طريقه تحليلية بعنقدة تنقيب البيانات لتقويم الاداء

Analytical Method by Clustering Data Mining for Performance Evaluation

المدرس غازي جوني Ghazi Johnny

جامعة بغداد/ الكلية التقنية الإدارية

الخلاصة

تنقيب البيانات يتعاملَ مع إكتشافِ المعرفةِ المخفيةِ، أنماط غير متوقّعة وقوانين(قواعد) جديدة مِنْ قاعدةِ البيانات الكبيرةِ، أساساً هو مُهتمّ بتحليلِ البياناتِ وإستعمالِ تقنياتِ البرامجِ لإيجاد الأنماطِ في مجموعاتِ البياناتِ، إحدى تقنياته الرئيسيةِ(العنقده)، والذي يعني عملية تَقسيم قاعدة بيانات إلى مجموعاتِ خاصه مشتركه(بخواص) تتقارب بالتشابه وتتباعد بالاختلاف بقدر الإمكان مِنْ أحدهما الآخر، حيث أنَّ المسافة تقاس حسب المتغيّرات المتوفرة.

تعتبر العنقده الاكثر استخداما في خوارزميات التنقيب عن البيانات ومفيده كتقنيه استطلاعيه وكروتين ثانوي للبحث عن اكثر البيانات تعقيدا وفي كثير من المجالات كاكتشاف القواعد ،الفهارس، التصنيف، الاختصار واكتشاف بعض الشاذ من البيانات ، في العقد الأخير بَدأتْ المؤسساتُ التربوية بتَطبيق معاييرِ الجوده والاعتماد، إحدى أهم المعاييرِ هوالتخطيط والتقويمُ، لذا كُلّ نشاطات المؤسساتِ التربويةِ يجب أنْ تُقومَ طبقاً لطرق علميه و منطقيه ، طرق العنقده طبقت في هذا البحث على حالات حقيقيه تتعلق بهيئه التعليم التقني لايجاد تقويم اداء منصف بين فروعها للعامين (٢٠٠٦-٢٠٠٩) للدورات التعيلميه المقدمه للمجتمع.

طريقه الارتباط الكامل اعطت تركيبه مثاليه و ذات معنى لطريقه تحليليه لتقويم اداء منصف يطابق معايير الجوده والاعتماد، بينما يَعتمدُ التقويمَ التقليديَ فقط على عددِ الدوراتِ.

Abstract

Data mining deals with the discovery of hidden knowledge, unexpected patterns and new rules from large database, Basically it is concerned with the analysis of data and the use of software techniques for finding patterns in sets of data, one of its main terms is the clustering, which is mean, the process of dividing a database into mutually exclusive groups such that the members of each group are as close as possible to one another , and different groups are as far as possible from one another , where distance is measured with respect to all available variables.

Clustering is perhaps the most frequently used data mining algorithm, being useful in its own right as an exploratory technique, and also as a subroutine in more complex data mining algorithms such as rule discovery, indexing, summarization, anomaly detection, and classification, In the last decade the educational institutions began to apply the standards of quality and accreditation, one of the most important standards is the(Planning and Evaluation) therefore; all the educational institutions activities must be evaluate according to logical and scientific methods, clustering methods have been applied in this research for a real cases concerned with Foundation Technical Institutes (FTE) to have fairly evaluation of performance among FTE branches in giving educational courses to the society(year 2006-2009),Complete Linkage

method gives perfect and meaningful structure to have analytical method for fairly evaluation of performance to be identical with the standards of quality and accreditation, while the traditional evaluation depends only on the number of courses.

1. Introduction:

The term cluster analysis (first used by Tryon, 1939) encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories. A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop taxonomies. In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation. In other words, cluster analysis simply discovers structures in data without explaining why they exist [1, 2].

