using WEKA data mining tool.

DETAILED EXPLANATION OF THE PROPOSED ALGORITHMS

Algorithm 1: Proposed algorithms to convert interesting areas to binary vectors
Input: The database that have all teachers
Output: Binary vectors for each teacher which is represented by an array of 2D where: (rows= id no. for each teacher, columns= string of binary) for the interesting areasa subjects in CS

Begin
Initialize an array (names subj[15]) which contains all the interesting areas in SC
Subj[1]="NLP" Subj[2]="fuzzy" Subj[3]="OS"….. Subj[15]="SSD"

Open database of teachers
For i= 1 to #teachers
Scan DB in interesting area field and read the int_subject from it
For j= 1 to #subjects
If int_subject= subj[j] then
vector(id=i,j)=1
else
vector(id=i,j)=0
next j
next i
End

Algorithm 2:_Proposed algorithm to find the relationship between id teacher and all the interesting areas in all other teachers
Input: id for the teacher, binary vectors for all teachers
Output: CS subjects with id that have a relationship with the entered teacher id

Begin
Initialize an array (names subj[15]) which contains all the interesting areas in SC
Subj[1]="NLP" Subj[2]="fuzzy" Subj[3]="OS"….. Subj[15]="SSD"

Call algorithm 1
vector(1, 1) = "100000001000001"
vector(2, 1) = "100100000000001"
vector (3, 1) = "000000000000100"
vector (4, 1) = "001000011000000"….. vector (n, m) = "011100101000111"
Enter the id no. for a specific teacher- Read id
For j = 1 To # teachers
For i = 1 To #subjects
aa(i) = (Mid(vector(id, 1), i, 1)) **And** (Mid(vector(j, 1), i, 1))

```
        If aa(i) = 1 Then
        Print names(id) & "----->" & names(j) & " " & subj(i)
        End If
        Next i
        Next j
End
```

Algorithm 3:_Proposed algorithm to find the relationship between id teacher and all the interesting areas in all other teachers for specific subject
Input: id for the teacher, id for the specific subject, binary vectors for all teachers
Output: CS subjects with id that have a relationship with the entered teacher id and subject

Begin
Initialize an array (names subj[15]) which contains all the interesting areas in SC
Subj[1]="NLP" Subj[2]="fuzzy" Subj[3]="OS"….. Subj[15]="SSD"

Call algorithm 1

```
        vector(1, 1) = "100000001000001"
        vector(2, 1) = "100100000000001"
        vector (3, 1) = "000000000000100"
        vector (4, 1) = "001000011000000"….. vector (n, m) =
        "011100101000111"

        Enter the id no. for a specific teacher- Read id
        Enter the id no. for a specific subject- Read subjectID

        found = false
        For j = 1 To # teachers
        For i = 1 To #subjects
        aa(i) = (Mid(vector(id, 1), i, 1)) And (Mid(vector(j, 1), i, 1))
        If aa(i) = 1 and i=subjectID Then
        found = true
        Print names(id) & "----->" & names(j) & " " & subj(i)
        End If
        Next i
        Next j
        If found = false Then
        MsgBox "No Teacher Found !"
        End If
End
```

Algorithm 4: SMC similarity measure of interesting area subjects in CS between specific teacher and all other teachers

Input: id for the teacher, vector of teacher names, vector of interesting area subjects in CS, binary vectors for all teachers that have their interesting areas
Output: SMC similarity measure for the id teacher with all other teachers
Begin

```
Enter the id no. for a specific teacher- Read id
For j = 1 To #teachers
sum01 = 0
sum10 = 0
sum00 = 0
sum11 = 0
For i = 1 To #subjects
 x = Mid(vector(id, 1), i, 1)
 y = Mid(vector (j, 1), i, 1)

 If x = 0 And y = 1 Then
  sum01 = sum01 + 1
  ElseIf x = 1 And y = 0 Then
  sum10 = sum10 + 1
  ElseIf x = 0 And y = 0 Then
  sum00 = sum00 + 1
  ElseIf x = 1 And y = 1 Then
  sum11 = sum11 + 1
  End If
Next i
            SMC = (sum11 + sum00) / (sum01 + sum10 + sum11 + sum00)
print names(id) & "----->" & names(j) & " " & SMC
```

