Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 10, October 2014, pg.944 – 949

RESEARCH ARTICLE

Automatic Clustering Old and Reverse Engineered Databases by Evolutional Processing

Hasan Asil
(Islamic Azad University, azarshahr branch, azarshahr, Iran
H.Asil@asd.ir)

Amir Asil
(University of Tabriz, Tabriz, Iran
Asil@asd.ir)

Abstract: ***Keeping of old databases is difficult. Especially when the documentaries of system are infirm or written documentaries are deleted. Reverse engineering methods are tried to improve and solving this problem. Various methods and conceptive models are presented for discovering and extraction old databases. Some of these methods are tried to do normalization up to different levels. But after normalization, perception conceptive of these databases and also the relations of system's different parts are considered. So for improve, perception and managing these databases better, various methods are presented for their automatic clustering. This research presents an automatic algorithm for clustering of the relations of database tables by evolutional processing. This suggested method tries to optimize the time of execution of this clustering for managing the databases better.***

Keywords: Database Normalization, Legacy Databases, Reverse Engineering, Database Design

1. Introduction

Old databases are very valuable for the organizations. Most of these databases are developed by old programming languages like COBOL and RPG and system file [Lee, 00].Some of these databases have been outdated by new technologies. Even these databases are hierarchy or Data model. So for solving these problems, new methods are presented for changing the structure and changing them into a modern database corresponding previous database [Cho, 10]. One of these methods uses reverse engineering. This method tries to design a system based on reverse engineering in order to present a modern and normal database instead of old database. In the database community, reverse engineering tries to extraction the meanings of domain like practical keys, dependants and available limitations in the structure of data bank [6,7]. As usual, reverse engineering changes the logical model into conceptive model. This method tries to change the old database into normal database up to 3rd Normal [Cho, 10]. This method tries to convert this database by reverse engineering and data mining. But the above method converts the database into a normal database. In big databases it's not simply possible to form subsystems and their relations accurate because of the high volume of data [Kun , 11]. This paper tries to present a method based on evolutional

processing for better management of normal and old converted databases in order to specify data bank of subsystems and different parts of database based on maximum relations by using the methods of clustering. This method tries to cluster the tables of database based on available relations and by this clustering, the modern converted database is managed better. In next sections the way of presentation of this algorithm will be explained.

2. Reviewing the past works

Various methods are presented for databases clustering. But most of these methods are based on ER diagram. One of these methods is Martin's method. He has used layering on ER clustering in his past works which each one is declared by a 1*n relation. In this method a root existence without 1*n relation is placed on top of each level. The clusters with the root existences are shaped as center and the children are shaped as elements in clustering. In this algorithm the main existences should not enjoy 1*n relation [Madjid , 01].

Two level clustering is the main advantage of martin's method.

Miller analyzed the Martin's clustering and changed grouping the existence's clusters by focusing on the properties. In another method which had presented by Terror.

All of these methods are depended on human's judgment [Kun , 11] for afresh solving of borders and specification of relations power. Huffman was the first person who has tried to make ER clustering automatic by spreading a system by using a chain of rules. Also there are other methods presented by Francis and AKak which gently deleted the dependence to the human based on some rules [Madjid , 01].

3. Aim

As it's said one of the most important advantages of databases clustering is better management of it. In fact by clustering we can specify several advantages from analysis of the database structure. Some of these advantages are pointed below [Madjid , 01]:

- They present a simpler method for spreading and apply and cause to further introduction of modules.
- It's simpler that save them as documentary. Because they are divided to smaller groups.
- Their evaluation is simpler.
- They are able to help to the management of project. Because they specify the modules and cause to the lucidity of the duties.
- They're able to help in recognition of the systems or cause to the deletion of unwanted parts and usage of subsystems in other places.

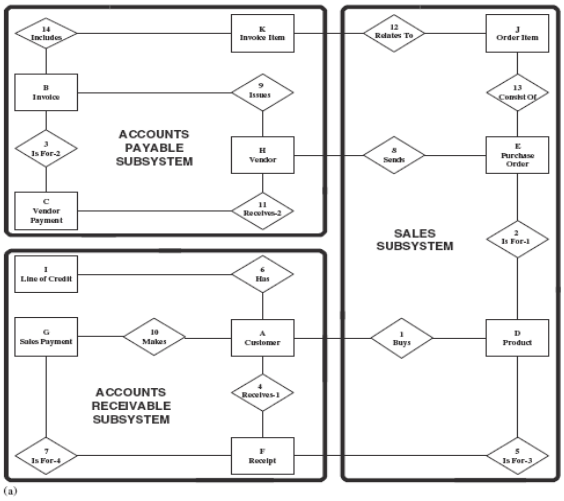
Notwithstanding these advantages, nowadays none of the tools of diagram drawing use clustering. This is because that the clustering algorithms are recalled as system.