2. Hierarchical Tree: Consider a Horizontal Hierarchical Tree Plot (figure 1), on the left of the plot, we begin with each object in a class by itself. Now imagine that, in very small steps, we "relax" our criterion as to what is and is not unique. Put another way, we lower our threshold regarding the decision when to declare two or more objects to be members of the same cluster.

As a result we link more and more objects together and aggregate (amalgamate) larger and larger clusters of increasingly dissimilar elements. Finally, in the last step, all objects are joined together. In these plots, the horizontal axis denotes the linkage distance (in Vertical Icicle Plots, the vertical axis denotes the linkage distance). Thus, for each node in the graph (where a new cluster is formed) we can read off the criterion distance at which the respective elements were linked together into a new single cluster. When the data contain a clear "structure" in terms of clusters of objects that are similar to each other, then this structure will often be reflected in the hierarchical tree as distinct branches. As a result of a successful analysis with the joining method, one is able to detect clusters (branches) and interpret those branches [3].

3. Distance Measures:

The joining or tree clustering method uses the dissimilarities (similarities) or distances between objects when forming the clusters. Similarities are a set of rules that serve as criteria for grouping or separating items. These distances (similarities) can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for grouping objects. For example, if we were to cluster fast foods, we could take into account the number of calories they contain, their price, subjective ratings of taste, etc. The most straightforward way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances. If we had a two- or three-dimensional space this measure is the actual geometric distance between objects in the space (i.e., as if measured with a ruler). However, the joining algorithm does not "care" whether the distances that are "fed" to it are actual real distances, or some other derived measure of distance that is more meaningful to

the researcher; and it is up to the researcher to select the right method for his/her specific application.

Euclidean distance. This is probably the most commonly chosen type of distance. It is simply the geometric distance in the multidimensional space. Which is computed as:

$$\text{Distance}(x,y) = \{ \Sigma_i (x_i - y_i)^2 \}^{½}$$

Note that Euclidean distances are usually computed from raw data, and not from standardized data. This method has certain advantages (e.g., the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers). However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed. For example, if one of the dimensions denotes a measured length in centimeters, and you then convert it to millimeters (by multiplying the values by 10), the resulting Euclidean distances (computed from multiple dimensions) can be greatly affected (i.e., biased by those dimensions which have a larger scale), and consequently, the results of cluster analyses may be very different. Generally, it is good practice to transform the dimensions so they have similar scales [3,4].

4. Clustering Methods.

The clustering methods as follows [4]:

1- **Single linkage (nearest neighbor).**
2- **Complete linkage (furthest neighbor).**
3- **Unweighted pair-group average.**
4- **Weighted pair-group average.**
5- **Unweighted pair-group centroid.**
6- **Weighted pair-group centroid (median).**
7- **Ward's method.**

5. How They Work (The General Algorithm):
Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering (defined by S.C. Johnson in 1967) is this:

1. **Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.**
2. **Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.**
3. **Compute distances (similarities) between the new cluster and each of the old clusters.**
4. **Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. Of course there is no point in having all the N items grouped in a single**

cluster but, once you have got the complete hierarchical tree, if you want k clusters you just have to cut the k-1 longest links.

Step 3 can be done in different ways, which is what distinguishes single-linkage from complete-linkage.

This kind of hierarchical clustering is called agglomerative because it merges clusters iteratively. There is also a divisive hierarchical clustering which does the reverse by starting with all objects in one cluster and subdividing them into smaller pieces. Divisive methods are not generally available, and rarely have been applied(figure 2)[5].

6. Single-Linkage Clustering: (The Algorithm):
Let's now take a deeper look at how Johnson's algorithm works in the case of single-linkage clustering.
The algorithm is an agglomerative scheme that erases rows and columns in the proximity matrix as old clusters are merged into new ones.

The N*N proximity matrix is D = [d(i,j)]. The clustering's are assigned sequence numbers 0,1,....., (n-1) and L(k) is the level of the kth clustering. A cluster with sequence number m is denoted (m) and the proximity between clusters (r) and (s) is denoted d [(r),(s)].

