Interactive KDD System for Fast Mining Association Rules[+]

Ghazi Johnny [*]

الخلاصه:

يستخلص التنقيب عن القواعد العلائقيه الارتباطات العلائقيه المهمه بين المجاميع الكبيره لعناصر البيانات ،هذا الموضوع مهم جدا في التحليل و تحديدالاتجاهات في كثير من القضايا وخصوصا في الاعمال التجاريه، في هذا البحث تم بناء نظام تفاعلي ليقدم تعامل تفاعلي مع البيانات والمعلوما ت وهو اكثر فائده في اتخاذ القرارات الحاسمه وكذلك مساندا للمحللين بالمعلومات الحديثه من خلال القواعد العلائقيه المستخلصه من القضيه المعطاه.

هذا النظام طبق على حاله مقترحه تتعلق بادويه معالجه ارتفاع ضغط الدم، النظام أعطى مجاميع ادويه مدمجه جديده تقلل عدد الحبوب التي يتناولها المريض يوميا.

Abstract:

Association rule mining finds interesting associations and/or correlation relationships among large sets of data items, this subject is very useful in analyzing and fixing the trends of many issues especially in business , In this research an interactive system have been built to offer interactivity dealing with data and information which is so useful in taking crucial decisions and also support the analysts with up to date information through the extracted association rules from the given issue, the system has been applied in this research for proposed case concerned with drugs used in hypertension management, the system gives anew combinations of joined drugs that reduces the number of tablets swallowed daily by the patient.

Introduction:

With the enormous amount of data stored in files, databases, and other repositories, is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Figure(1) shows data mining as a step in an iterative knowledge discovery process.

[+] **Date Received 21/11/2008 Date of Acceptance 8/6/2009**

[*] **Lecturer/ Staff Developing Center**

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, *data cleaning* and *data integration* can be performed together as a pre-processing phase to generate a data warehouse. *Data selection* and *data transformation* can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, data mining should have been called "knowledge mining" instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery [1, 2].

2. A-priori algorithm and its enhancement:

Agrawal et al. [3, 4] proposed the A-priori algorithm that is used to discover the association rules. The problem of discovering association rules can be decomposed into two sub problems: (1) Find the sets of items (itemsets) that have the supports above the minimum support. (2) Use the frequent itemsets to generate the desired association rules.

Definition 1 The *support* of an itemset I, sup(I), is defined as the number of transactions in the database containing I. The *minimum support*, min_sup, is a user predefined threshold. An itemset is *frequent* if its support is not less than the *min_sup*. An itemset with k items in called a *k*-itemset.

The Apriori algorithm scans the database multiple passes to discover frequent itemsets. The main steps of discovering frequent itemsets can be listed as follows:

(1) First of all, the supports of individual items are counted, and the items with the supports not less than the *min_sup* form frequent 1-itemsets.
(2) Next, let L_k be the set of all frequent *k*-itemsets and C_k be the set of candidate *k*-itemsets. We use L_k to generate C_{k+1} which is a superset of L_{k+1}. By scanning the databases once to check if c is frequent, for each $c \in C_{k+1}$, we can find L_{k+1}. This process repeats recursively until no frequent itemset is found.

Definition 2 Let D be a set of transactions and $I = \{i_1, i_2, \ldots, i_m\}$, an itemset is a subset of I. Given X and Y are itemsets, an *association rule* is of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \Phi$, where the $sup(X \cup Y) \geq min\_sup$, and the *confidence* of $X \Rightarrow Y$ is not less than a predefined threshold, *min_conf*, the *confidence* of $X \Rightarrow Y$ is $\frac{\sup(X \cup Y)}{\sup(X)}$.