End

Algorithm 5: Cos similarity measure of interesting area subjects in CS between specific teacher and all other teachers
Input: id for the teacher, vector of teacher names, integer vectors for all teachers that have their interesting areas
Output: Cos similarity measure for the id teacher with all other teachers
Begin

```
vectorG(1, 1) = "2423"
vectorG (2, 1) = "2622"
vectorG (3, 1) = "3423"
vectorG (4, 1) = "3634" ….. vectorG (n, m) = "1304"

Enter the id no. for a specific teacher- Read id
For j = 1 To #teachers
sum = 0

For i = 1 To #subjectGroup
```

```
x = Mid(vectorG(id, 1), i, 1)
y = Mid(vectorG(j, 1), i, 1)
sum = sum + (x * y)
Next i
sum1 = 0
For i = 1 To #subjectGroup
x = Mid(vectorG(id, 1), i, 1)
sum1 = sum1 + (x * x)
Next i
res1 = Sqr(sum1)

sum2 = 0
For i = 1 To #subjectGroup
y = Mid(vectorG(j, 1), i, 1)
sum2 = sum2 + (y * y)
Next i
res2 = Sqr(sum2)
res = sum / (res1 * res2)
print names(id) & "----->" & names(j) & " " & res
Next j
End
```

SYSTEM IMPLEMENTATION AND RESULTS

The system is implemented by using Visual Basic programming language. After collecting the teachers' survey and store them as tables like shown in table (1), the system automatically will convert the raw data of teacher interesting areas to binary vectors as explained in algorithm (2).

Figure (1) shows the main interface of the proposed system, when the user click the button: "Find teachers with same interesting area" the algorithm (3) is implemented by asking the user to enter his ID number as shown in figure (2) to find a relationship between this teacher and all other teachers in all interesting subjects the results in shown in figure (1) in the third column where all the teacher names and their subjects are shown together with the teacher ID name.

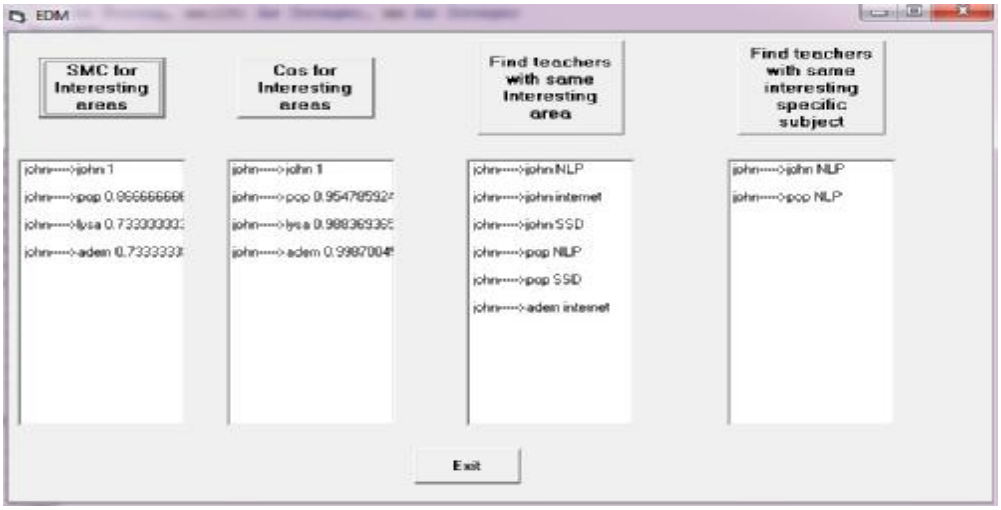


Figure (1) Main Interface of the Proposed System

When the user click the button: "Find teachers with same interesting specific subject" the algorithm (4) is implemented by asking the user to enter his ID number as shown in figure (2) and also enter the specific subject name to find a relationship between this teacher and all other teachers in that specific subject the results in shown in figure (1) in the forth column where all the teacher names with the specific subject are shown together with the teacher ID name.