The algorithm is composed of the following steps:

1. **Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.**
2. **Find the least dissimilar pair of clusters in the current clustering, say pair (r), (s), according to**
   d[(r),(s)] = min d[(i),(j)]
   where the minimum is over all pairs of clusters in the current clustering.
3. **Increment the sequence number : m = m +1. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)]**
4. **Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined in this way:**
   d[(k), (r,s)] = min d[(k),(r)], d[(k),(s)]
5. **If all objects are in one cluster, stop. Else, go to step 2.**

In single-linkage clustering (also called the connectedness or minimum method), we consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, we consider the similarity between one cluster and another cluster to be equal to the greatest similarity from any member of one cluster to any member of the other cluster. In complete-linkage clustering (also called the diameter or maximum method), we consider the distance between one cluster and another cluster to be equal to

the greatest distance from any member of one cluster to any member of the other cluster [6].

7. Standards for Accreditation.

There are eleven Standards as follows:

1. **Mission and Purposes.**
2. **Planning and Evaluation.**
3. **Organization and Governance.**
4. **The Academic Program.**
5. **Faculty.**
6. **Students.**
7. **Library and Other Information Resources.**
8. **Physical and Technological Resources.**
9. **Financial Resources.**
10. **Public Disclosure.**
11. **Integrity.**

The institution undertakes planning and evaluation appropriate to its needs to accomplish and improve the achievement of its mission and purposes. It identifies its planning and evaluation priorities and pursues them effectively.

2.1 Planning and evaluation are systematic, comprehensive, broad-based, integrated, and appropriate to the institution. They involve the participation of individuals and groups responsible for the achievement of institutional purposes. Results of planning and evaluation are regularly communicated to appropriate institutional constituencies. The institution allocates sufficient resources for its planning and evaluation efforts.

Planning.

2.2 The institution undertakes short- and long-term planning, including realistic analyses of internal and external opportunities and constraints. The institution systematically collects and uses data necessary to support its planning efforts and to enhance institutional effectiveness. It plans for and responds to financial and other contingencies, establishes feasible priorities, and develops a realistic course of action to achieve identified objectives. Institutional decision-making, particularly the allocation of resources, is consistent with planning priorities.

2.3 The institution has a demonstrable record of success in implementing the results of its planning.

Evaluation .

2.4 The institution regularly and systematically evaluates the achievement of its mission and purposes, giving primary focus to the realization of its educational objectives. Its system of evaluation is designed to provide relevant and trustworthy information to support institutional improvement, with an emphasis on the academic program. The institution's evaluation efforts are effective for addressing its unique circumstances. These efforts use both quantitative and qualitative methods.

2.5 The institution has a system of periodic review of academic and other programs that includes the use of external perspectives.

2.6 Evaluation enables the institution to demonstrate through verifiable means its attainment of purposes and objectives both inside and outside the classroom. The results of evaluation are used systematically for improvement and to inform

institutional planning, especially as it relates to student achievement and resource allocation.

Institutional Effectiveness.

2.7 The institution determines the effectiveness of its planning and evaluation activities on an ongoing basis. Results of these activities are used to further enhance the institution's implementation of its purposes and objectives [7].

8. The Implementation:

Foundation of Technical Education (FTE) runs (27) technical institutes and (13) technical colleges.

The main specializations offered by FTE are Engineering, Administration, Health and Medicine, Agriculture, and Applied Arts , in addition (FTE) has the FTE center and Staff Developing Center(SDC) .

All its branches give educational courses and updated information and skills in every aspect and specialization.

The aim of the research is to classify the FTE branches into closest groups by using (cluster analysis) implementing clustering methods, through which we can be able to evaluate the performance of the branches fairly.