4. Suggested algorithm

After conversion of database to normal state by the past methods[Cho, 10] we'll try to suppose the database as a matrix and then cluster this large matrix. Beside the separation case, another concept is clustering that there is no contacts among the different groups of community or the number of contacts between these two sets is minimum. Clustering problem is expressed in various ways[Lammari, 07]: the graph can be Oriented or not, the graph can be weighted etc. in some cases it's possible that the number of partitions will be important but in some cases only the separation owns importance to us. In these cases the important thing will be the least number of contacts between two sets and great number of the contacts between members (or different existences) .

In this algorithm it's tried to present a new method for clustering the tables of the database by using a collection of rules of the past works and also by evolutional processing and analyze the results.

The below figure shows a sample of clustering algorithm for ER diagram of a database and the clustering has been performed based on it's Vicinity matrix.



ENTITY	RELATIONSHIP													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	1	0	0	1	0	1	0	0	0	1	0	0	0	0
B	0	0	1	0	0	0	0	0	1	0	0	0	0	1
C	0	0	1	0	0	0	0	0	0	0	1	0	0	0
D	1	1	0	0	1	0	0	0	0	0	0	0	0	0
E	0	1	0	0	0	0	0	1	0	0	0	0	1	0
F	0	0	0	1	1	0	1	0	0	0	0	0	0	0
G	0	0	0	0	0	0	1	0	0	1	0	0	0	0
H	0	0	0	0	0	0	0	1	1	0	1	0	0	0
I	0	0	0	0	0	1	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	0	0	1	1	0
K	0	0	0	0	0	0	0	0	0	0	0	1	0	1

(b)

Weak/Subtype Entity	Strong/Supertype Entity
J	E
K	B

(c)

Figure 1- A sample of clustering[Madjid , 01]

Above figure shows a sample of clustering algorithm and the diagram has been categorized by the past methods. But the problem is the long time of clustering algorithm execution that has expressed as a NP hard problem. But the Genetic algorithm as a method that shows its ability in the clustering field can be an optimizer and simpler method for this type of implementation. We will use this method in the suggested algorithm. For solving the problem we suppose the relations of tables as a graph and each one of existences is supposed as a vertex and the relations are supposed as the graph's edge and we start the clustering.

Figure 2 shows a sample of graph and its clustering.

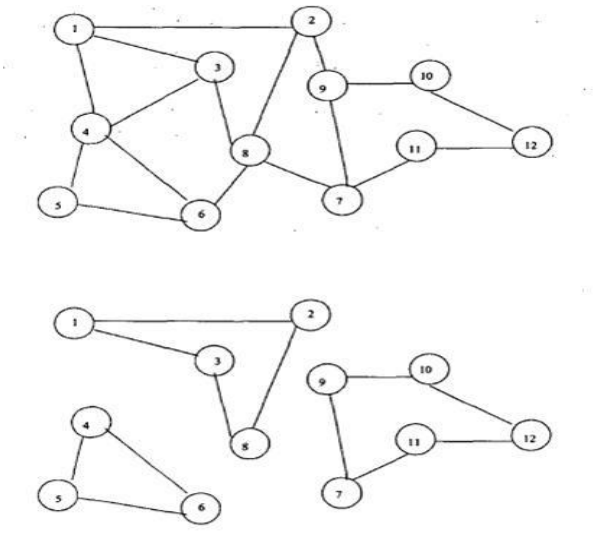


Figure 2- A sample of clustered graph[cincoti, 02]

The first stage in this algorithm is adding the primary population and we show it with P symbol. Below algorithm shows the manner of creating partitions. K symbol shows the number of partitions.

```
Algoritm Creat _ Partitions
Input A connected Graph G=(V,E), an integer  1<k<|V|=n, and a permutation p of V
Output : connected and pariwise disjoint k subgraphs of :G
For i=1 to k do
Assign pi to the partition i.
End_ for
While there are free vertices do
For i=k+1 to n do
If pi is free then
Assign pi to the smallest adjacent partition, if any;;
End for
End while
End algorithm
```

Algorithm 1- The manner of adding the clustering[cincoti, 02]

Our fitness in this problem for chromosomes is reached by the below relation (equation) which should be supposed minimum.

$$(K.M)/|V| \quad (1)$$

In this problem K shows the number of partitions, M is the number of the largest partition's vertexes and V shows the number of vertexes. Below relation is in this problem.

$$(K.M)/|V| \geq 1 \quad (2)$$

In this problem crossover operation is used as a special crossover with the title of OPDX. The reason of using this operation is the possibility of creation of unallowable chromosomes during this operation. We should prevent their production. For example the vertexes which are not adjacent should not be place in the same partition also should not be exist in various partitions.

