Why do organizations need business intelligence?

- BI systems are computer programs that provide valuable information for decision making.

Three primary BI systems:
- **tools** read data, process them, and format the data into structured reports (e.g., sorting, grouping, summing, and averaging) that are delivered to users. They are used primarily for assessment. RFM is one of the tools for reporting.
- **tools** process data using statistical, regression, decision tree, and market basket techniques to discover hidden patterns and relationships, and make predictions based on the results.
- **tools** store employee knowledge, make it available to whomever needs it. These tools are distinguished from the others because the source of the data is human knowledge.

Market Basket Analysis

- **Definition:** Market Basket Analysis (Association Analysis) is a mathematical modeling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items.

- **Purpose:** It is used to analyze the customer purchasing behavior and helps in increasing the sales and maintaining inventory by focusing on the point of sale transaction data.

- **Given a dataset, an algorithm (e.g., FPGrowth or Apriori Algorithm) trains and identifies product baskets and product association rules to find frequent items.**
Definitions and Terminology

• What items are bought together
  – What’s in each shopping cart/basket?
• Basket data consist of collection of transaction data and items bought in a transaction
  – Itemset
• How does this data differ from a transaction database?
  – Pivoting
• Retail organizations interested in generating qualified decisions and strategy based on analysis of transaction data
  – what to put on sale, how to place merchandise on shelves for maximizing profit, customer segmentation based on buying pattern

Definitions and Terminology (cont.)

• Transaction is a set of items (Itemset).
• Confidence: It is the measure of uncertainty or trust worthiness associated with each discovered pattern.
• Support: It is the measure of how often the collection of items in an association occur together as percentage of all transactions.
  – If you only want to find strong correlations, you can set it ___. If you want a wide range of unknown correlations, set it _____. Therefore, the lower support value the easier to get an solution.
• Frequent itemset: If an itemset satisfies minimum support, then it is a frequent itemset.
• Strong Association rules: Rules that satisfy both a minimum support threshold and a minimum confidence threshold
• In Association rule mining, we first find all frequent itemsets (i.e., meet a user-specified minimum support threshold) and then generate strong association rules (based on a user-specified minimum confidence) from the frequent itemsets

RapidMiner

• 1. Market(RM Format).xls – the data provided by the instructor
• 2a. With Itemcount
• 2b. Ignores Item Amounts (use the data provided in Market-Data repository)
  – ‘CreateAssociation” operator needs to be added in order to obtain ‘confidence’ data and perform its analysis
• 3. HW – ScubaGear
• 4. Project

Performing Market Basket Analysis (with ItemCount)

• To perform a Market Basket Analysis, we will begin by selecting “Open Template” from the main menu (Or by clicking “File>Open template) as is shown in Fig 1.

  2. From the “Open Template” window, we will scroll down to the bottom and choose “Market Basket analysis” as is shown in Fig 2-a. Proceed by clicking next (Fig 2-b), and then Finish. We will be modifying this skeleton format to suit our needs

  3. The Main Process window will have loaded the skeleton format for the Market Basket Analysis as is shown in Fig 3a. Our first modification will be to remove the “Retrieve” operator (Fig. 3-a), and replace it with a “Read Excel” Operator by right-clicking the Retrieve operator and choose “Select Operator” option then import → data → “Read Excel” operator (Fig 3-b) or deleting the “Retrieve Operator” the adding the “Read Excel” Operator (Fig 3-c). You may have to re-connect from “out” in “Read Excel” Operator to “thr” (for “through”) in “Define Item Count” Operator. Use the import wizard in the Read Excel operator to locate and Read the Market Basket file (file name: Basket (RM format).xls), and after clicking Next, choose the correct sheet (Numerical Data) as shown in Fig 3-c, and make note of the column header labels in the file (columnId, itemId, itemCount) then select NEXT, NEXT (Fig 3-d, 3-e) otherwise, only one attribute displayed. Do not need to identify “label” for a selected attribute as we did in the Life Insurance example) then Finish (Fig 3-f).
The figure shown on the left is the data file used for this example. The original data set contains 76 types of items and 89 customers with different combinations of items purchased (details can be found in the EXCEL file).

