CHAPTER 5

KNOWLEDGE INTEGRATION IN PARALLEL AND DISTRIBUTED DATA MINING WITH IMPROVED RESPONSE TIME

The processed information is obtained from the geographically distributed workstations in parallel. This knowledge is then integrated based on identifying the global exceptional patterns. The elapsed time between the end of an inquiry or request on a workstation and the beginning of a response is known as response time. An optimized response time for distributed data mining in parallel environment is broadly dependent on three factors:

1. Communication: The communication time is the time taken to transfer datasets or mobile agents carrying data mining primitive to remote workstations and the time taken to transfer results from remote workstations for integration.

2. Data Mining. This is the time taken to perform data mining on the distributed data sets and is a core factor irrespective of the DDM model.

3. Knowledge Integration. This is the time taken to integrate the results from the distributed datasets available in the workstations. The knowledge regarding the time taken to execute a data mining task at different distributed locations can be an effective basis for optimization of a DDM task. This research work concentrates on the rough sets algorithm to estimate application run times. This research work also presents
experimental evaluation of this technique to establish its estimation accuracy and validity.

5.1 Optimized Response Time

The architecture that is used for this research work is the client server model along with the agents as shown in Figure 5a.

![Client Server Architecture with agents](image)

Figure 5a: Client Server Architecture with agents

This model combines the best aspects of the agent model and the client-server approach by incorporating an agent framework with a dedicated data mining server. It brings with it the advantage of combining the concept of dedicated data mining resources. It also has the ability to circumvent the communication overheads associated with the client-server approach.

There are primarily three cost estimates that the optimizer needs to build and use in order to decide on the execution strategy for a distributed data mining task. The three estimates are:

- *Data transfer cost* estimates, which involve the time needed for the different datasets to be communicated to the geographically
distributed workstations and also the time involved by the agents to travel and interact.

- *Data mining cost* estimates, which is an estimate for performing the data mining task.
- *Knowledge integration cost* which estimates the time needed for integrating the knowledge (outcome of mined results) obtained from the distributed workstations with a global identification model.

Based on these three estimates the optimiser chooses the best option for a given data mining task. Therefore, the optimiser’s tasks involve building cost estimates and then choosing the best alternative. The design of the optimizer used in our research work is illustrated in Figure 5b.

![Figure 5b: Cost Computation](image)

**Planner.** The planner receives the task specification and information about environmental factor such as the network conditions. The planner is the controlling entity in the optimiser. The planner passes on the relevant
information to the different "calculators". Based on the different estimates received from the calculators, the planner devises the execution strategy for the parallel and distributed data mining task.

The response time is calculated based on

\[ T = t_{com} + t_{dm} + t_{ki} \]

where

- \( T \) is the response time
- \( t_{com} \) is the time involved in communication.
- \( t_{dm} \) is the time taken to perform data mining.
- \( t_{ki} \) is the time taken to perform knowledge integration.

### 5.2 XML Rule Mining Approach

Association rule mining from XML data has gained several techniques to solve this problem. The straightforward approach is to map the XML documents to relational data model and to store them in a relational database. This allows applying the standard tools that are used to perform rule mining from relational databases. Even though it makes use of the existing technology, this approach is often time consuming and involves manual intervention because of the mapping process. Due to these factors it is not quite suitable for XML data streams [WD, 03].

An XML document contains one root level element with the corresponding opening and closing tags. These tags surround all other data
content within the XML document. The format of a sample XML data used to test our algorithm is shown in Figure 5c. The transactions tag is the root element that contains many transaction elements. Each transaction element is uniquely identified by its id attribute. Each transaction element contains one items element which in turn contains many item elements. An item element has the name of the particular item in the given transaction.