The FTE branches have been coded as follows:

X1:Institute of Medical Technology/Baghdad X2: Institute of Technology /Baghdad

X3: Institute for Administration/Rassafa X4: Applied Arts Institute
X5: Technical Institute/Basrah X6: Technical Institute/Mosual
X7: Technical Institute/Kirkuk X8: Technical Institute/Babil
X9: Technical Institute for Administration X10: Technical Institute/Anbar
X11: Technical Institute/Najaf X12: Technical Institute/Aumara
X13: Technical Institute/Mussyab X14: Technical Institute/Shatra
X15: Technical Institute/Nasiria X16: Technical Institute/Kut
X17: Technical Institute/Hawija X18: Technical Institute/Kufa
X19: Technical Institute/Alsuwyra X20: Technical Institute/Baquba
X21: Technical Institute/Semawa X22: Technical Institute/Karbala
X23: Technical Institute/Door X24: Institute of Medical T/mansor
X25: Technical Institute/Dewanya X26: Technical Institute/Nainawa
X27: Technical Instructors Training Institute X28: Technical College/Baghdad
X29: Technical College/ Mosual X30: Technical College/ Basrah
X31: Health and Medical Technical College X32: Technical College/ Mussyab
X33: Technical College/ Kirkuk X34: Technical College/ Najaf
X35: Electrical and Electronic Tech College Administration/Bghdad X36: Technical College for
X37: FTE Center X38: Staff Developing Center (SDC)

Four colleges have been added recently

X39: Technical College for Administration/kufa
X40: Technical College for Administration/Mosual
X41: Technical College for Administration/Basrah
X42: College for Applied Arts

The variables using to measure the similarities have been covered all the effective sides of Educational Operations in offering Educational courses for all FTE branches

Where the variables as follows:

Y1: Establish year **Y2: Number of Departments**
Y3: Number of Planned Programs **Y4: Number of executed Programs**
Y5: Number of Added Programs **Y6: Number of Participants from FTE**
Y7: Number of Participants out of FTE **Y8: Number of Instructors**
Y9: Number of Technical Staff **Y10: Courses (in/out of) summer holiday (values (1, 0)).**

The size of matrix is (38 * 10) contains real data which has been collected from the annual report of the year(2006) Table(1)[8], matrix (42*10) from the annual report of the year(2009) Table(2)[9],the difference between the two years was so meaningful in showing the useful of the proposed analytical method.

Two methods have been applied (Single Linkage and Complete Linkage) to the proposed case, the results shown in figures (3, 4, 5), the meaningful structure is shown in figures (4,5) which means that Complete Linkage gives best results for the proposed case by dividing the FTE branches into five clusters .

9. Conclusions and Recommendations:

The following conclusions can be staffed:

1. Traditional evaluation depends on the number of courses given to the society where the proposed method gives fairly evaluation according to their performance, facilities and capabilities.

2. Complete linkage method shows five clusters, branches ordered within the clusters according to their priority depending on the scale of y-axis (linkage distance).

3. Year (2006) clusters with scale of y-axis (linkage distance 0-2000) (figure 4) as follows:

Cluster A:
- **X38: Staff Developing Center (SDC).**

Cluster B:
- **X6: Technical Institute/Mosual**
- **X11: Technical Institute/Najaf**
- **X12: Technical Institute/ Aumara**
- **X22: Technical Institute/ Karbala.**

Cluster C:
- **X5: Technical Institute/Basrah**
- **X7: Technical Institute/Kirkuk**
- **X2: Institute of Technology/Baghdad.**

Cluster D:
- **X37: FTE Center**
- **X16: Technical Institute/Kut**
- **X26: Technical Institute/ Nainawa**
- **X36: Technical College for Administration/Bghdad**
- **X28: Technical College/ Baghdad**
- **X33: Technical College/ Kirkuk**
- **X35: Electrical and Electronic Tech College**

X34: Technical College/ Najaf
X17: Technical Institute/Hawija
X30: Technical College/ Basrah
X23: Technical Institute/Door
X32: Technical Institute/Mussyab
X18: Technical Institute/Kufa
X10: Technical Institute/Anbar
X4: Applied Arts Institute.