Figure (2) Each Teacher will have ID Number to Enter it in that Input Box

In the analysis phase the SMC and Cos similarity measures are implemented between specific teacher and all other teachers as shown in Tables: (2), (3) and (4). For example in table (2) when the id=1, the SMC between id=1 and id=2 is equal 0.866 and SMC between id=1 and id=3 is equal 0.733 this will indicates that id=1 is more similar to interesting areas to id=2 than to id=3 because SMC (id=2) > SMC (id=3).

Table (2) SMC Similarity Measure between Interesting Areas of Teachers

id	1	2	3	4	…	100
1	**1**	**0.866**	**0.733**	**0.733**	**…**	**0.6**
2	**0.866**	**1**	**0.733**	**0.6**	**…**	**0.466**
3	**0.733**	**0.733**	**1**	**0.733**	**…**	**0.6**
4	**0.733**	**0.6**	**0.733**	**1**	**…**	**0.733**
…	**…**	**…**	**…**	**…**	**1**	**…**
100	**0.6**	**0.466**	**0.6**	**0.733**	**…**	**1**

Table (3) Cos Similarity Measure between Interesting Areas of Teachers

id	1	2	3	4	…	100
1	**1**	**0.954**	**0.988**	**0.998**	**…**	**0.880**
2	**0.954**	**1**	**0.936**	**0.966**	**…**	**0.730**
3	**0.988**	**0.936**	**1**	**0.988**	**…**	**0.872**
4	**0.998**	**0.966**	**0.988**	**1**	**…**	**0.869**
…	**…**	**…**	**…**	**…**	**1**	**…**
100	**0.880**	**0.730**	**0.872**	**0.869**	**…**	**1**

Table (4)Compares between SMC and Cos Similarity Measure for 4 Teachers out of 100

id	1		2		3		4	
Similarity	*SMC*	*Cos*	*SMC*	*Cos*	*SMC*	*Cos*	*SMC*	*Cos*
1	**1**		**0.866**	**0.954**	**0.733**	**0.988**	**0.733**	**0.998**
2	**0.866**	**0.954**	**1**		**0.733**	**0.936**	**0.6**	**0.966**
3	**0.733**	**0.988**	**0.733**	**0.936**	**1**		**0.733**	**0.988**
4	**0.733**	**0.998**	**0.6**	**0.966**	**0.733**	**0.988**	**1**	

After calculating the two similarity measures a K-means clustering analysis is done by implementing a WEKA 3.4.7 data mining tool. The interesting area fields of teachers that are explained in table (1) are input to the K-means clustering in WEKA data mining tool as shown in figure (3) where the information are stored in Microsoft Excel sheet. Figure (4) and (5) show the obtained results of WEKA clustering in K-Means method. In figure (4) the teachers are grouped in 3 clusters and in figure (5) the teachers are grouped in 4 clusters.

WEKA data mining tool is also implemented to extract association rules from interesting areas of teachers by using Apriori algorithm as shown in figure (6), where for example the first rule shows that the teacher who is not interested in web and image and interested in programming is also not interested in SSD.