- <CATALOG>
- <PLANT>
  <COMMON>Bloodroot</COMMON>
  <BOTANICAL>Sanguinaria canadensis</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Mostly Shady</LIGHT>
  <PRICE>$2.44</PRICE>
  <AVAILABILITY>031599</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Columbine</COMMON>
  <BOTANICAL>Aquilegia canadensis</BOTANICAL>
  <ZONE>3</ZONE>
  <LIGHT>Mostly Shady</LIGHT>
  <PRICE>$9.37</PRICE>
  <AVAILABILITY>030699</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Marsh Marigold</COMMON>
  <BOTANICAL>Caltha palustris</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Mostly Sunny</LIGHT>
  <PRICE>$6.81</PRICE>
  <AVAILABILITY>051799</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Cowslip</COMMON>
  <BOTANICAL>Caltha palustris</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Mostly Shady</LIGHT>
  <PRICE>$9.90</PRICE>
  <AVAILABILITY>030699</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Dutchman's-Breeches</COMMON>
  <BOTANICAL>Dicentra cucullaria</BOTANICAL>
  <ZONE>3</ZONE>
<LIGHT>Mostly Shady</LIGHT>
<PRICE>$6.44</PRICE>
<AVAILABILITY>012099</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Ginger, Wild</COMMON>
  <BOTANICAL>Asarum canadense</BOTANICAL>
  <ZONE>3</ZONE>
  <LIGHT>Mostly Shady</LIGHT>
  <PRICE>$9.03</PRICE>
  <AVAILABILITY>041899</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Hepatica</COMMON>
  <BOTANICAL>Hepatica americana</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Mostly Shady</LIGHT>
  <PRICE>$4.45</PRICE>
  <AVAILABILITY>012699</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Liverleaf</COMMON>
  <BOTANICAL>Hepatica americana</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Mostly Shady</LIGHT>
  <PRICE>$3.99</PRICE>
  <AVAILABILITY>010299</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Jack-In-The-Pulpit</COMMON>
  <BOTANICAL>Arisaema triphyllum</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Mostly Shady</LIGHT>
  <PRICE>$3.23</PRICE>
  <AVAILABILITY>020199</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Mayapple</COMMON>
  <BOTANICAL>Podophyllum peltatum</BOTANICAL>
  <ZONE>3</ZONE>
  <LIGHT>Mostly Shady</LIGHT>
  <PRICE>$2.98</PRICE>
  <AVAILABILITY>060599</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Phlox, Woodland</COMMON>
  <BOTANICAL>Phlox divaricata</BOTANICAL>
  <ZONE>3</ZONE>
  <LIGHT>Sun or Shade</LIGHT>
  <PRICE>$2.80</PRICE>
  <AVAILABILITY>012299</AVAILABILITY>
</PLANT>
- <PLANT>
  <COMMON>Phlox, Blue</COMMON>
  <BOTANICAL>Phlox divaricata</BOTANICAL>
  <ZONE>3</ZONE>
  <LIGHT>Sun or Shade</LIGHT>
  <PRICE>$5.59</PRICE>
  <AVAILABILITY>021699</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Spring-Beauty</COMMON>
  <BOTANICAL>Claytonia Virginica</BOTANICAL>
  <ZONE>7</ZONE>
  <LIGHT>Mostly Shady</LIGHT>
  <PRICE>$6.59</PRICE>
  <AVAILABILITY>020199</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Trillium</COMMON>
  <BOTANICAL>Trillium grandiflorum</BOTANICAL>
  <ZONE>5</ZONE>
  <LIGHT>Sun or Shade</LIGHT>
  <PRICE>$3.90</PRICE>
  <AVAILABILITY>042999</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Wake Robin</COMMON>
  <BOTANICAL>Trillium grandiflorum</BOTANICAL>
  <ZONE>5</ZONE>
  <LIGHT>Sun or Shade</LIGHT>
  <PRICE>$3.20</PRICE>
  <AVAILABILITY>022199</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Violet, Dog-Tooth</COMMON>
  <BOTANICAL>Erythronium americanum</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Shade</LIGHT>
  <PRICE>$9.04</PRICE>
  <AVAILABILITY>020199</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Trout Lily</COMMON>
  <BOTANICAL>Erythronium americanum</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Shade</LIGHT>
  <PRICE>$6.94</PRICE>
  <AVAILABILITY>032499</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Adder's-Tongue</COMMON>
  <BOTANICAL>Erythronium americanum</BOTANICAL>
  <ZONE>4</ZONE>