After discovering all frequent itemsets, the algorithm of generating association rules uses the subsets of a frequent itemset as antecedents to generate the rules. For example, given a frequent itemset *ABCD*, we first consider the subset *ABC*, then *AB*, etc. If the support of a subset υ of a frequent itemset l dividing that of l is less than the minimum confidence, the support of any subset of a dividing that of l is less than the minimum confidence too. That is because the support of any subset $\acute{\upsilon}$ of υ must be as great as the support of a. Therefore, the confidence of the rule $\acute{\upsilon} \rightarrow (l - \acute{\upsilon})$ cannot be greater than the confidence of $\upsilon \rightarrow (l - \upsilon)$.

The enhancement of the algorithm is to reduce the number of scans to one, and to do that let us discuss the following idea, when choosing set of items contains A, B, C the A-priori scans the database three times one for the sets which contains one item like A, B, C and the second scan is for the sets which contains two items like {A,B} , {A,C} , {B,C}, the third scan is for the sets with three items like {A, B, C}.
Since the number of itemsets can be obtained as follows:

no of itemsets = 2^n -1 where n is the number of items .

And by using the permutation of the giving items which are obtaining by the concatenations programming technique we can determine firstly the itemsets and we

can simply compute the frequent of each itemset in the database by just one main scan.

So the Algorithm as follows:

```
Algorithm: finding all association rules
Input: all frequent itemsets
Output: all association rules
Begin
      Select items of association rules.
      Generate K-itemset
         Input: Selected items.
                              Output: K-itemset-table.
          Begin
                K= 1
                For J=1 to length (S) do
                Begin
                      For I= 1 to length (S) – (K- 1) do
                       Begin
                            Get Concatenates of substring (S, I, K);
                             Write K-itemset-table;
                       End;
                        K = K+1
                 End;
           End;
        Applying A-priori.
             Calculate frequency for each K-itemset.
             Calculate K-support to determine min-support.
             Calculate confidence.
             Determine final association rules.
END.
```

3. The proposed system and Algorithm:

The proposed system offers interactivity dealing with data and information which is so useful in taking crucial decisions and also support the analysts with up to date information through the extracted association rules from the given issue, the proposed system mines seven items each cycle and can extracts association rules from at least two items from the seven up to the whole items together in the inner cycle.

The algorithm of the proposed system will be as follows:

Algorithm: finding all association rules by interactive KDD system
Input: all required items to be mine from database

Output: all association rules

Begin

Open database to get transactions.

Choosing items.

Determine seven items to be mine.

Scan database to Get transactions of the selected items.

Coding.

Coding by using Boolean properties and creating Data file.

Data mining.

Determine items(not less than two up to seven) for extracting association rules.

Determine the confidence threshold value.

Get the permutation of the determined items.

Compute the frequent value for each itemset by using the datafile .

Compute the support values for each itemset.

Compute the confidence values for each related itemset.

Find the association rules.

Reporting.

Display all association rules.

If there are another choose for mining within the selected items go to Data mining else next sentence .

Ending.

If there are another new seven items for mining go to Choosing items else closing.

END.

4. The Implementation:

We applied the system for drugs used in hypertension management where each drug is an example of group acting by the same mechanism but differs slightly among each other ,the aim is to extract anew combinations of joined drugs to reduce the number of tablets swallowed daily by the patient .

The drugs ordered as follows:

A: ASPIRIN	B: ATENOLOL
C: CAPTOPRIL	D: CHLORTHALIDONE
E: AMLODIPINE	F: SIMVASTATIN
G: DIAZEPAM	

We entered more than one hundred real medical description as transaction database records, we fixed the confidence values to 100%, the following steps will illustrates the whole execution of the proposed system.

1. step1:
   Shows the screen of choosing drugs for mining Figure (2).

2. step2:
   Shows the permutation of the chosen drugs Figure (3).

3. step3:
   Shows the frequent of each drug Figure (4).
4. step4:
   Shows the support of each drug Figure (5).
5. step5:
   Shows the confidence and display out the final association rules of the chosen drugs Figure (6).