Cluster E:

X3: Institute for Administration/Rassafa
X29: Technical College/ Mosual
X20: Technical Institute/Baquba
X1: Institute of Medical Technology/Baghdad
X31: Health and Medical Technical College
X19: Technical Institute/Alsuwyra
X21: Technical Institute/Semawa
X9: Technical Institute for Administration
X25: Technical Institute/Dewanya
X8: Technical Institute/Babil
X24 Institute of Medical T/mansor
X13 Technical Institute/Mussyab
X27: Technical Instructors Training Institute
X15: Technical Instructors Training Institute
X14 Technical Institute/Shatra.

4. As mentioned before and illustrated by the results, the evaluation of performance must take the facilities and capabilities of each branch into consideration and each branch must be evaluated within its cluster.

5. X38, X6, X5, X37, X3, have the high priority in the annual performance of the year (2006).

6. Year (2009) clusters with scale of y-axis (linkage distance 0-1200) (figure 5) as follows:

Cluster A: X38 , X6.
Cluster B: X26.
Cluster C:X36 ,X42 ,X29 ,X34 ,X33 ,X41 ,X21 ,X19 ,X28 ,X23 ,X4 ,X35 ,X30 ,X37 X10 , X17, X9, X40, X39.
Cluster D:X25 ,X22 ,X13 ,X18 ,X3 ,X12 ,X31 ,X24 ,X32 ,X27 ,X16 ,X15.
Cluster E:X11, X7, X2, X1, X8, X5, X20, X14.

7. X38, X6 (have the same distance), X26, X36, X25, X11, have the high priority in the annual performance of the year (2009).

8. FTE supplied its branches with facilities and capabilities; it is clearly, when we notice the difference between the performance of the year (2006) and the year (2009).

9. X: 38 Staff Developing Center (SDC) still have the first rank in clusters because it is specialist in offering educational courses to the society (2006 and 2009), where X6: Technical Institute/Mosual have the first rank in second cluster (B) (year 2006) and jumps to the first cluster (A) competes with X: 38 in the same distance (year 2009) which mean that it precedes all the branches.

10. We can easily follow up the performance of all the branches by analyzing the clustering diagrams.

11. The institution determines the effectiveness of its planning and evaluation activities on an ongoing basis. Results of these activities are used to further enhance the institution's implementation of its purposes and objectives therefore; we recommended that this method should be taken into consideration, since it gives wide analyzing and fairly evaluation according to the standards for accreditation.