119
<LIGHT><b>Shade</b></LIGHT>
<PRICE>$9.58</PRICE>
<AVAILABILITY>041399</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Anemone</COMMON>
  <BOTANICAL>Anemone blanda</BOTANICAL>
  <ZONE>6</ZONE>
  <LIGHT><b>Mostly Shady</b></LIGHT>
  <PRICE>$8.86</PRICE>
  <AVAILABILITY>122698</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Grecian Windflower</COMMON>
  <BOTANICAL>Anemone blanda</BOTANICAL>
  <ZONE>6</ZONE>
  <LIGHT><b>Mostly Shady</b></LIGHT>
  <PRICE>$9.16</PRICE>
  <AVAILABILITY>071099</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Bee Balm</COMMON>
  <BOTANICAL>Monarda didyma</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Shade</LIGHT>
  <PRICE>$4.59</PRICE>
  <AVAILABILITY>050399</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Bergamot</COMMON>
  <BOTANICAL>Monarda didyma</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Shade</LIGHT>
  <PRICE>$7.16</PRICE>
  <AVAILABILITY>042799</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Black-Eyed Susan</COMMON>
  <BOTANICAL>Rudbeckia hirta</BOTANICAL>
  <ZONE>Annual</ZONE>
  <LIGHT>Sunny</LIGHT>
  <PRICE>$9.80</PRICE>
  <AVAILABILITY>061899</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Buttercup</COMMON>
  <BOTANICAL>Ranunculus</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Shade</LIGHT>
  <PRICE>$2.57</PRICE>
  <AVAILABILITY>061099</AVAILABILITY>
</PLANT>

120
- <PLANT>
  <COMMON>Crowfoot</COMMON>
  <BOTANICAL>Ranunculus</BOTANICAL>
  <ZONE>4</ZONE>
  <LIGHT>Shade</LIGHT>
  <PRICE>$9.34</PRICE>
  <AVAILABILITY>040399</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Butterfly Weed</COMMON>
  <BOTANICAL>Asclepias tuberosa</BOTANICAL>
  <ZONE>Annual</ZONE>
  <LIGHT>Sunny</LIGHT>
  <PRICE>$2.78</PRICE>
  <AVAILABILITY>063099</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Cinquefoil</COMMON>
  <BOTANICAL>Potentilla</BOTANICAL>
  <ZONE>Annual</ZONE>
  <LIGHT>Shade</LIGHT>
  <PRICE>$7.06</PRICE>
  <AVAILABILITY>052599</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Primrose</COMMON>
  <BOTANICAL>Oenothera</BOTANICAL>
  <ZONE>3 - 5</ZONE>
  <LIGHT>Sunny</LIGHT>
  <PRICE>$6.56</PRICE>
  <AVAILABILITY>013099</AVAILABILITY>
</PLANT>

- <PLANT>
  <COMMON>Cardinal Flower</COMMON>
  <BOTANICAL>Lobelia cardinalis</BOTANICAL>
  <ZONE>2</ZONE>
  <LIGHT>Shade</LIGHT>
  <PRICE>$3.02</PRICE>
  <AVAILABILITY>022299</AVAILABILITY>
</PLANT>
</CATALOG>

Figure 5c: XML Input File Format

The input XML document can have a very complicated structure, containing the transaction data at different depths. In this research work, the input document is preprocessed by using an XML style sheet language,
like XSLT, to convert it into a simply structured document format as shown in Figure 5d [Bru, 04]. This preprocessing is carried out quickly and easily.

```
INPUT FILE NAME=data.xml
OUTPUT FILE NAME=plants.xml
MINIMUM SUPPORT=0.6
MINIMUM CONFIDENCE=1.0
FIRST LEVEL ELEMENT NAME=CATALOG
SECOND LEVEL ELEMENT NAME=PLANT
THIRD LEVEL ELEMENT NAME=common
SECOND LEVEL ELEMENT UNIQUE
ATTRIBUTE NAME=not used
```

