Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

BPI 2020 Assignment – Part 2

Introduction

This assignment guides you through the analysis of a real data set using the techniques and tools provided in the course. The assignment tests your understanding of process mining concepts and techniques. It is necessary to follow the assignment in the given order since the results of specific questions might depend on answers to previous questions.

Introducing the Event Log

This data originated from a large multinational company operating from The Netherlands in the area of coatings and paints and we ask students to investigate the purchase order handling process for some of its 60 subsidiaries. In particular, the process owner has compliance questions. In the data, each purchase order (or purchase document) contains one or more line items. For each line item, there are roughly four types of flows in the data:

1. **3-way matching, invoice after goods receipt**: For these items, *we should have first record goods receipt and then invoice receipt* (indicated by both the GR-based flag and the Goods Receipt flags set to true).
2. **3-way matching, invoice before goods receipt**: Purchase Items that do require a goods receipt message, while they do not require GR-based invoicing (indicated by the GR-based IV flag set to false and the Goods Receipt flags set to true). For such purchase items, invoices can be entered before the goods are receipt, but they are blocked until goods are received. This unblocking can be done by a user, or by a batch process at regular intervals. Invoices should only be cleared if goods are received and the value matches with the invoice and the value at creation of the item.
3. **2-way matching (no goods receipt needed)**: For these items, the value of the invoice should match the value at creation (in full or partially until PO value is consumed), but

there is no separate goods receipt message required (indicated by both the GR-based flag and the Goods Receipt flags set to false).

4. **Consignment**: For these items, there are no invoices on the PO level as this is handled fully in a separate process. Here we see the GR indicator is set to true but the GR IV flag is set to false and also we know by item type (consignment) that we do not expect an invoice against this item.

Overall, for each line item, the amounts of the line item, the goods receipt messages (if applicable), and the invoices have to match for the process to be compliant. Of course, the log is anonymized, but some semantics are left in the data, for example:

- The resources are split between batch users and normal users indicated by their names. The batch users are automated processes executed by different systems. Normal users refer to human actors in the process.
- The monetary values of each event are anonymized from the original data using a linear translation respecting 0, i.e. addition of multiple invoices for a single item should still lead to the original item worth (although there may be small rounding errors for numerical reasons).
- The company, vendor system, and document names and IDs are anonymized consistently throughout the log. The company has the key, so any result can be translated by them to business insights about real customers and real purchase documents.

For each purchased item (or case) the following attributes are recorded:

- **Concept:name**: A combination of the purchase document id and the item id,
- **Purchasing Document**: The purchasing document ID,
- **Item**: The item ID,
- **Item Type**: The type of the item,
- **GR-Based Inv. Verif.**: Flag indicating if GR-based invoicing is required (see above),
- **Goods Receipt**: Flag indicating if 3-way matching is required (see above),
- **Source**: The source system of this item,
- **Doc. Category name**: The name of the category of the purchasing document,
- **Company**: The subsidiary of the company from where the purchase originated,
- **Spend classification text**: A text explaining the class of purchase item,
- **Spend area text**: A text explaining the area for the purchased item,
- **Sub spend area text**: A text explaining the area for the purchased item,
- **Vendor**: The vendor to which the purchase document was sent,
- **Name**: The name of the vendor,
- **Document Type**: The document type,
- **Item Category**: The category as explained above (3-way with GR-based invoicing, 3-way without, 2-way, consignment).

Question 1 - Knowing the Event log (10 points)

The first step to start analyzing and providing value for the business owners is to know the event log attributes and features. However, in most of the real event logs, there are **incomplete** traces that could affect the understandability of the results. These traces should be removed from the event log before starting further analysis.

Divide the original event log into four event logs with respect to the four mentioned types of item categories. Afterwards, you should have 5 event logs (the original event log + four sub-event-logs with respect to the four categories).

Apply a filtering technique to remove the incomplete traces from the sub-event-logs. Now, you should have 9 event logs (the original event log + four sub-event-logs + four filtered sub-event-logs).

For those 9 event logs, answer the following questions to discover the general characteristics of each event log **before** and **after** filtering. Also, explain the details of your filtering technique.

1. How many cases and events are in the event log?
2. How many unique trace variants are in the event log?
3. What is the number of unique activities and unique resources?
4. What are the minimum and the maximum number of activities in a trace in the process?
5. What is the set of start and the set of end activities in the cases?

For the rest of the assignment, use the filtered sub-event-logs or the original event log as indicated in the questions.

Question 2 - Process Discovery (25 points)

The business owners are interested in discovering how their processes are actually being executed. However, the processes clearly show much variability, and the activities that are actually executed depend on the outcome. In the following, you are asked to provide different process models.

1. Discover four process models with respect to the four types of item categories. Those models need to be sound, cover round about 80% of the traces and should have a precision as high as possible.
2. Compare the discovered models for **3-way match, invoice after goods receipt** and **3-way match, invoice before goods receipt**. Explain at least five differences in the models.
3. What is the most frequent variant in each category? What are the similarities and differences between those that you discovered?
4. Generate a C-net of the **3-way matching, invoice after goods receipt** process containing roughly 15% of the directly follow relations. Explain the parameters that you use and the resulting C-net.
5. (Continued from 2-4) Based on your analysis of the C-net, create a Petri net of the **3-way matching, invoice after goods** receipt process on your own (**manually**).
6. Consider the **3-way matching categories (invoice before and after goods receipt).** In how many cases (percentage) "*Create Purchase Order Item*" is not eventually followed by "*Record Goods Receipt*"? Explain what happens in these cases.

