VKDP: A New Approach in Knowledge Discovery Process

Ass. Prof. Dr. Hussein K. Khafaji

Received on:11/1/201
Accepted on:25/5/2011

Abstract

Visualization is the post processing stage in Knowledge Discovery Process (KDP) to simplify the process of understanding, abstraction, and diminishing the size of mined information, patterns, and/or knowledge. Pre-mining and mining stages of KDP seem as preprocessing steps for visualization engine. Visualization is a complex process because it needs a formal definition of complicated rules to translate large volumes of data into graphic formats. In this research, the term Visual KDP, VKDP was suggested, in which the benefits of visualization techniques have been utilized before, during, and after the data mining stage. To prove the validity and applicability of VKDP approach it is applied to the most important task of Data Mining; Association Rules Mining. The process of finding the appropriate visualization techniques is not a trivial one. Therefore, many visualization techniques are proposed for different levels of Association Rule (AR) mining, i.e., for database under mining, intermediate result or mining level, and mined rule level. Bipartite graph is proposed as a new technique to visualize the database under mining in addition to many variations in horizontal and vertical layouts. Also, networks of concepts are proposed as a new visualization technique to visualize the mined frequent itemsets, while the *two-dimensional matrix, directed graph*, and *rule-item approach* are adopted as visualization techniques to visualize the mined rules.

Keywords: Knowledge Discovery Process, KDP, Data Minind, DM, Visualization, Association Rule

ترئيةُ الاستكشاف: اتجاهٌ جديدٌ في استكشاف المعارف في قواعدِ البيانات

الخلاصة

التَرئيةُ تُعَدّ معالجة بَعدية لعملية استكشاف المعارف,هدفُها عَرض المعارف المُستكشفة،تَسهيل فَهمها من قِبل المُستخدم,إعطاء نظرةً تجريديةً لهاو خَفض حجمِها.عملياتُ ما قبل تَعدين البيانات وعمليات التَعدين في عمليةِ الاستكشافِ تَبدو وكأنها عملياتُ قَبلية أو عملياتُ تحضيرية لماكنة التَرئية. التَرئية عملية صعبة لحاجتها إلى صياغةٍ رياضيةلقواعد معقدة لتحويل البيانات من صيغتها النصية عادةً إلى صيغ رسومية. هذا البحث يقدم اصطلاحاً جديداً هو *الاستكشاف المرئي للمعارف في قواعدِ البيانات* لتحويل عملية الترئية إلى أداةٍ تخدم عملية الاستكشاف في جميع أطواره التعدينيةِ وقبليةِ وبعديةِ التعدين لتساعد في فهم وتجريد وانتخاب وتنقية المعارف والبيانات وأدراك كَنَهِ خوارزميات هذه الأطوار. لبرهنة إمكانية الاستكشاف المرئي والاستفادة منه, اختيرت المهمة الأصعب في تعدينِ البيانات لتطبيق هذا المفهوم عليها, إلا وهي استكشاف قواعد الارتباط. إن اختيار تقنية مناسبة للاستكشاف ليس بالأمر السهل الهين,لذا فان عدة خوارزميات قد اقُترحت في هذه البحث ولأطوار الاستكشاف المختلفة, فقداقُترحت الصيغة العمودية والأفقية ومخططات الانشطار الثنائي لترئية قواعد البيانات الخاضعة للتعدين,إقُترِ حَت المخططات الشبكية لترئية قواعد مجاميع العناصر المتكررة باعتبارها مرحلة وسطية للتعدين. وتم تبني تقنيات المصفوفة الثنائية الأبعاد، المخططات الموجهة، وقاعدة-عنصر لترئية قواعد الارتباط المستكشفة.

[1] Al-Rafidain University College/Software Engineering, Baghdad, Iraq.

1. Introduction

The ***Knowledge discovery Process (KDP)***, also called knowledge discovery in databases, is a new field depending on ideas from statistics, machine learning, databases, parallel computing, computer graphics, data visualization, and other fields. KDP systems generally use methods, algorithms, and techniques from all of these fields. It has been materialized due to the extraordinary growth of data in all specialties of human activities, disability of database management system (DBMS) to extract hidden knowledge in databases, and the need for economic and scientific tools to extract such knowledge. KDP includes techniques and tools to address this need.

KDP seeks new knowledge in some application domain. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1].

KDP consists of many steps such as **Data selection, Cleaning, Enrichment, Coding, Data mining, and visualization**. The first four steps are called **pre-mining**, and the last step, **visualization**, is called **post mining**. Figure (1) abstracts the KDP process.