Figure 5d: XML Input General Format

The implementation outputs of this research work in the association rules in XML format is shown in Figure 5e. The root level element name is rules which contain zero or more rule elements. Each rule element has one antecedent and one consequent elements, and each rule has two attributes: support and confidence.

```
<rules>

<rule support="0.6" confidence="1.0">

<antecedent>

  <COMMON>Pimrose</COMMON>

</antecedent>

<consequent>

  <COMMON>Cardinal Flower</COMMON>

</consequent>

</rule>

...

</rules>
```

Figure 5e: XML Output File Format
5.3 Knowledge Integration

The knowledge integration is carried out when a particular company has multiple branches. Consider an organization which has many branches in different locations. Each branch has its own database, and the bank data is widely distributed and thus becomes a distributed issue. In Figure 5f, the top level is an interstate company (IC). This IC is responsible for the development and decision-making for the entire company. The middle level consists of n branches LB₁, LB₂, LB₃... LBₙ. The bottom level consists of n local databases DB₁, DB₂, DB₃,...DBₙ of the n branches.

![Diagram of IC, LB₁, LB₂, LBₙ, DW₁, DW₂, DWₙ]

Figure 5f: An interstate company and its branches

The local branches mining process is carried out by allocating the datasets based on the CPU idle time and the allocation algorithm. This (knowledge) mined result obtained from the distributed workstations is integrated based on the interestingness measures.
The knowledge discovery process takes the raw results from data mining and carefully and accurately transforms them into useful and understandable information. The mined outcome obtained from the distributed workstations in the form of knowledge for XML data is integrated based on the following schema.

![Diagram](image)

Knowledge Integration

Figure 5g: Schematic representation of knowledge integration

5.3.1 Integration by Identification of exceptional patterns

Let \( D_1, D_2, \ldots, D_n \) be \( n \) data warehouses in the \( n \) branches \( B_1, B_2, \ldots, B_n \) of an organization, respectively; and \( LI \) be the set of local patterns (local instances) from \( D_i (i=1,2,\ldots,n) \). The global exceptional patterns of interest in the local patterns are identified. Let \( LP_1, LP_2, \ldots, LP_n \) be the corresponding local patterns which are mined from every database. And \( \text{minsupp} \) be the user specified minimal support in the database \( D_i (i=1,2,\ldots,n) \). For each pattern \( P \), its support in \( D_i \) is denoted by \( \text{Supp}(P) \). The average vote of local patterns is given by

124
\[
\text{Average votes} = \frac{\sum_{i=1}^{\text{Num}(\text{GP})} \text{Num}(P_i)}{\text{Num}(\text{GP})}
\]

GP means the Global Patterns, set of all patterns from each data warehouse i.e. \( \text{GP} = \{ \text{LP}_1 \cup \text{LP}_2 \cup \ldots \cup \text{LP}_n \} \) and \( \text{Num}(\text{GP}) \) is the number of patterns in \( \text{GP} \) [CMW+04]. The global support of a pattern is given by

\[
\text{Supp}_G(P) = \frac{\sum_{i=1}^{\text{Num}(P)} \text{Supp}_i(P) - \text{minsupi}}{1 - \minsupi}
\]

\( \text{Supp}_G(P) \) is the global support of a pattern. This gives a method to compute the global patterns from the locally generated knowledge. The global knowledge could be based on the interestingness measures. Exceptional patterns reflect the individuality of branches within an interstate company. Identify exceptional pattern algorithm is used to search all the significant exceptional patterns from the given \( n \) local patterns.
Algorithm 1: IdentifyExPattern

