Стогний Е.А., Барсегян К.Д.

*Process Mining*

**Отчет по лабораторной работе №2**

*О журнале событий*

Основой для выполнения лабораторной работы выступает более полутора миллиона данных от крупной многонациональной компании, работающей в Нидерландах в области покрытий и красок. Рассматривается процесс обработки заказов на поставку для некоторых из ее 60 дочерних компаний. Журнал содержит данные с 1948 по 2018 гг.

Для работы с журналом событий определим основные атрибуты событий в журнале (Табл. 1).

Таблица 1 – Основные атрибуты событий в журнале

|  |  |  |
| --- | --- | --- |
| **Атрибут** | **Наименование** | **Роль** |
| Идентификатор случая (case id) | case:concept:name | Заказ |
| Деятельность  (activity name) | concept:name | Изменение статуса заказа |
| Ресурс (resource) | org:resource | Исполнитель изменения (система или оператор) |
| Отметка времени (timestamp) | time:timestamp | Дата и время регистрации изменения |

К остальным атрибутам относятся тип документа, наименование компании, поставщика и другие позиции, не являющиеся важными для нашего анализа.

*Обработка пропущенных значений*

В ходе анализа данных было выявлено, что пропущенных значений в столбцах не имеется, однако в org:resource для некоторых исполнителей указано None. Исходя из описания данных, мы знаем, что None указывается в случаях, когда пользователь не был зарегистрирован. Заменим None на Unregistered для большей информативности.

*Статистики по журналу событий*

При обработке данных журнала была получена следующая информация:

* Число уникальных заказов составляет **251 734**. Самое длительное выполнение заказа составило **70,5** лет (у заказов с id 4507004931\_00030, 4507004931\_00020, 4507004931\_00010, 4507004931\_00040 и 4507004931\_00050 – год их создания датируется 1948-м). Некоторые заказы еще не исполнены, поэтому время их выполнения составляет **0** минут. При этом минимальное время исполненного заказа составило **2** дня 2 часа 46 минут у заказов 4508046096\_00040, 4508046096\_00030, 4508046096\_00020 и 4508046096\_00010.
* Число уникальных статусов заказа – **42**. Всего **11 973** различных путей исполнения заказа. Максимальное число изменений статуса у заказов – **990** (у заказа id4507027474\_00010), минимальное – **1**. Среднее время между созданием заказа и последующим изменением его статуса составляет **13.6** дней. Самая часто встречающаяся последовательность событий (встречается **50 286**) выглядит следующим образом:
* **«Заказ создан»** (Create Purchase Order Item) **->**
* **«Создаётся счет-фактура»** (Vendor creates invoice) **->**
* **«Зафиксирован приход товара» (**Record Goods Receipt) **->**
* **«Зафиксировано получение счет-фактуры» (**Record Invoice Receipt) **->**
* **«Счет очищен»** (Clear Invoice).
* Самый частый статус заказа – Регистрация прихода товара (Record Goods Receipt). Данный статус повторялся **314 097** раз в журнале.
* В среднем самое длительное действие – Ошибка передачи (SRM: Transfer Failed (E.Sys.)), оно составляет **103.5** дня. Самые короткие в среднем по длительности действия:
* SRM: Ожидает Одобрения (SRM: Awaiting Approval),
* SRM: Завершено (SRM: Complete),
* SRM: Не завершено (SRM: Incomplete),
* SRM: Оформлено (SRM: Ordered),

их длительность составляет **0** минут. Суммарно самое продолжительное действие в журнале – Запись квитанции об оплате (Record Invoice Receipt), оно заняло **6 940 385** дней (около 19 000 лет).

* Число уникальных исполнителей равно **627**, при этом выделяется группа незарегистрированных в системе пользователей; всего 607 пользователей-людей и 20 системных. За весь журнал событий больше всего действий выполнял user\_002 – **166 353** действия. Некоторые пользователи выполняли по **1** действию, и это количество является минимальным. Больше всего уникальных исполнителей было задействовано на статусе Регистрация прихода товара (Record Goods Receipt) – **256** исполнителей.