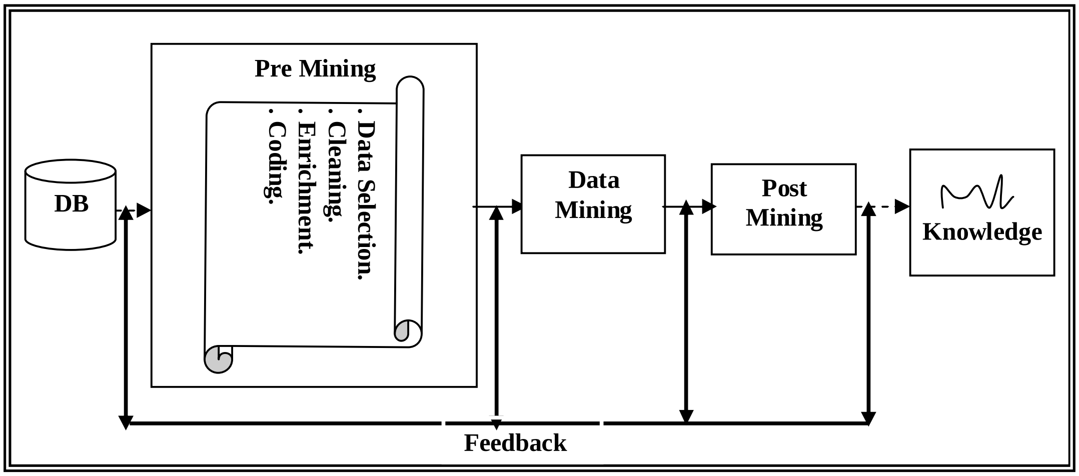


Figure (1) the stages of KDP

The pre-mining stage is responsible for selecting a suitable set of data for the mining, cleaning it from noisy, erroneous, and polluted data, strengthening the contents of the selected set, and transforming data to mining-algorithm-based format. Post-mining stage, i.e., Visualization, consumes 15% or more out of 85% of the time and efforts which are usually consumed by the pre and post-mining stages [2, 3].

Human beings look for structure, features, patterns, trends, anomalies, and relationships in data. Visualization supports this by presenting the data in various forms with differing interactions. Visualization can provide a qualitative overview of large and complex data sets, summarize data, and assist in identifying regions of interest and appropriate parameters for more focused quantitative analysis. In an ideal system, visualization harnesses the perceptual capabilities of the human visual system [4, 5]. Figure (2) shows the relation of data mining and visualization stage [5]. In spite of the benefits of visualization, it is not utilized in KDP with the exception of the visualizing of the discovered knowledge [1, 5]. It is a useful research trend to simplify the understanding, analyzses, abstracting of the data, information and knowledge that are attained from the different stages of KDP. Also, the graphical representation of intermediate result can simplify understanding the behavior of mining algorithms. From this simple introduction one can conclude that there are no obstacles in using the visualization techniques to support all the steps of KDP. And this may make a qualitative outburst in KDP process to be Visual KDP to attain the above advantages of visualization in addition to achieving of a high system transparency.

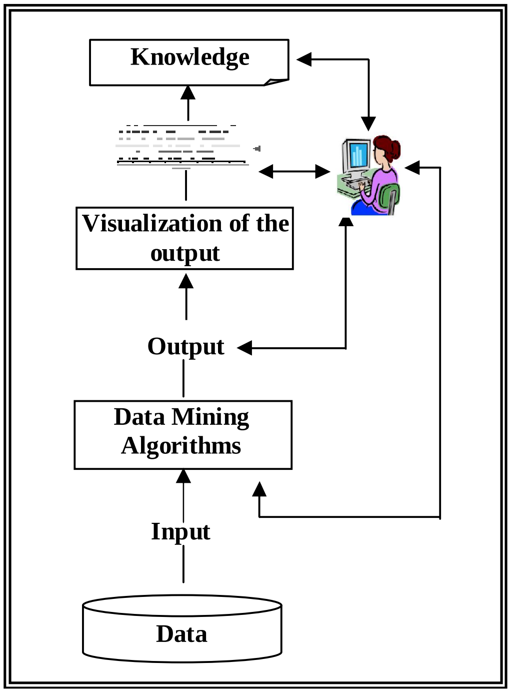


Figure (2) Visualization and Data Mining

This research presents a new viewpoint for the complementary relation of data mining and visualization by forcing visualization processing in different KDP stages starting from input databases to mined knowledge. This concept helps the integration of KDP and visualization techniques and presents a new term; i.e. visual KDP, VKDP. To verify the possibility of accomplishing a VKDP. The mining steps of Association Rules (AR) are chosen as a case study for the visualization due to the importance of ARs and their mining complexity. This imposes suggestions of many techniques to visualize the input database, mined itemsets, and mined rules.

2. The Proposed KDP Approach

In this section, the concept of visual KDP, VKDP, is proposed, which connects the visualization stage to all the steps of KDP to attain the benefits of the visualization technique.

As previously mentioned, KDP consists of three stages; pre-mining, data mining, and post-mining stage. In turn, pre-mining stages are composed of many complicated steps, such as data selection, cleaning or cleansing, enrichment, and coding. These stages are not working in a trajectory manner; there is a feedback from a step to another which enhances the performance of KDP process. And it is not a witticism that data mining techniques may be used in pre-mining steps, for example, to predict missing values or to determine a corrupted data. Also, pattern recognition may be used to classify a corpus of data values to a specified or unspecified category before the beginning of the mining step.

Usually, KDP or DM deals with a large volume of data, therefore, many algorithms for sampling are required to choose a suitable subset which substitutes the presence of all the data. Coding is the process of reformulating the data layout to be in a format that can be processed by the mining algorithm. Data mining stage is the real discovery process. It is a mechanized process of identifying or discovering useful structure in data. The term structure refers to data pattern, models, or relations over the data. A pattern is a description of a subset of data points. A model is a description of the entire data set. A relation is simply a property specifying some dependency between fields over a subset of the data.