A1 id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	NLP	fuzzy	OS	network	security	info	neural	compiler	internet	web	structure	programm	image	AI	SSD
2	ID12101	YES	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	YES
3	ID12102	YES	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES
4	ID12103	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO
5	ID12104	NO	NO	YES	NO	NO	NO	NO	YES	YES	NO	NO	NO	NO	NO	NO
6	ID12105	NO	YES	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	YES	NO	NO
7	ID12106	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	YES	NO	YES	YES	NO
8	ID12107	NO	NO	YES	NO	YES	YES	NO	NO	YES	NO	YES	NO	NO	YES	NO
9	ID12108	YES	YES	YES	NO	NO	NO	YES	YES	YES	NO	NO	YES	YES	YES	NO
10	ID12109	YES	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	YES	NO	NO	NO
11	ID12110	YES	YES	YES	NO	NO	NO	YES	YES	YES	NO	NO	YES	YES	YES	NO
12	ID12111	NO	YES	YES	NO	NO	NO	NO	YES	YES	NO	NO	NO	YES	YES	NO
13	ID12112	YES	YES	YES	YES	NO	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
14	ID12113	NO	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	NO	YES	YES	YES
15	ID12114	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
16	ID12115	NO	YES	YES	YES	NO	NO	NO	YES	YES	YES	NO	NO	YES	YES	YES
17	ID12116	YES	YES	YES	YES	NO	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
18	ID12117	NO	NO	NO	YES	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	YES
19	ID12118	NO	YES	NO	YES	NO	NO	NO	YES	NO	YES	NO	NO	YES	NO	YES
20	ID12119	NO	YES	NO	NO	YES	YES	NO	YES	NO	NO	YES	NO	YES	NO	NO
21	ID12120	YES	YES	YES	NO	NO	NO	YES	YES	YES	NO	NO	YES	YES	YES	NO
22	ID12121	NO	NO	NO	YES	YES	NO	NO	YES	NO	NO	YES	NO	NO	YES	YES
23	ID12122	YES	NO	YES	YES	YES	YES	NO	YES	YES	NO	NO	NO	NO	YES	YES
24	ID12123	NO	NO	NO	NO	YES	NO	NO	YES	YES	NO	NO	YES	YES	YES	YES
25	ID12124	NO	NO	NO	YES	NO	NO	YES	YES	YES	YES	NO	YES	NO	NO	YES