Input: \(LP_i\): set of local patterns; \(\text{minsup}_i\): minimal support threshold in \(D_i\) \((i = 1, 2, \ldots, n)\);
Output: \(EP\): the set of exceptional patterns;

begin
(1) \(GP \leftarrow \{LP_1 \cup LP_2 \cup \cdots \cup LP_n\}\); \(CEP=\emptyset\);
(2) For each pattern \(P\) in \(GP\) do
\hspace{1cm} Count \(P\)'s votes, \(Num(P)\); And Record which database support it, using \textit{from} to note them.
\hspace{1cm} Calculate the average votes using Formula 1: \(\text{AverVotes} = \frac{\sum_{i=1}^{Num(GP)} Num(P_i)}{Num(GP)}\)
(3) For each pattern \(P\) in \(GP\) do
\hspace{1cm} if \((Num(P) < \text{AverVotes})\) \(CEP = CEP \cup P\)
(4) For each candidate exceptional pattern \(P\) in \(CEP\) do
\hspace{1cm} \(Suppc(P) \leftarrow \frac{\sum_{i=1}^{Num(P)} Supp_i(P) - \text{minsup}_i}{1 - \text{minsup}_i}\)
(5) Rank all the patterns \(P\) in \(CEP\) by their \(Suppc(P)\);
(6) Output the high rank patterns in \(CEP\) and the databases which support them;

End.

Figure 5h: Identification of exceptional patterns algorithm

Step 1: Generates the set of patterns from each data warehouse.

Step 2: Counts each pattern’s votes and the average votes of patterns

Step 3: Generates the candidate exceptional patterns

Step 4: Calculate all the candidate exceptional patterns by their \(Suppc(P)\) values.

Step 5: Rank the candidate exceptional patterns by their \(Suppc(P)\) value.

Step 6: Output all the exceptional pattern which satisfy the users requirement and have high rank.

126
Based on this algorithm the patterns are taken and globalized integrated result is obtained by eliminating the uninteresting patterns. Exceptional patterns reflect the individuality of branches within an interstate company.

Consider 5 datasets D_1, D_2, D_3, D_4, D_5 and their corresponding patterns is in the following. Patterns are denoted by A-F, the value after each colon is the pattern’s support \( \text{minsup}_1 = 0.49, \text{minsup}_2 = 0.48, \text{minsup}_3 = 0.82, \text{minsup}_4 = 0.20, \text{minsup}_5 = 0.15 \) are 5 databases’ minimal support respectively.

\[
\begin{align*}
LP_1 &= \{(A: 0.69) ; (C: 0.68) ; (F: 0.52)\} \\
LP_2 &= \{(A: 0.50); (B: 0.62); (C: 0.91); (E: 0.82); (F: 0.76) ; (G : 0.86)\} \\
LP_3 &= \{(A: 0.87) ; (C : 0.85); (D: 0.86) ; (E : 0.86); (F : 0.95)\} \\
LP_4 &= \{(B : 0.36) ; (C : 0.31) ; (E : 0.28)\} \\
LP_5 &= \{(E : 0.22)\}
\end{align*}
\]

The algorithm Identify-ExPattern is then used in this research work to search all the exceptional patterns from the given local patterns. According to Step (1) and Step (2), \( \text{GP}=\{A,B,C,D,E,F,G\} \), and the Avervotes=18/7=2.57. Because Pattern B,D,G have less votes than the Avervotes. After pruning by Avervotes \( \text{CEP}=\{B,D,G\} \). The \( \text{Supp}_G(P) \) value of each pattern in CEP are shown as follows.

\[
\begin{align*}
\text{Supp}_G(B) &= 0.235, \text{ Pattern B comes from } \{D_2, D_4\} \\
\text{Supp}_G(D) &= 0.222, \text{ Pattern D comes from } \{D_3\} \\
\text{Supp}_G(G) &= 0.73, \text{ Pattern G comes from } \{D_2\}
\end{align*}
\]
It is obvious that pattern \( \{G\} \) has the highest global support and it is supported by only a database. So it can be regarded as an exceptional pattern. After finding such exceptional patterns, the head company can use the patterns to assist making special decision for the corresponding subsidiary company.