Visualization stage depends on a visualization techniques and human cognition to identify the mined structure. Therefore, conveniently data mining seemed to be a tool or preprocessor for visualization engine which provides the structures to be visualized. Visualization in turn harnesses the perceptual capabilities of humans to provide visual insight into structure. In spite of the important role of the visualization, the data mining- visualization scenario is still unchanged. Recall figure (1) and figure (2) which depict this scenario.

This research suggests another scenario which deals with what has been called visual KDP, VKDP. VKDP can be accomplished by actual connection of KDP stages and visualization techniques and in this way the input, intermediate, and final result can be visualized. We think that there are many benefits will be gained from this scenario such as:

a- Visualization has been related to visualization of large volume of structure. The basic assumption is that large and normally incomprehensible amount of data can be reduced to a form that can be understood and implemented by human through the use of visualization techniques. Visualization helps widely in such data abstraction and understanding.

b- It helps in specifying the required fields and/or attributes for mining because it provides the graphical representation of the data which stimulates human cognition capabilities.

c- It eases the detection of the polluted data, missing values, and incorrect data; therefore, it speeds up the process of cleansing step.

d- If there is a suitable technique capable of visualizing more than one source of data, it is expected that VKDP will help widely in the enrichment process.

e- It assists the expert users of mining systems to determine the suitable data coding format according to the mining algorithm.

f- It elucidates the progress of mining process because it visualizes the intermediate and final results of it. Also, it is expected that VKDP eases determining the mining algorithm due to the obtained neatness, understandability, and analyzability of undermining database.

g- VKDP enables the user to visually tune the mining user-defined thresholds which affect the type, quality, and quantity of discovered knowledge.

Logically the above mentioned benefits enable the expert user to make an appropriate decision. Figure (3) depicts the proposed architecture of VKDP in which all the stages are connected to the visualization module which visualize the input and output of each step, while figure

(4) shows the suggested connection between DM and visualization module.

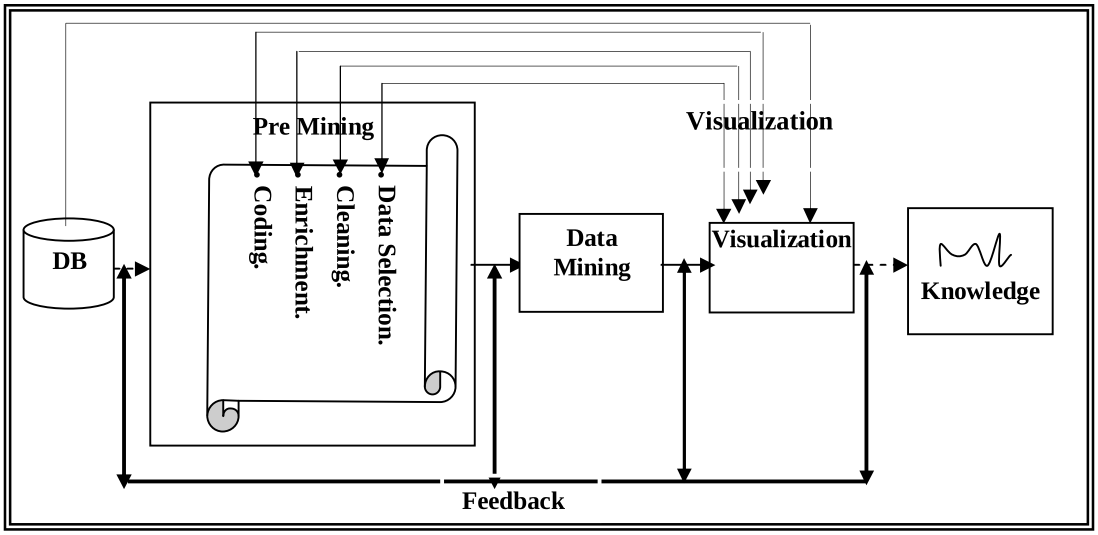


Figure (3) The VKDP Process Architecture

To prove and elucidate the concept of VKDP, different visualization techniques are proposed or adopted to design a visual association rule mining system, VAR. AR is chosen for this illustration because it is the most recent and important task of KDP. Its mining complexity allows explaining the concept of VKDP, and the limited success in the field of AR visualization.

The proposed techniques to visualize database under mining, the frequent itemsets database, and the database of extracted AR are presented in sections (3.1, 3.2, 3.3) respectively. Note that the proposed techniques are graph based techniques such as bipartite graph, concept, context, and lattice mathematics. Figure (4) explains the suggested connection between DM and visualization module, so that it is built to graphically present the input database and the result knowledge of DM module. Also, this process can be generalized if there is an intermediate result attained from the algorithm.

3. Proposed Visualization Technique of AR

The AR is selected to verify the concept of VKDP; therefore, it is required to suggest many techniques to visualize the structures directed between the modules of the AR miner, such as the input database, itemset database, and AR database. The next section presents the suggested visualization technique. Accordingly, we divide the step of VAR to three levels; database level, itemsets level, and association level. Note that the literatures of AR are not presented in this research to avoid overlaboring, but the reader can reference to many excellent references such as [1, 6, 7, 8].