Figure (3) Microsoft Excel Sheet for the Interesting Area Field of Teachers

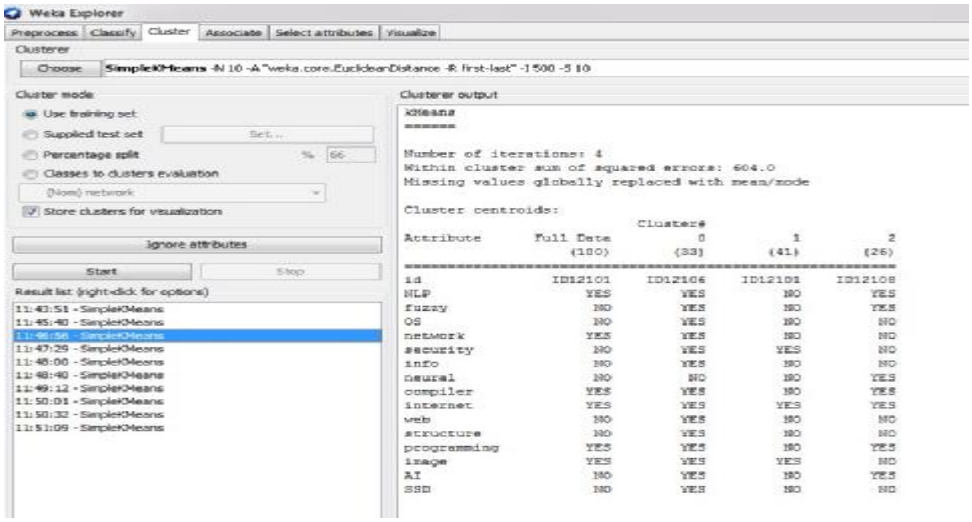


Figure (4) K-Means Clustering for Interesting Area of Teachers with 3 Groups

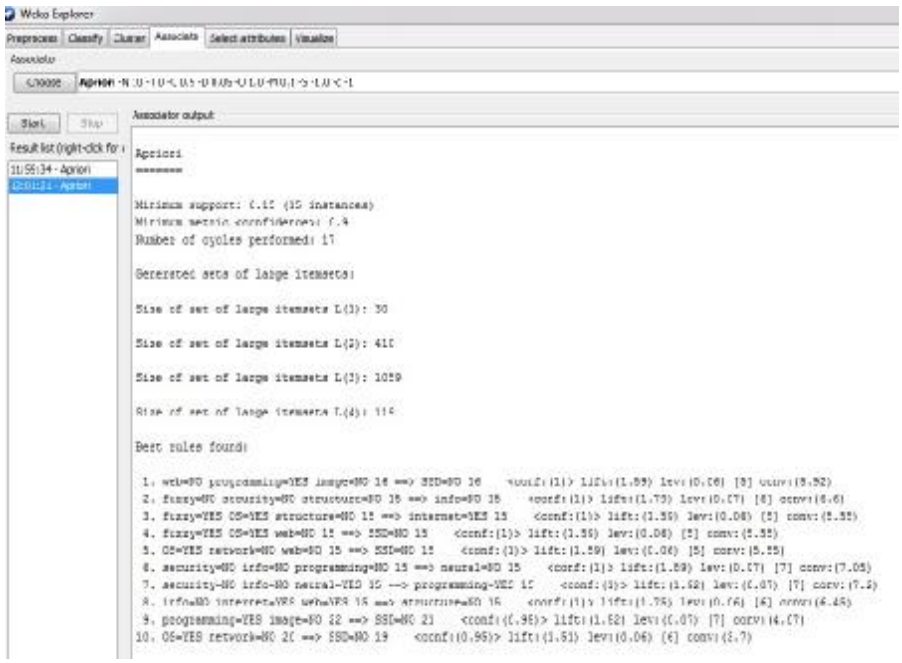


Figure (5) K-Means Clustering for Interesting Area of Teachers with 4 Groups

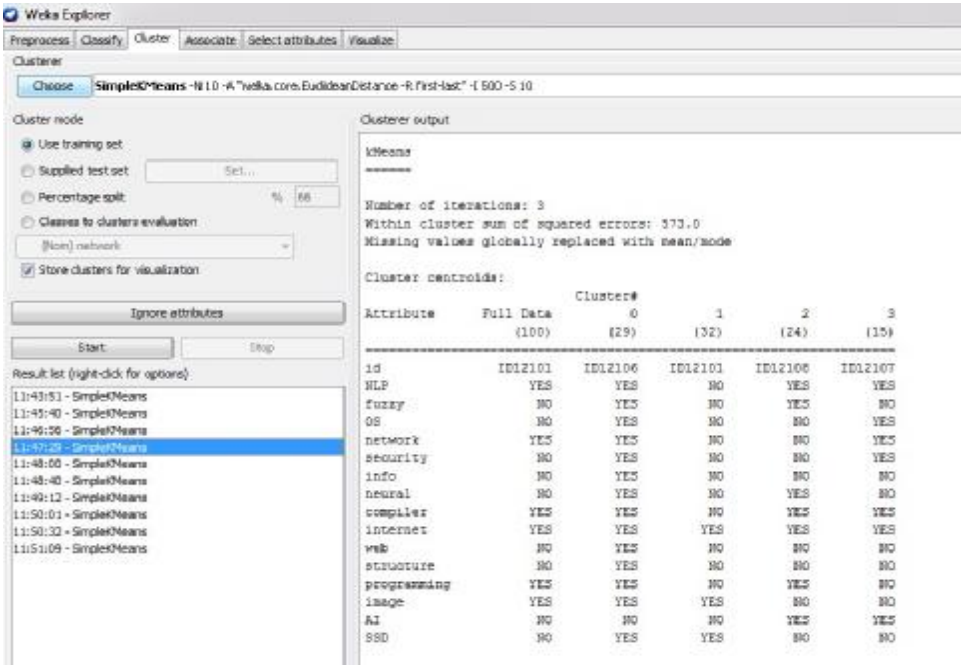
DISCUSSION AND CONCLUSIONS

This work has presented a solution to the tutor in writing a researched paper, the work can mine statically information and answer many educational questions, like: Which is the subject that is much interested by the teachers? How many teachers are interested in programming and wish to write a research on information hiding? and so on.