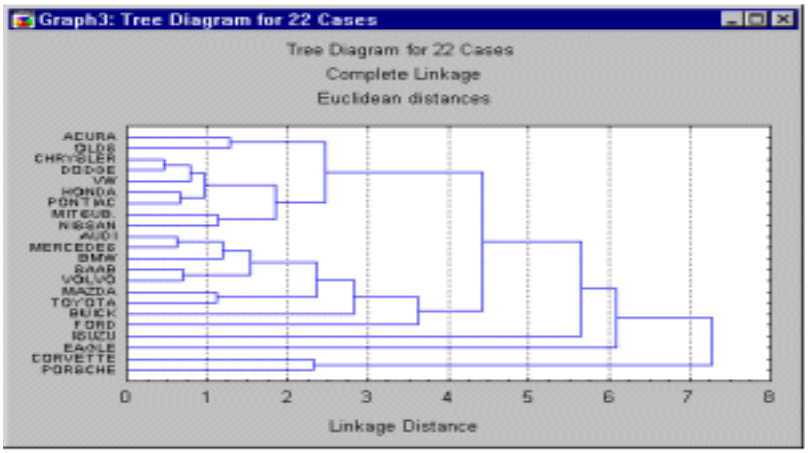


Figure (1) shows the tree diagram

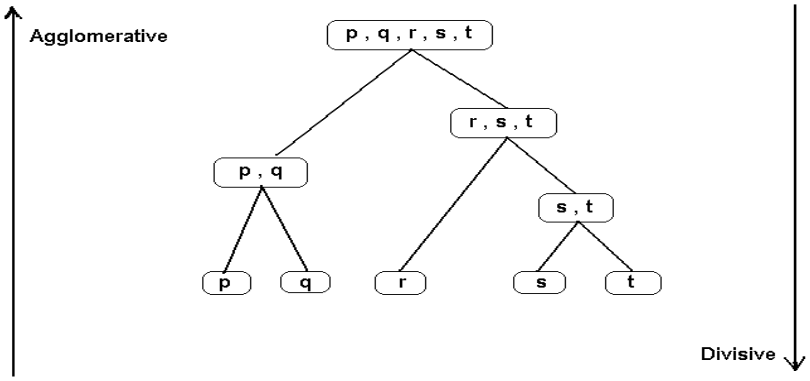
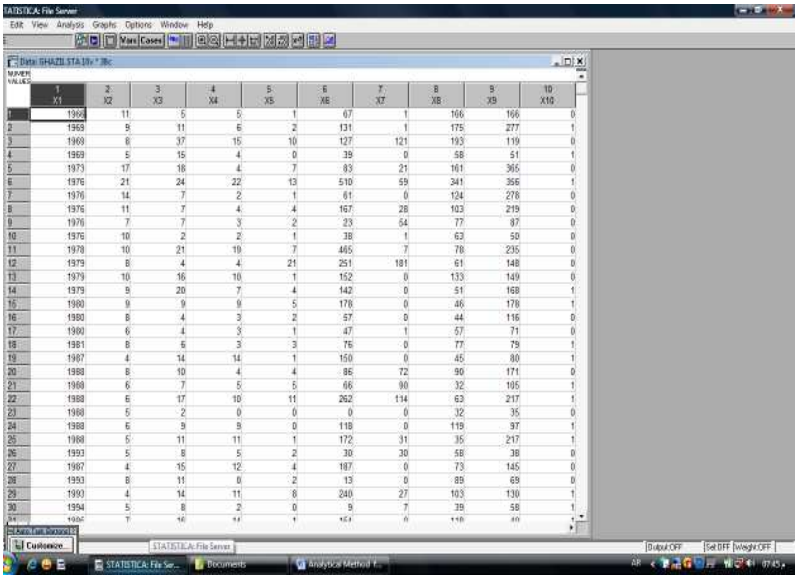
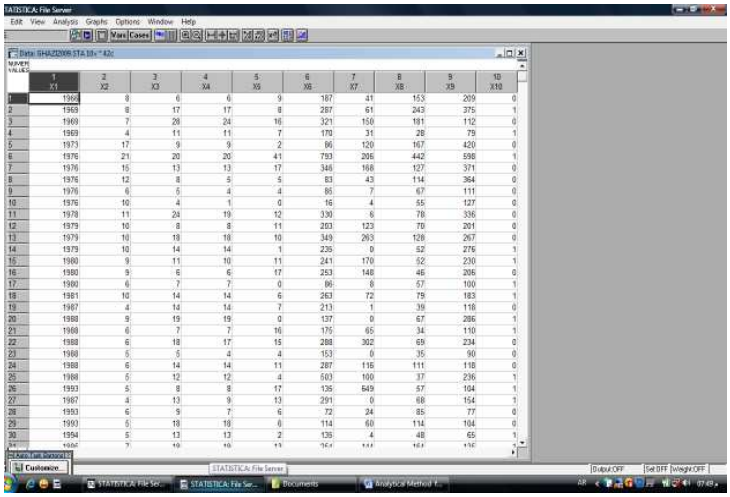