However, the itemCount is generated using a random generator function of \((=RANDBETWEEN(1,10))\) from the Excel.

When the process is run, the results show the association rules created in the form: Premises \(\Rightarrow\) Conclusion, as shown in figure 7-e. Your results will depend on the minimum confidence and support chosen in the FP-Growth, and the Create Association rules operators respectively. For this example, a minimum support of 0.0001 has been used, and a minimum confidence of 0.8.

Row no. 1 can be read as: The purchase of item 37 implies the purchase of item 2. These items are bought together in 4.5% of the transactions, and 80% (probability) of the times that Item 37 is bought, Item 2 is bought as well.

What is the business “implication”? 

Interpretation on the results: (Basket(RM).xls)

Your Turn …

• Follow the handout.
• Continue on Step 4
Dr. Chen, Business Intelligence

Part II

• Develop market basket analysis using RapidMiner (Ignore ItemCounts)

---

Interpretation on the results: (Basket[RM].xls)

• Row no 33 can be interpreted as: The purchase of item 2 implies the purchase of items 70 and 39. These three items are bought together in 10% of the transactions, and 20% of the times that item 2 is bought, items 70 and 39 are bought together as well.
• Note: Generally, with large amounts of data, the lower the minimum confidence chosen is, the longer the program will take to process.
• What is the business “implication”?

<table>
<thead>
<tr>
<th>No</th>
<th>Premises</th>
<th>Conclusion</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>sum/Item2 Item12</td>
<td>sum/Item1 Item2 Item11</td>
<td>0.100</td>
<td>0.000</td>
</tr>
<tr>
<td>32</td>
<td>sum/Item2 Item12</td>
<td>sum/Item1 Item2 Item11</td>
<td>0.100</td>
<td>0.000</td>
</tr>
<tr>
<td>33</td>
<td>sum/Item2 Item12</td>
<td>sum/Item1 Item2 Item11</td>
<td>0.100</td>
<td>0.000</td>
</tr>
</tbody>
</table>

---

Grocery Store Scenario

• Here we have an Excel-based dataset containing information about customers who have shopped in a grocery store. We are trying to detect relationships or associations between specific items in a large “catalog” of objects using market basket analysis.
• A simple example would be the occurrence of beers and diapers in the same sales transaction.
• However the real value from association rules analysis is finding connections between seemingly non-intuitive items: it would be indeed surprising (and valuable) to the retailer if it was found that there was a strong association between beer and diapers.

---

RapidMiner

• 1. Market Basket Analysis (Grocery store example) use data provided by your instructor
  – a) follow the tutorial to produce the “support” from the data set
  – b) ‘CreateAssociation” operator needs to be added in order to obtain both ‘support’ and ‘confidence’ information and perform its analysis
• 2. HW – ScubaGear
  – Data file: Grocery.xlsx

---

Market Basket Analysis (Ignores Item Amount)

• 1. We will begin by starting a new project from the main menu by clicking the “New Process” Icon when starting RapidMiner as shown in Figure 1-a.
• 2. On the left hand side of the screen, we will click on Repositories, and will follow by expanding (clicking the plus (+) sign) Samples > processes > 02_Preprocessing. Double-click the 23_Transactional2Basket Process (Figure 1-b) the process model of ”Transactional2Basket Process” is automatically created as in Fig 1-c.

---

Grocery Store Scenario

• In this simple scenario that there are three customers (John, Mark and Alex) with six different items they purchased. Information about each customer’s purchasing history is included in the dataset as shown in Table 1.
• In order to produce the result from market basket analysis, we are using the RapidMiner software (www.rapidminer.com). RapidMiner supports many different data mining techniques, but we will focus only on market basket analysis here. Please note that TID and ITEM should be in ‘upper’ case.