3.1 Database Level

This level is related to visualize the transaction databases. **Bipartite graph** is suggested as a new technique to visualize the transaction database.

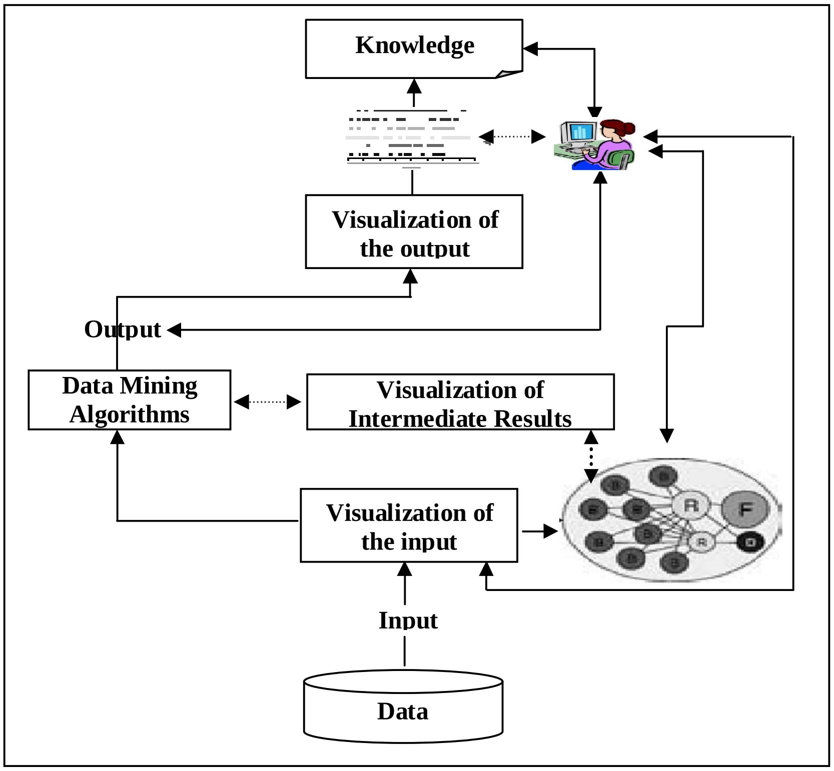

Knowledge
Visualization of the output
Output
Data Mining Algorithms
Visualization of Intermediate Results
Visualization of the input
Input
Data

Figure (4) Suggested Visualization and Data Mining Connection

Bipartite Graph

This section presents the concept of bipartite graph and how we can prove that a bipartite graph can be used as a useful tool to visualize the transaction databases.

A graph is called *bipartite* if the set of its vertices can be represented as a union of two disjoint sets such that no two nodes of the same set are connected by an edge. i.e., edges only go from the nodes of one set to the nodes of another [9]. The following is a mathematical definition of bipartite graph:

A simple undirected graph $G=(V,E)$ is called bipartite if there exists a partition of the vertex set $V=V1\cup V2$ such that both $V1$ and $V2$ are independent sets. One often writes $G=(V1+V2, E)$ to denote a bipartite graph whose partition has the parts V1 and V2. If that $|V1|=|V2|$, that is, if the number of elements in V1 is equal to the number of elements in V2, then G is called a *balanced* bipartite graph [10, 11].

Bipartite Graph As Proposed Visualization Technique

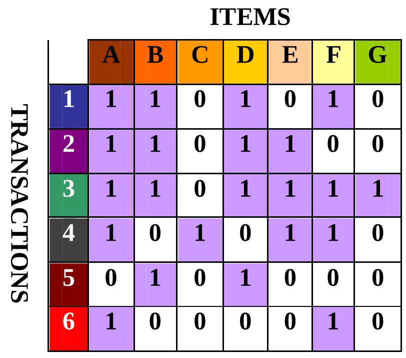
Recall that a bipartite graph G=(V,E) where V is the set of vertices and E is the set of edges. V must consist of two independent sets of vertices, V1 and V2. $V1=(v_{11}, v_{12}, \dots, v_{1n})$ and $V1=(v_{21}, v_{22}, \dots, v_{2m})$, so that $V1\cap V2=\Phi$. Remember that edges represent a relation between the vertices; such that there is no relationship between two vertices in the same set, but there are relationships between the vertices of V1 and V2.

Recall the definition of transaction database, there is unique set of items and unique set of transactions. Each transaction is defined by its Transaction Identifier, TID. A transaction contains one or more items. A transaction cannot contain another transaction and no item can contain another item.

So, we can represent the set of transactions as T={$TID1,TID2,\ldots TID_n$} and the item set I={$i1$, $i2$,…, i_m} and the relations, edges, which connect between T and I have the form E= "TID_t *contains item* i_i".

In other words T and I are independent sets, T∩I= Φ. Therefore we can write a transaction database as *D=({T+I}, E)* submitted to the same definition of the bipartite graph *G=({T+I}, E)*. .

Accordingly, D can be drawn as a bipartite graph. This taming of bipartite graph to transaction database layouts will be used as proposed technique to visualize the transaction database.