The proposed binary representation of the database for teachers will reduce the data size that are stored in the tables and also reduce the time that is required to find the results, so the results will be found faster since the compares will be reduced by using a binary logic **And** operation. Also, the proposed binary representation of the database for teachers is flexible in which a teacher can add or remove any subject from his own CV hence it is easy to update the binary vector of the interesting area fields by turning "0" to "1" when adding and "1" to "0" when removing.

The SMC and Cosine similarity measures are important in the proposed system to give an indication about the similarity between interesting areas of teachers the

more similarity value found means the best teacher to choose to develop a



researched paper in specific computer science subject. Two similarity measures are

Figure (6) Association Rules Extracted from Interesting Areas of Teachersby using Apriopi Algorithm in WEKA

implemented in the proposed system because when the SMC is failed to match teachers (means SMC = 0), Cosine similarity will find a teacher within the same group of interesting areas and not on specific subject.

The clustering implementation in the proposed system can create and organize a team work of programming teachers to develop a researched paper for a specific computer science subject or more since each cluster would have a specific property on specific subject features like security team work.

The analysis phase in K-Means clustering with the similarity measures that are applied on the same teacher database give a good relationships between teachers since the clustering only arrange the teachers in groups but with the similarity measures that are implemented in the proposed system a results will be more accurate.

In order to develop this system the database that is used in this project is for one department which is the computer science department, the same work can be implemented to include all the departments of the University of Technology by adding the database for each department and finding a new relational attributes between them.

Acknowledgement

This work has been implemented to the Department for Scientific Affairs and Higher Studies in Computer Science Department in the University of Technology. The author wishes to acknowledge the help provided by the teachers' staff of the Computer Science Department in the University of Technology in the data collection stage of the work by providing the needed information and teachers' survey that is used in this work.

REFERENCES

[1] Chandra E. and Nandhini K.; "Knowledge Mining from Student Data"; European Journalof Scientific Research, Vol.47, No.1, pp.156-163, 2010. http://www.eurojournals.com/ejsr.htm

[2] Sona M. and Bertan B.; "Analyzing Teaching Performance of Instructors using Data Mining Techniques"; Department of Management Information Systems, Informatics in Education, Vol.10, No.2, 245-257, 2011.

[3] Barracosa J. and Antunes C.; "Anticipating Teachers' Performance"; Department of Computer Science and Engineering, Technical University of Lisbon, Portugal, 2011.

[4] Agathe M. and Kalina Y.; "Interestingness Measures for Association Rules in Educational Data"; Proceedings of the 1st International Conference on Educational Data Mining, Canada, June 20-21, 2008.

[5 Pang-Ning T., Micheal S. and Vipin K; "Introduction to Data Mining"; Addison-Wesley, 2006 by Pearson Education, Inc.

[6] Zhang; K-Bing "Visual Cluster Analysis in Data Mining"; Department of Computing Division of Information and Communication Sciences, Australia, October, 2007.http://www.comp.mq.edu.au/hdr/current/kebing_thesis.pdf

[7]Bashar Al-Shboul, and Sung-Hyon Myaeng;" Initializing K-Means using Genetic Algorithms"; World Academy of Science, Engineering and Technology, 2009.http://www.waset.org/journals/waset/v54/v54-21.pdf

[8] Eibe F., Geoffrey H., Bernhard P., Peter R. and Ian H. Witten; "The WEKA Data Mining Software: An Update"; Published by ACM 2009 Article, Vol. 11, No. 1, June, 2009.http://www.kdd.org/explorations/issues/11-1-2009-07/p2V11n1.pdf

[9] The University of Waikato, Hamilton, New Zealand http://www.cs.waikato.ac.nz/~ml/weka/