---

Note: Generally, with large amounts of data, the lower the minimum confidence chosen is, the longer the program will take to process.
• What is the business “implication”?
Step 3. If you would like to see the data that we will be working with, you can expand the “data” folder also shown in Fig 1-b, and double-click the Market-Data repository (Fig 2-a) or click View and select Perspective → Results as shown in Fig 2-b.

TID: Customer ID
ITEM: Item purchased ID

<table>
<thead>
<tr>
<th></th>
<th>TID</th>
<th>ITEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
<td>3.0</td>
</tr>
<tr>
<td>5</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>6</td>
<td>2.0</td>
<td>4.0</td>
</tr>
<tr>
<td>7</td>
<td>3.0</td>
<td>5.0</td>
</tr>
<tr>
<td>8</td>
<td>3.0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

How about confidence?

- Step 10. While the previous result shows the support of items occurring together, we need to add an additional “operator” to be able to create and obtain association rules between the items. As shown in Fig 7-a, we must add a “Create Association Rules” operator between the FPgrowth operator and the result. However, if we go ahead to RUN the process, an error was detected as shown in Fig 7-b.

Interpretation on the results: (Ignores Data Amounts)

- Row no. 1 can be read as: The purchase of item 1 implies the purchase of item 3. These items are bought together in 33% of the transactions, and 50% probability of the times that Item 1 is bought, Item 3 is bought as well.

- If we look at row no. 4, we see that the purchase of Item 3 implies the purchase of Item 1. Again, this combination happens in 33% of the transactions, but we see that every time Item 3 is bought, Item 1 is bought as well (100% confidence).

- What is the business “implication”?
**HW**

- Hints:
  1. Follow the handout of “Ignores Item Amounts”
  2. Create your data file
     - modify the data file from the previous handout
     - Use “actual” customer name rather than customer id (numerical values in TID)
     - Please note that TID and ITEM should be in ‘upper’ case.
  3. Figure out how to “Read” your excel file into the model

**FP-Growth (RapidMiner Core)**

- The frequent-itemsets problem is that of finding sets of items that appear together in at least a threshold ratio of transactions. This threshold is defined by the ‘minimum support’ criteria.
- This operator has two basic working modes:
  - finding at least the specified number of itemsets with highest support without taking the ‘min support’ into account. This mode is available when the find min number of itemsets parameter is set to true. Then this operator finds the number of itemsets specified in the min number of itemsets parameter. The min support parameter is ignored in this case.
  - finding all itemsets with a support larger than the specified minimum support. The minimum support is specified through the min support parameter. This mode is available when the find min number of itemsets parameter is set to false.

**Create Association Rules (RapidMiner Core)**

- **confidence**: The confidence of a rule is defined \( \text{conf}(X \implies Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \). Be careful when reading the expression: here \( \text{supp}(X \cup Y) \) means “support for occurrences of transactions where X and Y both appear”, not “support for occurrences of transactions where either X or Y appears”. Confidence ranges from 0 to 1
- Confidence is an estimate of \( \Pr(Y | X) \), the probability of observing Y given X. The support \( \text{supp}(X) \) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.

**Data files used between “Ignores Item Amounts” and HW (but, HW is with four customers – TID; see detail on the HW)**

<table>
<thead>
<tr>
<th>TID</th>
<th>ITEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>2</td>
<td>3.0</td>
</tr>
<tr>
<td>2</td>
<td>4.0</td>
</tr>
<tr>
<td>3</td>
<td>5.0</td>
</tr>
<tr>
<td>3</td>
<td>6.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TID</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>John</td>
<td>Mask</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>John</td>
<td>Tank</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>John</td>
<td>Fins</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>Amanda</td>
<td>Mask</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
<td>Alexa</td>
<td>Weight</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5</td>
<td>Alexa</td>
<td>Computer</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>6</td>
<td>Alexa</td>
<td>Suit</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>7</td>
<td>Alexa</td>
<td>Suite</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>8</td>
<td>Alexa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>