ITEMS

TRANSACTIONS

	A	B	C	D	E	F	G
1	1	1	0	1	0	1	0
2	1	1	0	1	1	0	0
3	1	1	0	1	1	1	1
4	1	0	1	0	1	1	0
5	0	1	0	1	0	0	0
6	1	0	0	0	0	1	0

Figure (5) transaction database with Maximal Unit Sub-Matrix of a Binary Matrix

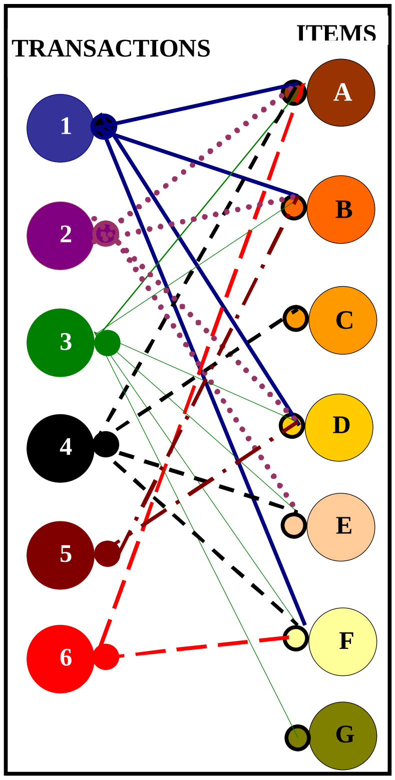


Figure (6) Maximal constrained Bipartite Clique

For example, the input database for association mining is essentially a very large bipartite graph, with *I* as the set of items, *T* as the set of TIDs, and each (item, tid) pair as an edge. The problem of enumerating all (maximal) frequent itemsets corresponds to the task of finding all (maximal) constrained bipartite cliques, $I_{subset} \times T_{subset}$, where $I_{subset} \subseteq I$, $T_{subset} \subseteq T$, *and* $|T_{subset}| \geq$ min-sup. There is one-to-one correspondence between bipartite graphs, binary matrices and hypergraphs [9], therefore, we can also view it as the problem of finding all (maximal) unit sub-matrices in a binary matrix, or as the problem of enumerating all (minimal) transversals of a hypergraph, satisfying the support constraints. Figure (5) and figure (6) show a database as a binary matrix, and as a bipartite graph, respectively. They also show the maximal unit sub-matrix and maximal constrained bipartite clique {A, B, D, E}×{1, 3, 5} (i.e., the maximal frequent itemset ABDE).

3.2 Itemsets Level

This section presents a novel representation for the itemsets database; the output of the first stage of AR mining. We'll domestic the *concept and context* definitions to itemset's definition. Then the set of all itemsets will be regarded as a set of *concepts*.

Concept and Context Definitions

A ***concept*** is considered to be determined by its *extent* and its *intent*: the extent consists of all objects belonging to the concept (as the reader belongs to the concept 'living person') while the intent is the collection of all attributes shared by the objects (*as all living persons share the attribute 'can breathe'*) [11]. Recall figure (5), suppose that the database presented in Figure (5) represents six nominees for world's beauty queen competition with their characteristics, (items). The abbreviations of the characteristics are illustrated in table (1) so that the nominee#1, (TID#1), has the features A, B, D, and F, i.e., lady#1 is virgin, has svelte body, owns protruding rear, and possesses round tit.

Table(1) Items' meaning of database presented in Figure(5)

Abbreviation (Item)	Plain Characteristic
A	*Virgin*
B	*Svelte body*
C	*swordlike eyelashes*
D	*Protruding rear*
E	*Snowy teeth*
F	*Round tit*
G	*Cultured and multilingual*

Now, the objects are the ladies numbers while the attributes are the seven indicated beauty attributes. That the *i*th lady, (object), possesses the *j*th attribute is indicated by an × in the *ij*-position of the database table. A concept of this context will consists of an ordered pair (E, I), where E (the extent) is a subset of the six ladies and I (the Intent) is a subset of the seven properties. To demand that the concept is determined by its extent and by its intent means that I should contain just those properties by all the ladies in E and similarly, the ladies in E should be precisely those sharing all the properties in I. A simple procedure for finding a concept is as follows: take an object, say lady#1 and let I be the set of attributes which she possesses, in this case *I={virgin, svelte body, Protruding rear, Round tit }*, then let E be the set of all ladies possessing all the attributes in I, in this case *E={1, 3}*. Then (E, I) is a *concept*. More generally, it is possible to begin with a set of objects rather than a single object. Concepts may also be obtained via a similar process commencing with a set of attributes. A concept (A_1, B_1) is regarded as being '*less general*' than a concept (A_2, B_2) if the extent A1 is contained in the extent A2. Thus an order is defined on the set of concepts by [12, 13]: $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2$.

This order is very clear when we tam it to the itemsets principles, for example, recall figure(5), *({A,B},{1,2,3)* $\subseteq$ *({A,B,D,E},{2,3})* $\Leftrightarrow$*{A,B}*$\subseteq$ *{A,B,D,E}* $\Leftrightarrow$ *{1,2,3}* $\supseteq$*{2, 3}*. This means that when an itemset is subset of another itemset, then its tid set is super set of the tid set of the second one.