Figure (2) shows the Agglomerative and Divisive



	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	1968	11	5	5	1	67	1	166	166	0
2	1968	9	11	6	2	131	1	175	277	1
3	1969	8	37	15	10	127	121	193	119	0
4	1969	5	15	4	0	39	0	58	51	1
5	1973	17	18	4	7	83	21	161	365	0
6	1976	21	24	22	13	510	59	341	395	1
7	1976	14	7	2	1	61	0	124	278	0
8	1976	11	7	4	4	167	28	103	219	0
9	1976	7	7	3	2	23	54	77	87	0
10	1976	10	2	2	1	38	1	63	50	0
11	1978	10	21	19	7	465	7	78	235	0
12	1979	8	4	4	21	251	181	61	148	0
13	1979	10	16	10	1	152	0	133	149	0
14	1979	9	20	7	4	142	0	51	168	1
15	1980	9	9	9	5	178	0	46	178	1
16	1980	8	4	3	2	57	0	44	116	0
17	1980	6	4	3	1	47	1	57	71	0
18	1981	8	6	3	3	76	0	77	79	1
19	1987	4	14	14	1	150	0	45	80	1
20	1988	8	10	4	4	86	72	90	171	0
21	1988	6	7	5	5	66	90	32	105	1
22	1988	6	17	10	11	262	114	63	217	1
23	1988	5	2	0	0	0	0	32	35	0
24	1988	6	9	9	0	118	0	119	97	1
25	1988	5	11	11	1	172	31	35	217	1
26	1993	5	8	5	2	30	30	58	38	0
27	1987	4	15	12	4	187	0	73	145	0
28	1993	8	11	0	2	13	0	89	69	0
29	1990	4	14	11	8	240	27	103	130	1
30	1994	5	8	2	0	9	7	39	58	1

Table (1) shows data extracted from annual report (2006).



	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	1968	8	6	6	9	187	41	153	209	0
2	1969	8	17	17	8	287	61	243	375	1
3	1969	7	28	24	16	321	150	181	112	0
4	1969	4	11	11	7	170	31	28	79	1
5	1973	17	9	9	2	86	120	167	420	0
6	1976	21	20	20	41	793	206	442	598	1
7	1976	15	13	13	17	348	168	127	371	0
8	1976	12	8	5	5	83	43	114	364	0
9	1976	6	5	4	4	86	7	67	111	0
10	1976	10	4	1	0	16	4	55	127	0
11	1978	11	24	19	12	330	6	78	336	0
12	1979	10	8	8	11	263	123	70	291	0
13	1979	10	18	18	10	349	263	128	267	0
14	1979	10	14	14	1	235	0	52	276	1
15	1980	9	11	10	11	241	170	52	230	1
16	1980	9	6	6	17	253	148	46	206	0
17	1980	6	7	7	0	86	8	57	100	1
18	1981	10	14	14	6	263	72	79	183	1
19	1987	4	14	14	7	213	1	39	118	0
20	1988	9	19	19	0	137	0	67	286	1
21	1988	6	7	7	16	175	65	34	110	1
22	1988	6	18	17	15	288	302	69	234	0
23	1988	5	5	4	4	153	0	35	90	0
24	1988	6	14	14	11	287	116	111	118	0
25	1988	5	12	12	4	503	100	37	236	1
26	1993	5	8	8	17	135	649	57	104	1
27	1987	4	13	9	13	291	0	68	154	1
28	1993	6	9	7	6	72	24	85	77	0
29	1990	5	18	18	6	114	60	114	104	0
30	1994	5	13	13	2	135	4	48	65	1

Table (2) shows data extracted from annual report (2009).