*Узкие места*

К узким местам можно отнести те изменения статуса заказа, которые долго (по сравнению с остальными процедурами) длятся при расчете на 1 изменение статуса, то есть имеют наибольшую среднюю длительность. К ним можно отнести следующие статусы (Табл. 2).

Таблица 2 – Узкие места в процессе

|  |  |  |
| --- | --- | --- |
| **Статус** | **Средняя длительность** | **Исполнитель** |
| SRM: Ошибка передачи (E.Sys.) (SRM: Transfer Failed (E.Sys.)) | 103 days 10:59:00 | user\_003 |
| SRM: Транзакция завершена (SRM: Transaction Completed) | 77 days 22:36:00 | user\_022, user\_026, user\_016, user\_009, user\_010 |
| Поставщик создает дебетовое авизо (Vendor creates debit memo) | 41 days 16:59:48 | Unregistered |

Также к узким местам можно отнести самое длительное выполнение заказа – **70,5** лет у заказов с id 4507004931\_00030, 4507004931\_00020, 4507004931\_00010, 4507004931\_00040 и 4507004931\_00050, созданных в 1948-м.

Еще к узкому месту можно отнести суммарно самое продолжительное действие в журнале – это Запись квитанции об оплате (Record Invoice Receipt), оно заняло **6 940 385** дней (заняло около 19 лет).

К узкому месту можно отнести и работу пользователя user\_002, так как за весь журнал событий он изменял больше всего (по сравнению с другими исполнителями) статусов заказа – **166 353**.

*Удаление аномальных значений*

Так как у пяти заказов с 1948 года аномально длительная продолжительность выполнения, мы решили удалить их из датасета, и посмотреть, как изменятся статистики. Число заказов уменьшилось на 5 и стало **251 729.** Самый загруженный исполнитель потерял 5 заказов, его загруженность также уменьшилась на 5 и составила **166 348**. Самое длительное выполнение заказа составило **24 года**. Самое быстрое выполнение заказа составило **2 дня 9 часов**. Среднее время изменения с 1го на 2й статус заказа составило **13.1** дней. Узким местом стал Статус «Снятие блокировки платежа» – он длится в среднем **более 35 дней.**

*Построение моделей*

По нашему набору данных было построено четыре модели процессов с помощью различных алгоритмов обнаружения процесса: Inductive miner, Heuristics miner, Alpha Miner и DFG граф. Полученные метрики приведены ниже (Табл. 3).

Таблица 3 – Метрики моделей обнаружения процесса

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Simplicity** | **Fitness** | **Precision** | **Generalization** |
| **Inductive miner** | 0.584046 | 0.999995 | 0.23604 | 0.903164 |
| **Heuristic miner** | 0.450262 | 0.829126 | 0.839973 | 0.770133 |
| **Alpha miner** | 0.495935 | 0.340944 | 0.220888 | 0.911182 |
| **DFG graph** | 0.373793 | 0.656737 | 1 | 0.659717 |

*Выводы*

Как видно из полученных метрик, по двум показателям из четырех (Simplicity, Fitness) лидирует Inductive miner, его обобщенность (Generalization) вторая среди оставшихся алгоритмов, однако при этом у него самая низкая точность (Precision). Поэтому, на наш взгляд, он не подходит для выбора в качестве лучшей модели для анализа процесса.

По нашему предположению, лучшим выбором будет Heuristic miner, поскольку ни одна из его метрик не является худшей по сравнению с остальными алгоритмами, а в большинстве случаев она близка к лучшим показателям. Heuristic miner помогает бороться с шумом (например, в нашем журнале событий шумом могут выступать единовременное создание заказа и его удаление). Также данный алгоритм работает с большими журналами (в нашем журнале более полутора миллиона данных).