A ***context*** is a triple (G, M, I) where G and M are sets and I $\subseteq$ G × M. The element of G and M are called *objects* and *attributes* respectively. *(g, m)* $\in$ *I* we write gIm and say '*the object g has the attribute m*'. A context can be represented by a *cross table* [12]. To tame and generalize these definitions to the AR problem we regard G as the set of Transactions Identifiers (TIDs), i.e. the ladies numbers in the database of figure (5); {1,2,3,4,5,6} and M is regarded as the set of attributes; Items, {A,B,C,D,E,F,G}. Hence one can say '*lady#1 has the attributes {virgin, svelte body, Protruding rear, Round tit}*'. This fact belongs to I. Hence the database presented in figure(5) represents a context. To complete the domestication of concept and context to the problem of AR mining, let’s present the following basic definitions which are adopted from [11].

Definition 1. For a set A$\subseteq$G of objects we define A' := {m $\in$ M | g I m $\forall$ g $\in$ A}(the set of attributes familiar to the objects in A). In the same way, for a set B of attributes we define B' := {g $\in$ G | g I m $\forall$ m $\in$ B}.(the set of objects which have all attributes in B).

Definition 2. A *formal concept* of the context (G,M,I) is a pair (A,B) with A$\subseteq$G, B$\subseteq$M, A’=B and B' = A. We call A the *extent* and B the *intent* of concept (A,B). ***B****(G,M,I)* denotes the set of all concept of the context (G,M,I) .

Let (G,M,I) be a context and (A1,B1) and (A2,B2) are two concepts in ***B****(G,M,I)*. One can write (A1,B1) ≤ (A2,B2) , if A1$\subseteq$A2 which implies that A’1$\supseteq$A’2, and the reverse implication is valid too, because A’’1=A1 and A’’2=A2. Therefore, one can conclude that $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2$. Hence, the relation ≤ is an order on ***B****(G,M,I)* and (***B****(G,M,I)*, ≤) is a complete lattice [11]. So, concept lattice is two lattice connected together in essence. The diagram of the lattice can be generated by use of the partial order relation. If $(A_1, B_1) < (A_2, B_2)$ and there is no other elements such that $(A_1, B_1) < (A3,B3) < (A_2, B_2)$, there is an edge from *(A_1, B_1) to (A_2, B_2)*. It reveals the generalization/specialization relationship between the concepts and could be used as an efficient tool for data mining and knowledge acquisition.

Before introducing the algorithm, we give some definitions and notations.

Let **K** =(G,M,I) be a formal context., for an object g$\in$G , we write g' instead of {g}’ for the *abject intent {m* $\in$*M | g I m }* of the object *g*. Correspondingly, *m*’*: ={ g* $\in$*G | g I m}* is the *attribute extent* of the attribute *m*. A *Basis* B is the set of all attribute extent of **K,** i.e., $\mathrm{B} = \{m' | m \in M\}$. We denote by F_B the family generated by joint form the basis B, i.e., $F_{\mathrm{B}} = \{\bigcap_{m' \in I} m' | I \in 2^{\mathrm{B}}\}$. For each F $\in$$F_{\mathrm{B}}$, we denote by γ *(F)* the subset of M, such that for each element m in it , F is included in m’ , i.e. $\gamma(F) = \{m \in M \mid F \subseteq m'\}$.

For example, in the Context of Figure (5) , *B* = {a’={12346}, b’={1235}, c’={4}, d’={1235}, e’={234}, f’={1346}, g’={3}}, F_B = {{123456}, {12346}, {1235}, {4}, {1235},{234}, {1346}, {3},{123}, {4}, {123}, …,Φ } and *{* γ *(F)*

| $F \in \mathcal{F}_B$ } = {Φ,{a},{b},{c},{d},{e},{f},{g},{ab}, {ac}, {ad}, …, {abcdefg}}

The following theorem can be obviously derived from the above definition.

Theorem 1: Let **K=** (G, M, I) be a formal context, for each $F \in \mathcal{F}_B$, (F, γ (F)) is a concept of **K,** and **B** ≡ {(F, γ (F))| $F \in \mathcal{F}_B$ } = **B**(G,M,I).

Now our problem is constructing the concepts set from the given context **K**=(G, M, I). The proposed algorithm can be complete by the following three steps:

(1) Compute the basis of the context B;

(2) Generate the family **B =** {(F, γ (F))| $F \in \mathcal{F}_B$ };

(3) Construct the concepts set from **B.**

Figure(7) present a novel algorithm to visualize the mined itemsets database depending on the proven affinities of itemset, concept, and extent definitions.

Proposed Algorithm to Visualize The Itemsets Database

After showing the one-to-one correspondence of transaction database and context cross-table and one-to-one correspondence of itemset×TIDset and concept definition, this section presents a proposed algorithm to draw the set of concepts, i.e., the database of itemsets. This algorithm is named Itemsets visualization Algorithm, (IVA).