5. Conclusions and Recommendations:

The following conclusions can be staffed:

1. The proposed system offers one main scan instead of multi-scanning process.
2. The proposed system(medicinal system) gives the following combinations of joined drugs as follows:

A: ASPIRIN + ATENOLOL + CAPTOPRIL
B: ASPIRIN + ATENOLOL + CHLORTHALIDONE
C: ASPIRIN + ATENOLOL + AMLODIPINE
D: ASPIRIN + ATENOLOL + CAPTOPRIL + SIMVASTATIN
E: ASPIRIN + ATENOLOL + DIAZEPAM
F: ASPIRIN + CHLORTHALIDONE + AMLODIPINE
G: ASPIRIN + CHLORTHALIDONE + SIMVASTATIN
H: ASPIRIN + CAPTOPRIL + CHLORTHALIDONE + SIMVASTATIN
I: ASPIRIN + AMLODIPINE + SIMVASTATIN
J: ASPIRIN + CAPTOPRIL + SIMVASTATIN
K: ASPIRIN + CAPTOPRIL + CHLORTHALIDONE
L: ASPIRIN + CAPTOPRIL + CHLORTHALIDONE + DIAZEPAM

3. The proposed system offers interactivity dealing with data and information which is so useful in taking crucial decisions and also supports the analysts to have wide view about the given issue figure(7).
4. The proposed system discovers that there is no relation between some given items at the same run time ,in the medicinal system when the combinations has been extracted we can return back to choose another items like CAPTOPRIL and AMLODIPINE to find that there is no relationship between them figure (8).
5. The results can help the government to assist the patients with chronic diseases by reducing the number of tablets and costs of medicines.