5.3.2 Integration by Bayes networks

A Bayesian network is a graphical model for probabilistic relationships among a set of variables. Over the last decade, the Bayesian network has become a popular representation for encoding uncertain expert knowledge in expert systems. More recently, researchers have developed methods for learning Bayesian networks from data. In this work the Bayesian techniques for extracting and encoding knowledge from data is dealt.

Finding all attribute sets with given minimum interestingness

<table>
<thead>
<tr>
<th>Input: Collection of attributed sets ( C ), Bayesian network ( BN ) over attributes ( Z ).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Output:</strong> Distributions ( P_i^{BN} ) for all ( I \in C ).</td>
</tr>
<tr>
<td>1. Let ( S \leftarrow \text{Bd}^+ (C) ).</td>
</tr>
<tr>
<td>2. While ( S \neq \emptyset );</td>
</tr>
<tr>
<td>3. Let ( I \leftarrow ) an attribute set from ( S ).</td>
</tr>
<tr>
<td>4. For all ( A ) in ( Z \setminus I );</td>
</tr>
<tr>
<td>5. Compute gain ( (I \cup {A}) ).</td>
</tr>
<tr>
<td>6. Pick ( A^* ) for which the gain in step5 was maximal.</td>
</tr>
<tr>
<td>7. If gain ( (I \cup {A^*}) &gt; \text{gain}(I) ) then</td>
</tr>
<tr>
<td>8. Let ( I \leftarrow I \cup {A^*} ).</td>
</tr>
<tr>
<td>10. Compute ( P_i^{BN} ) from ( BN ) using bucket elimination</td>
</tr>
<tr>
<td>11. Compute ( P_j^{BN} ) for all ( J \in S, J \notin I ).</td>
</tr>
<tr>
<td>12. Remove from ( S ) all attribute sets included in ( I ).</td>
</tr>
<tr>
<td>13. Compute ( P_i^{BN} ) for all ( J \in C \setminus \text{Bd}^+ (C) ).</td>
</tr>
</tbody>
</table>

Figure 5i: Identification of attribute sets
The above algorithm is used for finding all attribute sets with interestingness greater than or equal to a specified threshold $\varepsilon$ given a dataset $D$, and a Bayesian network $BN$.

In the second stage all itemsets frequent in the Bayesian network are found, and their joint probability distributions in the data are computed using an extra database scan. To find all itemsets frequent in the Bayesian network the Apriori algorithm is used.

\begin{center}
\begin{tabular}{|l|}
\hline
\textbf{Input:} Bayesian network BN, minimum support $\varepsilon$.  \\
\textbf{Output:} sets of attributes whose support in BN is $\geq \varepsilon$.  \\
\textbf{1.} Let $k \leftarrow 1$.  \\
\textbf{2.} Let $C_{\text{and}} \leftarrow \{I: |I| = 1\}$.  \\
\textbf{3.} Compute $\text{supp}_{BN}(I)$ for all $I \in C_{\text{and}}$ using the algorithm i  \\
\textbf{4.} Let $\text{Freq}_k \leftarrow \{I \in C_{\text{and}}: \text{supp}_{BN}(I) \geq \varepsilon\}$.  \\
\textbf{5.} Let $C_{\text{and}} \leftarrow$ generate new candidates from $\text{Freq}_k$.  \\
\textbf{6.} Remove attribute sets with infrequent subsets from $C_{\text{and}}$.  \\
\textbf{7.} Let $k \leftarrow k + 1$; Go to 3  \\
\hline
\end{tabular}
\end{center}

Figure 5j: Frequent itemsets in Bayesian network

The knowledge integration is carried out efficiently from the distributed workstations in parallel environment using the above mentioned algorithms. The detailed analysis of the results obtained and its interpretations are discussed in Chapter 6.