To explain IVA figure(8) is given to illustrate it depending on the database presented in figure (7). At the beginning there are three maximal columns; column a', b', and d'. When a' has been eliminated from the table the columns e' and f' become maximal and the proposed algorithm can manipulate any one of b', d', e' and f' in the subsequently step. It is useful to handle the old maximal column before considering the new one which indicated by bold horizontal lines in the Attribute/Extents table. According to the step 1.2.2, the row [b, d | STUW] match b'=d'={STUW}. The row [g|U] was dropped when U already available in the list. Also, according to step 1.2.2, the lable g was added to the Attribute cell of the previous occurrences.

4.3.4 Rule Level

Three approaches are used to visualize the mined AR, which are the *two-dimensional matrix*[14], *directed graph* [15], and *rule-item approach*[16].

4. Discussion, Conclusions, and Future Works

4.1 Discussion and Conclusions

In this research, we catch many birds by one brick, but its chest is suggesting of the VKDP approach and taming the problem of AR mining to the bipartite, context, and concept mathematics. To summarize and discuss the research, its contributions will be divided into two divisions:

i- A new look for KDP architecture is suggested which emphasizes on regarding the visualization step as an interactive process needed for visualizing the data to be mined, pre-mining step, intermediate result of mining algorithm, and the mined knowledge. In this way the visualization step became a complementary step for all the steps of KDP, furthermore, visualizing the input data. We called this architecture as ***visual KDP (VKDP).*** Many VKDP benefits were listed and discussed in this research.

ii- To confirm the concept of visual KDP, we proposed many techniques to visualize each step of AR mining. These proposed techniques are:

a- Visualization of transactional database: A novel techniques is presented depending on a bipartite graph to represent the input database, transaction database. There are many benefits are gained from this technique some of these are:

- Visualization of input databases resizes the understanding of the content of the data and the expert user can easily determined its nature i.e. *dense* or *sparse*. Determining database nature is very useful in specifying the suitable algorithm for mining, because some algorithm are not suitable for dense database and vise versa.
- Visualizing of input database diminished the time and efforts required for data selection step of pre-mining stage which influence the positively the next steps such as cleaning, enrichment, and coding.
- This representation reflects the nature of the database because it shows the arcs related to each item and transaction.
- It shows the support of each item without any counting or accessing to the original database. It shows the two relations of transactional database that are *"the transaction X contains the set Y of items*", and *"the item X is included in the set of transactions Y*" that refer to the vertical and horizontal layout respectively. Hence, this research fabricates a multi-layout representation for transactional databases.
- It converts the database to mathematical model and one can apply bipartite theories to obtain new results which will release new researches in KDP field.

b- Intermediate Results Visualization: The '*concept*' and the network of concepts are presented as a novel technique to represent item sets and transaction sets. The gained benefits of this representation are listed below:

- It enables the user to determine suitable *minconf* and *minsup* thresholds.
- It enables the user to determine the frequent itemset and maximal itemsets which easies the process of infrequent itemsets pruning.
- It depicts the progress of algorithm execution and shows the intermediate result of the processing.
- It enables the expert user to determine the maximal itemsets that enable the user to determine the frequent itemsets according to the fact that is "all the subsets of frequent itemset are frequent".

- It enhances the performance of AR extractor by excluding some of uninteresting rules.
- The main benefits of this visualization are converting the mined itemsets to mathematical model which will open new era of research and techniques in fields of data mining.

In spite of the concluded and listed benefits above, the implementation of the proposed techniques is not easy work and we spent considerable efforts and time to accomplish the VKDP architecture especially the proposed algorithm of visualization of itemsets which has an $O((|G|+|M|)*|M|*|F_B|)$ time complexity. Also, we think that the output of this algorithm is very useful for the familiar and expert data mining users, but the decision makers require some training or explanation from the mining users.

```
Algorithm: Itemsets visualization Algorithm D (TID, Item)
Input: Horizontal Transaction Database
//Cross table of a context, the TID set represents G, while the Items represent
        the attributes M.
Output: visual Associative set of concepts
{
    Step 1: Find all the itemsets of context D.
            1.1 Construct a table with two columns headed attributes and
                    Extents.
                  Insert G in 1st cell of extent column and Leave the 1st cell of
                      the attribute column empty;
            1.2 Find a maximal attribute-extent, i.e., items-tids, say m'.
                  1.2.1 If the set m' is not already in the extents column, add the
                          row [m|m'] to the table. Intersect the set m' with all
                          previous extents in the extents column. Add these
                          intersections to the extents column (unless they are
                          already in the list) and leave the corresponding cells in
                          the attribute column empty.
                  1.2.2 If the set m’ is available in Extents, the label m will be
                          added to the attribute cell of the row where m’
                          previously occurred.
            1.3 Remove the column below m from the table.
            1.4 If the last column has been deleted, stop, else goto (1.2).
    Step 2: Graph drawing with m and m' labels.
            2.1 At the top of the graph draw a point labeled with G.
            2.2 Use the list of Extents in the table obtained from Step#1, for
                    each set S in the list draw a fittingly positioned point in
                    the graph. Label the point of S with element of it. If S is
                    an attribute-extent, add the label m above the point of S.
    Step 3: Graph redrawing with g and m labels
            3.1 Redraw the graph; add the m labels as in the first graph.
            3.2 For each g in G, write a label g below the point on the graph
                    which has the minimal extent including the object g.
}
```