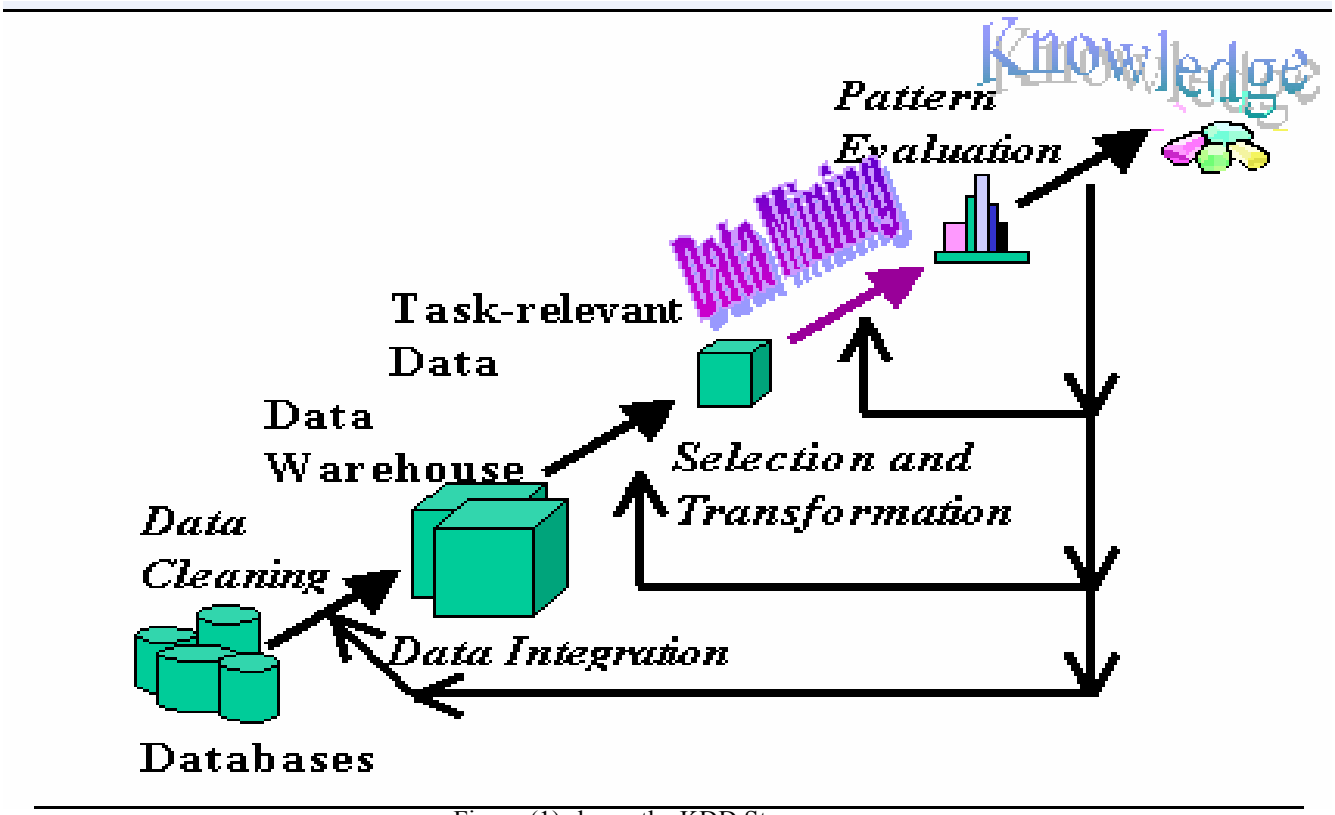

Figure (1) shows the KDD Stages.