The main algorithm presented for this problem is shown in algorithm 2.

We can cluster the diagram by the presented algorithm then solve the problem by converting the graph to database.

5. System Evaluation

This system has been designed and implemented as absolutely object oriented system and some of the important elements of this system are: the manner of clustering, the manner of creation of clusters and updating them and execution of algorithm based on vicinity matrix etc.

System evaluation process is done by execution of some experiments which the vicinity graphs and normal databases are used in various sizes. For this experiment we need to some databases which 4 vicinity matrixes with different sizes are used in order to compare with francis method.

The GA for Graph Partitioning

Input: A connected Graph G= (V, E)

Output: connected k subgraphs of G whose Total cardinality is "close" to the average value.

Initialize randomly a population P of 2*n elements

For i=0 to Max Gen do

Compute fitness

Sort the population whit respect to fitness value

Delete half of population whith lower fitness

Crossover

End for

End Algorithm

Procedure compute fitness

For each p ε P do

Call creat_partition(p);

Fitness (p)=k.m(p)/n;

Comment: M(p) is the partition of maximal cardinality

among the k partitions created givin the permutation p;

End for

End procedure

Procedure crossover

For i=1 to n/2 do

Select to parent Pa,P_b ε P randomly

Add 4 individuals Producted By ODPX(P_a,P_b) to P

END for

END Procedure

Algorithm 2- The manner of clustering and adding the cluster[cincoti, 02]

Figure 3- System performance based on time

As it's shown in the above figure, this is new method for clustering the database and in addition to advantages of diagram clustering, enjoys the lower cost and execution time in comparison to presented algorithms and francis algorithm.

6. Result

in this paper it's tried to present a new method for clustering the modern databases which are created by conversion of old databases. Using this method can solve the complexity of converted databases' complete perception for users and even the database engineers and educated designers. For improvement, perception and managing the large databases an absolutely automatic algorithm is presented in order to limit personal judgments. Also this algorithm enjoys less execution time in comparison to the past algorithms. In future the accuracy of this algorithm cane be increased by adding the states like the type of the existences relations and giving weight to them.

References

[cincoti, 02] D. A cincoti and v. cutello ,m pavondept Graph Partitioning using Genetic Algoritm Using Genetec Algorithm whit ODPX :. IEEE 78037282 ,2002

[Cho, 10] j. Cho, Database Reverse Engineering based on Association Rule Mining, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 2, 2010

[Madjid , 01] Madjid Tavana , Prafulla Joglekar , Michael A. Redmond , Entity model An automated entity–relationship clustering algorithm , 0306-4379/$ - see front matter r 2006 Elsevier B.V. All rights reserved.doi:10.1016/j.is.2006.07.001

[Lee, 00] H. Lee, and C. Yoo, "A form driven object-oriented reverse engineering methodology", Information Systems, Vol.25,No.3, 2000, pp. 235-259.

[Francalanci, 94] C. Francalanci, B. Pernici, Abstraction levels for entity relationship schemas, in: P. Loucopoulus (Ed.), Proceedings of 13th International Conference on the Entity–Relationship Approach, Springer, Berlin, Heidelberg, 1994, pp. 456–473.

[Chiang, 97] R. Chiang, T. M. Barron, and V. C. Storey, "A framework for the design and evaluation of reverse engineering methods forrelational databases", Data & Knowledge Engineering,Vol.21, 1997, pp. 57-77.

[Lammari, 11] Likun Liu, Cheng Chen ,Yongwei Wu and Guangwen Yang ,Metadata changes in large file systems: a metadata querying perspective, International Journal of Computer Systems Science and Engineering, Vol 26 No 5 September 2011

[Lammari, 07] N. Lammari, I. Comyn-Wattiau, and J. Akoka, "Extracting generalization hierarchies from relational databases: A reverse engineering approach", Data & Knowledge Engineering, Vol.63, 2007, pp. 568-589.

[Kun , 11] Kun Yue, Wei-Yi Liu and Li-Ping Zhou,"Automatic keyword extraction from documents based on multiple content-based measures", International Journal of Computer Systems Science and Engineering, Vol 26 No 2 March 2011