Figure (7) The proposed Itemsets Visualization Algorithm, IVA

4.2 Future Works

Many researches can be emerged from this research in future, some of these are:

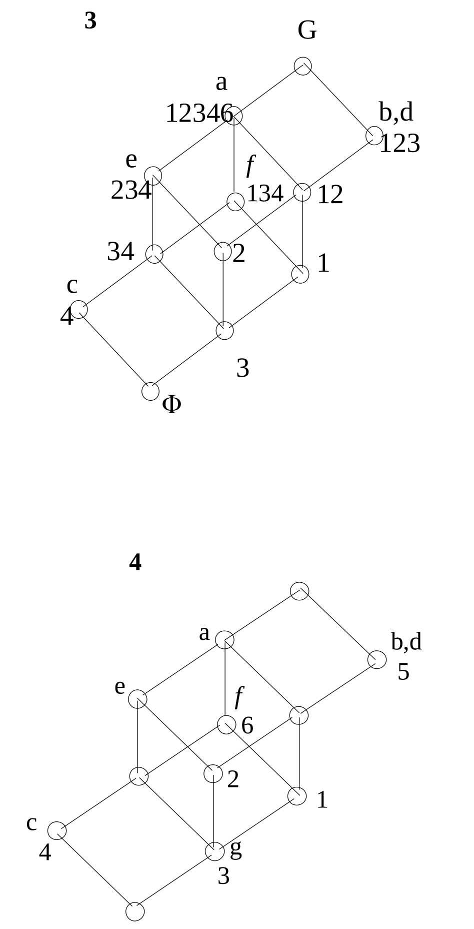
- Suggesting of visualization techniques for the rest of KDP and mining tasks such as classification, clustering, regression, etc.
- The research converts the transactions and itemsets databases to mathematical models depending on bipartite graph and concept; therefore we believe that these models will be a corner to producing new mining algorithms in this field.

1

	A	b	C	d	e	f	G
1	1	1	0	1	0	1	0
2	1	1	0	1	1	0	0
3	1	1	0	1	1	1	1
4	1	0	1	0	1	1	0
5	0	1	0	1	0	0	0
6	1	0	0	0	0	1	0

2

Attribute	Extents
	G
A	12346
b,d	1235
	123
E	234
	23
F	1346
	13
	34
g *(added later)*	3
C	4
	Φ
G	3



Figure(8) Concepts Net of Database Presented In Figure (5)

References

[1] Krzysztof J. Cios, Witold Pedrycz, `Roman W. Swiniarski, Lukasz A. Kurgan, "***Data Mining: A Knowledge Discovery Approach***", Springer, 2008.

[2] Usama M. Fayyad,"***Advances In Knowledge Discovery and Data Mining***", MIT Press, 2006.

[3] Jiawei Han, Micheline Kamber, "***Data Mining Concepts and Techniques***", Morgan Kaufmann, 3rd ed., 2006.

[4] R. Redpath, B. Srinivasan, "***Criteria for Comparative Study of Visualization Techniques in Data mining***", IEEE 3rd int. Conf. On Intelligent System, Tulsa, USA, 2003.

[5] Usama M. Fayyad, G. Grinstein, "***Information Visualization in Data Mining and Knowledge Discovery***", Morgan Kaufmann, San Francisco (CA) 2005.

[6] Margaret H. Dunham, Yongqiao Xiao Le Gruenwald, Zahid Hossain, "***A SURVEY OF ASSOCIATION RULES***", Department of Computer Science and Engineering Department of Computer Science Southern Methodist University University of Oklahoma, 2004.

[7] Christian Borgelt, "***An Implementation of the FP-growth Algorithm***"**,** Department of Knowledge Processing and Language Engineering, School of Computer Science, 2004.

[8] S. Parthoasarorthy, W. Li, M. Ogihara, ***"New Algorithms for Fast Discovery of Assoication Rules"***, In 3rd Intl. Conf. on KDD and DM, Augest 2000.

[9] Armen S Asratian, Tristan M J Denley, Roland Haggkvist, "***Bipartite Graphs and Their Applications***", Cambridge University Press , 2009.

[10] Wen Luo-Sheng, Zhong Jiang and Yang Xiao-Fan, ***"Sexually Transmitted Diseases on Bipartite Graph",*** *Chinese Phys. Lett.*, 2009.

[11] Khee Meng Koh, "***Introduction To Graph Theory***", World Scientific Publishing Co., 2007.

[12] Steven Roman, "***Lattice and Order***", Springer, 2008.

[13] B. A. Davey, H. A. Priestley, "***Introduction to Lattices and Order***", 2nd ed., Cambridge University Press, 2002.

[14] D. Keim, "*Designing Pixel-Oriented Visualization Techniques: Theory and Applications*", Transaction on Visualization and computer Graphics 6, 1 (Jan- Mar 2000), 59-78.

[15] D. Bruzzese, C. Davino, "*Visual Post-Analysis of Association Rules*", Dept. of Mathematics and Statistics, University of Naples Federico, Italy, {dbruzzes, cdavino}@unina.it, 2002.

[16] P. C. Wong, P. Whitney, J. Thomas, "*Visualizing Association Rules for Text Mining*", Pacific Northwest National Laboratory, 2000.