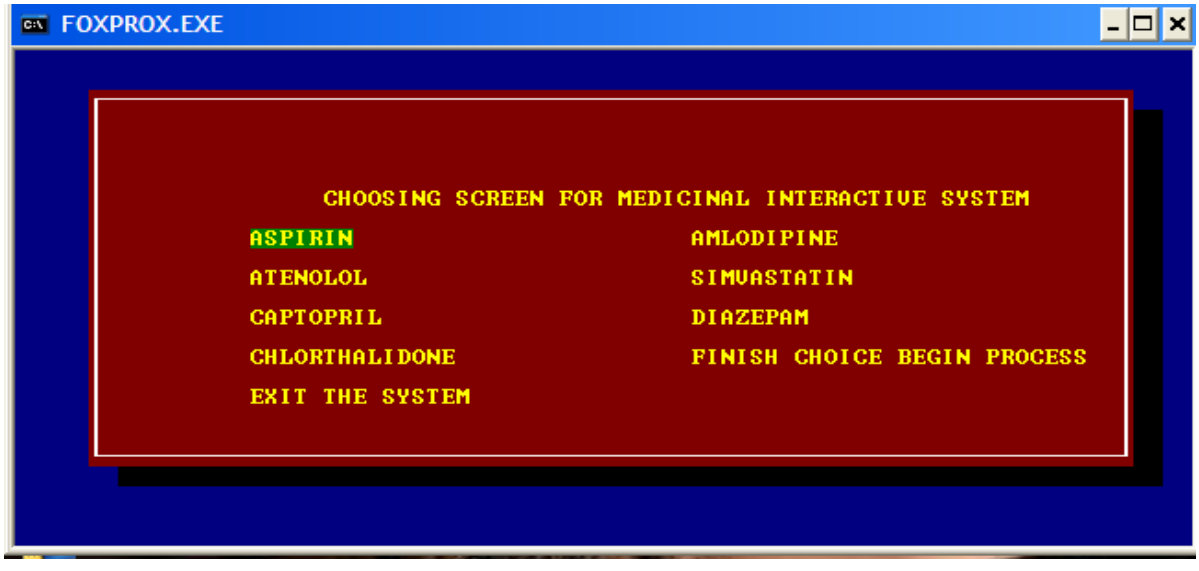

Figure (2) Shows the chosen screen

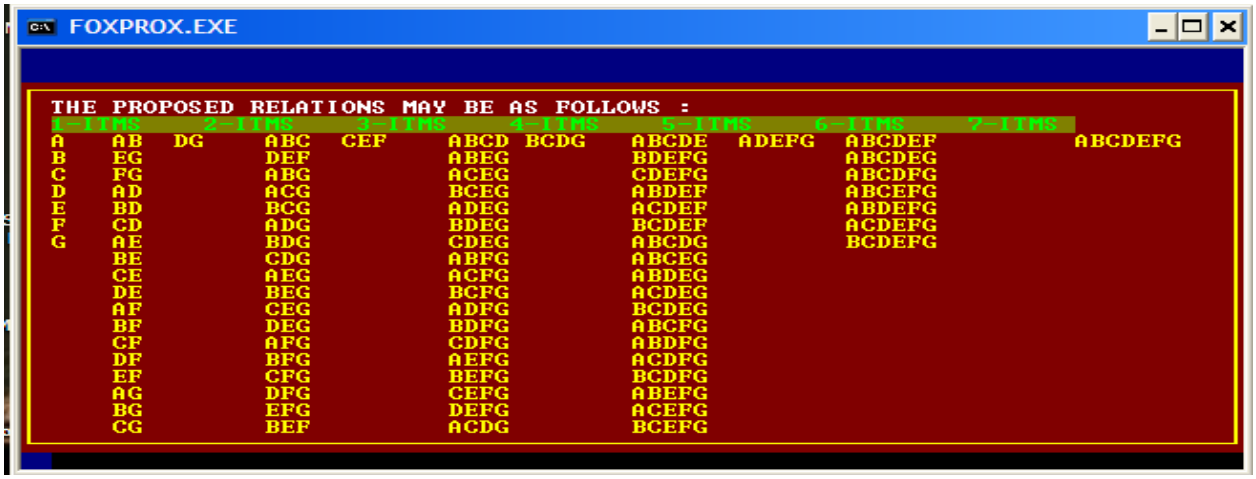

Figure (3) shows the permutation of selected items

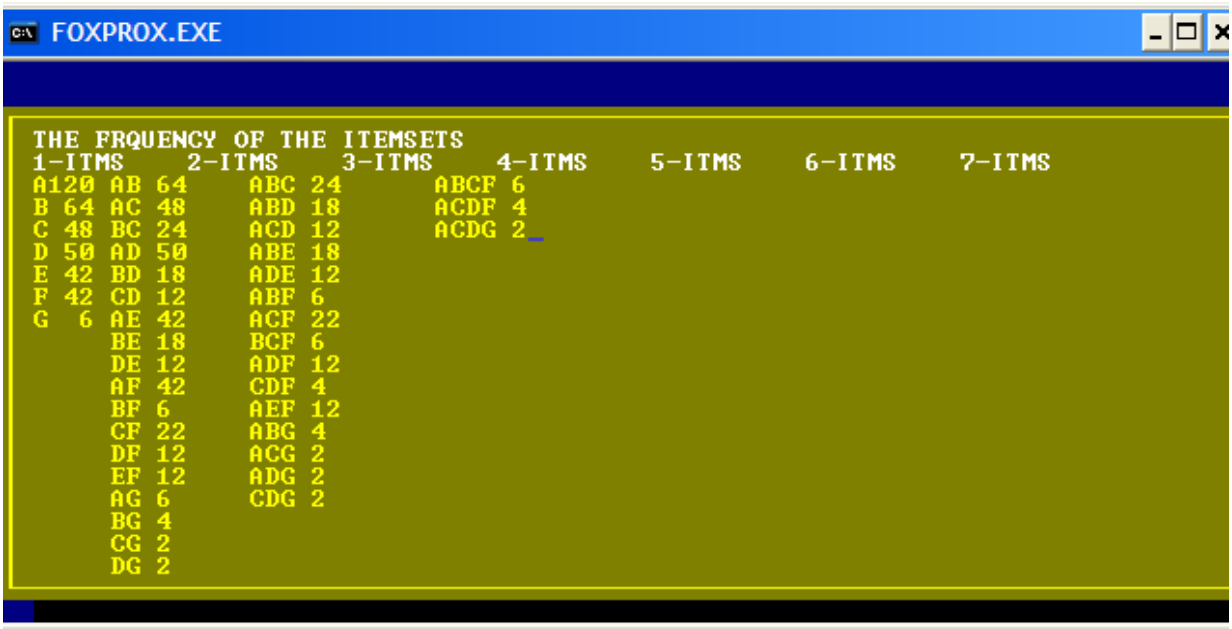

Figure (4) shows the frequent of each itemset