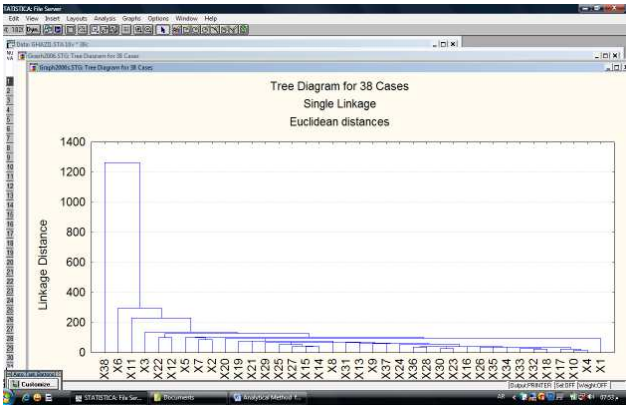


Figure (3) shows single linkage distance of 2006.

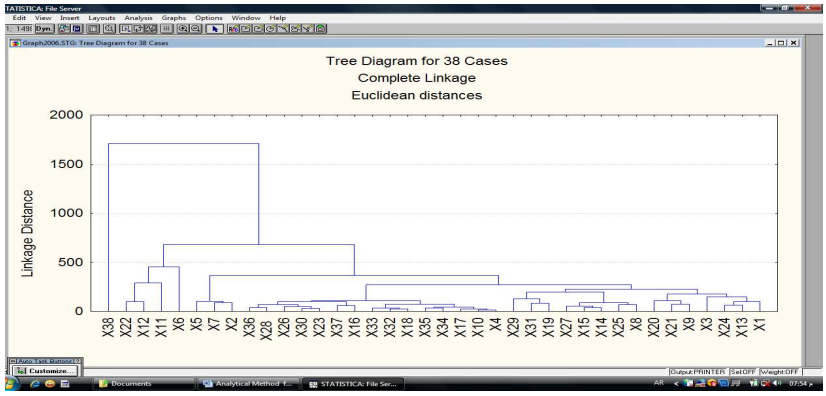


Figure (4) shows complete linkage distance of 2006.

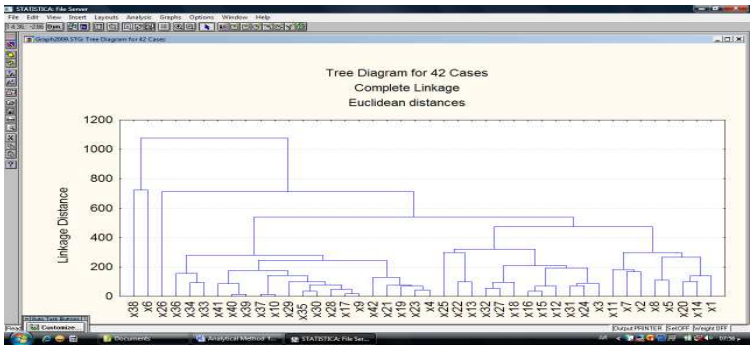


Figure (5) shows complete linkage distance of 2009.

10. References:

1. S. C. Johnson (1967): "Hierarchical Clustering Schemes" Psychometrika, 2:241-254.

2. R. D'andrade (1978): "U-Statistic Hierarchical Clustering" Psychometrika, 4:58-67.

3. Andrew Moore: "K-means and Hierarchical Clustering - Tutorial Slides" http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html.

4. Osmar R. Zaïane: "Principles of Knowledge Discovery in Databases - Chapter 8: Data Clustering" http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html.

5.Stephen P.Borgatti: "How to explain hierarchical clustering" http:/ /www.analytictech .com /networks/hiclus.htm.

6.Maria Irene Miranda: "Clustering methods and algorithms" http://www.cse.iitb.ac.in/dbms/Data/Courses/CS632/1999/clustering/dbms.html.

7. Standards for Accreditation "Commission on Institutions of Higher Education", 209 Burlington Road , Bedford, MA 01730/ 2006.

8. Foundation of Technical Education/Annual Report 2006/Baghdad-Iraq.

9. Foundation of Technical Education/Annual Report 2009/Baghdad-Iraq.