Question 3 - Conformance Checking (20 points)

The process owner wants to have process models with good qualities (i.e., process models that reflect the reality properly). To that end, they want the evidence on the quality of these models.

1. For the four event logs belonging to the four item categories (**3-way matching (Invoice after goods receipt), 3-way matching (Invoice before goods receipt**), **2-way matching (no goods receipt needed), Consignment**) and their corresponding models discovered in Question 2.1, perform alignment-based conformance checking. Furthermore, indicate where the main deviations happen according to the replay. (Note that you should have four results.)
2. Consider the three business rules for the following three item categories:
   a. **3-way matching (invoice after goods receipt) -** *Invoice happens after good receipt*,
   b. 3-way matching (invoice before goods receipt) - *Invoice happens before good receipt,*
   c. 2-way matching (no goods receipt needed) - *no goods receipt*,

   Explain in how many cases (percentage) the rules are violated, or in case, there is no violation, explain why.

Question 4 – Decision Mining (10 points)

1. Mark possible decision points on the discovered process model for **3-way match, invoice after goods receipt** in Question 2-1.
2. (Continued from 4-1) Using any attributes that you consider relevant, find a meaningful decision tree for any of the possible decision points. Explain why the decision tree is meaningful in terms of quality measures (e.g., precision, recall and F-measure).
3. (Continued from 4-2) Interpret the discovered decision tree (e.g., explain transition guards derived from the decision tree).

Question 5 – Performance Analysis (15 points)

1. Analyze the performance of **3-way matching, invoice before goods receipt** based on the process model that you discover in Question 2-1. What are the bottlenecks? What are your recommendations for the company to increase the performance of the process?
2. Consider **3-way matching, invoice before goods receipt**. Compare the mean and median of case durations between **(a)** *the instances whose case duration belongs to the top 20%* and (**b)** *the instances whose case duration belongs to the bottom 20%*.
3. (Continued from 5-2) Discover process models of **(a)** and **(b)** and explain at least two differences in the models. You can use any algorithms to discover the models.
4. Assume that the process owner is interested in the following performance measures.
    a. For Internal Customer: From PO Item to Goods (Time between “Create Purchase Order Item” and “Record Goods Receipt”)
    b. For Vendor Relationship: From invoice to invoice receipt (Time between “Vendor creates invoice” and “Record Invoice Receipt”)

    For **3-way matching, invoice before goods receipt**, what is the mean and median of those performance measure?

Question 6 – Organizational Mining (10 points)

1. In the event log belonging to the category **3-way matching (invoice after goods receipt)**, extract the main variant (i.e., the most frequent variant) for which multiple record goods receipts are performed. Analyze the resource perspective of the cases having this variant (e.g., analyze who serves those cases most frequently).
2. For the following sub-questions, consider **2-way matching category**, and NONE as one of the resources in the process:
   a. By changing the threshold of the correlation coefficient (the slider of the edges), discover and interpret two social networks of the roles (similar-task) that show some resources do similar tasks.
   b. List the set of activities, which are done by the smallest non-individual role (i.e., the role that contains the minimum number of resources among the ones having more than one resource).

      Hint: you can find the smallest non-individual role by changing the threshold (slider) of the correlation coefficient.

3. Analyze the resource perspective (including the NONE user) of the top 2 cases in the original event log that have the longest throughput time (in case there are more than 2 cases having the longest throughput time, you are free to choose any of them):
   a. What is the list of resources involved in these cases?
   b. Show and interpret the handover of work social network for the resources obtained in 3-a.
   c. Show the social network of the roles (similar-task) based on the correlation coefficient for the resources obtained in 3-a.
   d. Based on the social network obtained in 3-c, which resources can be specified as the ones who perform similar tasks (i.e., the same roles)?

Deliverables

The deadline for the assignment is on **Friday 10/07/2020.** You will need to hand in your submission via **Moodle**. Note that the deadline is strict (i.e., there is no extension possible, and late submissions will not be considered). Do not risk last-minute technical problems due to internet problems or RWTHmoodle failures. Your submission should be a **PDF File**, which presents your results, explanations, and screenshots of the used tools.

Report requirements:

- All the names and student numbers of the group members should be included on the first page of your report. Otherwise, your assignment will not be graded.
- The answers to the questions should be succinct but clear. Avoid being verbose while not answering the question.
- For this assignment, you are free to choose any tool introduced in the course. Specify the tool that you are using for the task.
- The screenshots are necessary to convincingly support your findings by evidence. Any finding that is not supported by a screenshot will not be considered.
- The report file should not exceed **30 pages**. The font size of your report should not be smaller than **12**.
- The screenshots are not acceptable if they are not readable. Therefore, the findings supported by those screenshots will not be considered.
- If you make some assumptions, please mention it in your report explicitly. Any assumption is accepted as long as it is reasonable and mentioned.
  - → Put a screenshot of the parameters you are using for the algorithms.
- The structure and quality of the report will also be assessed and graded.
  - → Ensure that the report is of sufficient quality.
  - → The report should be well-structured, start with an introduction, and end with a conclusion (keep them concise).

Note that only one of the group members should upload the assignment.

Grading

Participation in the assignment is one of the prerequisites for taking the written exam. The two

Assignments and the exam form a whole and it is not possible to retake parts of the course, i.e., the results of the assignments expire after the written exam. Furthermore, assignments can only be redone in the next academic year.

10% of the grading of this assignment is based on the **reporting style** (answers should be well-structured and explanations should be clear).

The grade of this assignment counts 20% towards the final grade (see the study guide for details). Please note that the correctness and completeness of your results, and also the accuracy of your explanation, are essential.