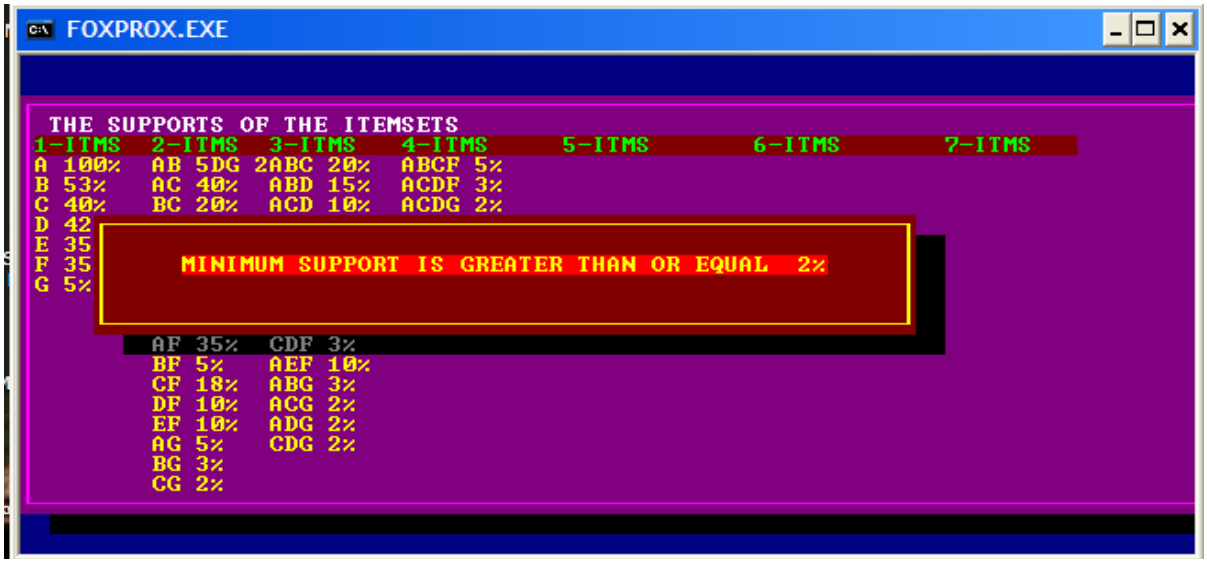

Figure (5) shows the support of itemsets

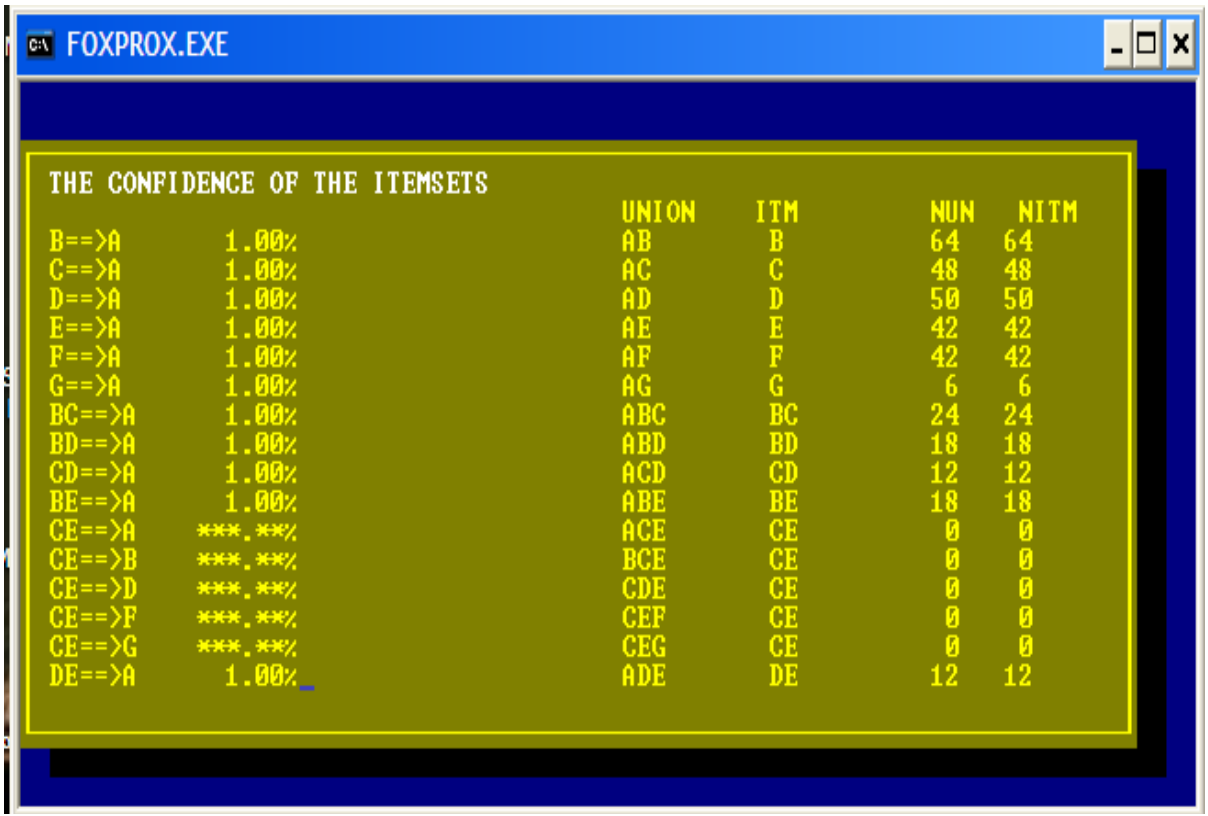

Figure (6) shows some of the confidences and the final association rules

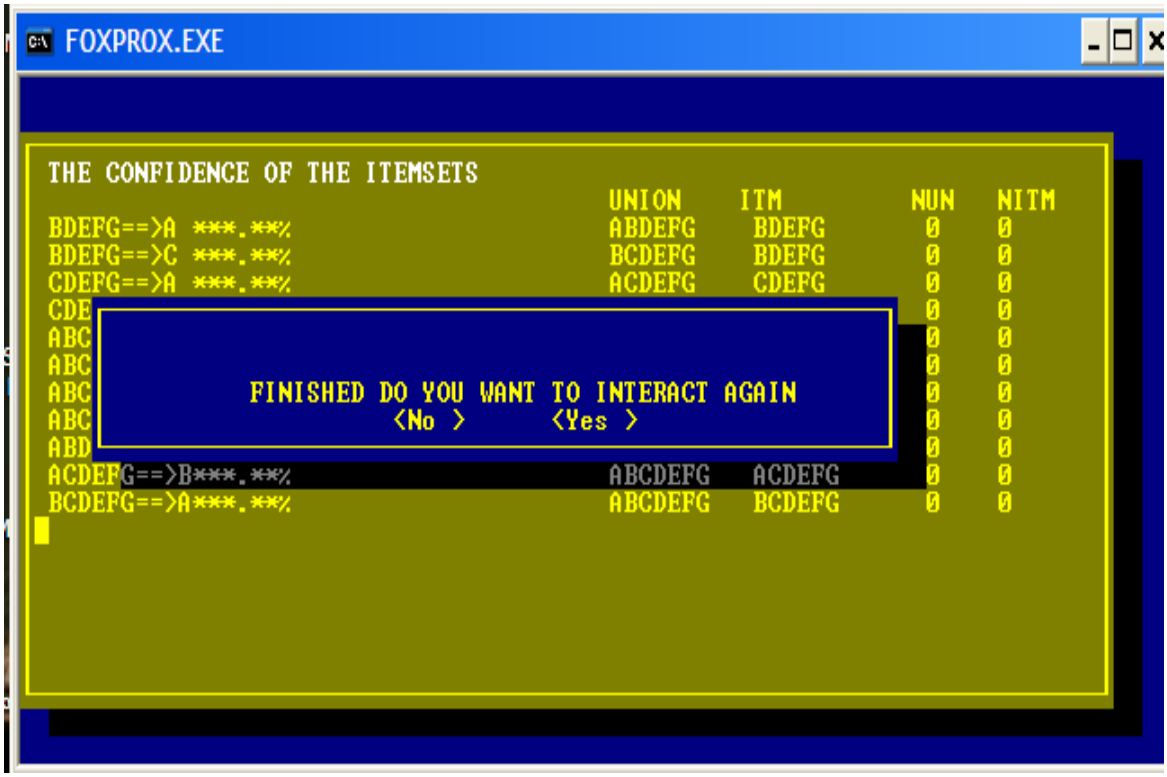

Figure (7) shows the interactivity of the system

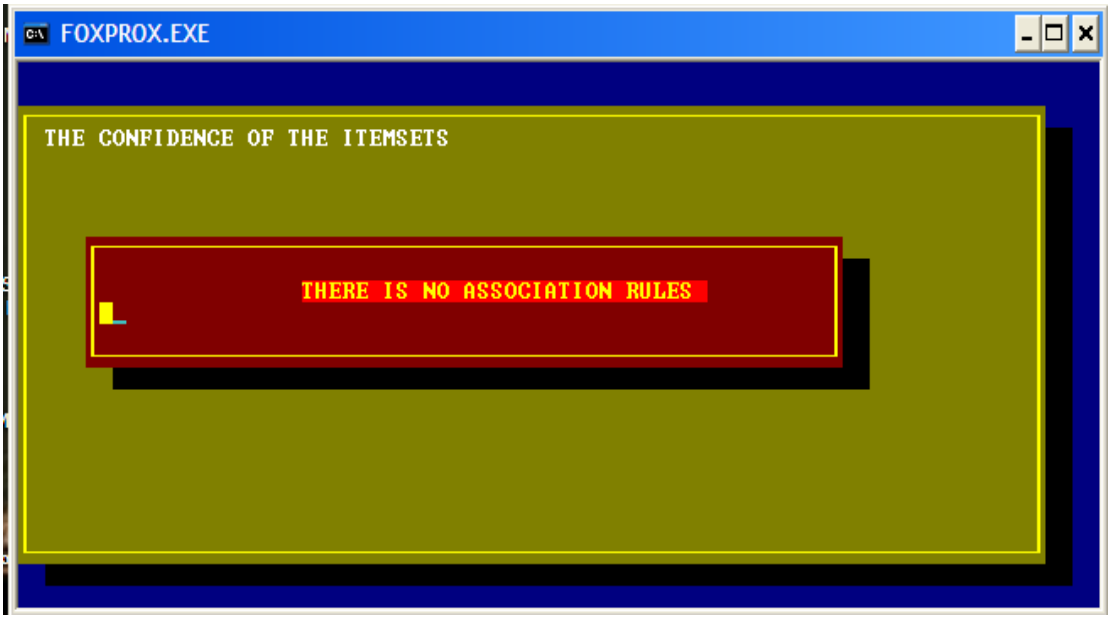

Figure (8) shows that there is no relation rules between some chosen items.

6. References:

[1] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[3] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In Proc. of Int. Conf. on Very Large Data Bases (VLDB'94), Santiago, Chile, pp. 487-499, September 1994.

[4] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", In Proc. 1993 ACM-SIGMOD Int. Conf. on Management of Data, Washington, D.C, pp. 207-216., May 1993.