Exploring different aspects of users behaviours in the Dutch autonomous administrative authority through process cubes

Amin Jalali

Stockholm University, Sweden,
aj@dsv.su.se

Abstract. This report demonstrates how storing events in a data warehouse can facilitate mining business process models. The data warehouse is designed based on dimensional modelling principles for the Business Processing Intelligence Challenge (BPIC) 2016, and it is populated by extracting and cleaning, transforming and loading data from the challenge log files. In addition, the data warehouse is stored in Multidimensional OnLine Analytical Processing (MOLAP) storage format to investigate its potential use. The data are queried based on different process cube operations, and the result is used by different process mining techniques. In this way, the business process models are investigated at different levels of abstraction. The result shows that it could be beneficial to structure event logs in a multidimensional structure to apply different process mining techniques more effectively.

Keywords: Process Mining; process cubes; Business Processing Intelligence Challenge;

1 Introduction

Process Mining is an area of research that aims to support managing business processes by analysing their event logs. Enacting business processes can result in huge number of events, which can be recorded by log files. These log files can include the enactment result of many process instances with different variations. Thus, applying process mining techniques can be a challenging and a difficult task in practice. Structuring the events in a way to select the relevant subset for mining and analysis is an important challenge in process mining. W.M.P. van der Aalst proposed Process Cubes to store events in dimensional structure, which enable analysts to select the relevant part of the events based on different operators [4].

In this study, we show how storing events in a data warehouse can support analysis of different aspects of business processes using process mining techniques. The event logs which are provided for the Business Processing Intelligence Challenge (BPIC) 2016 [5–9] are Extracted and cleaned, Transformed and

Loaded (ETL) into a data warehouse including several dimensions and fact tables. The data warehouse is used as a base to select different events which are relevant for analysis. These events are mined by process mining techniques. The result shows that it could be beneficial to structure event logs in a multidimensional structure to apply different process mining techniques more effectively.

The rest of this report is structured as follow. Section 2 describes the ETL process briefly. Section 3 describes the customers properties very briefly. Section 4 describes how customers use the services. Section 5 explore the users behaviour who have been complained at least once. Finally, Section 6 concludes this report.

2 Preparing the stage

In this section, we briefly take a look at the overall picture of our approach including steps that we performed to build and use our data warehouse. In addition, we will see the structure of the data warehouse that is designed and implemented to support the whole study.

2.1 The overall picture

Fig. 1 shows the overall picture of our approach. First, we designed and implemented a multidimensional data warehouse to store events for the BPIC 2016. The events are extracted from the five log files [5–9] provided for the challenge. The events are cleaned, transformed and loaded into the data warehouse. We also populated the data into cubes with the Multidimensional OnLine Analytical Processing (MOLAP) storage format to investigate its potential use.

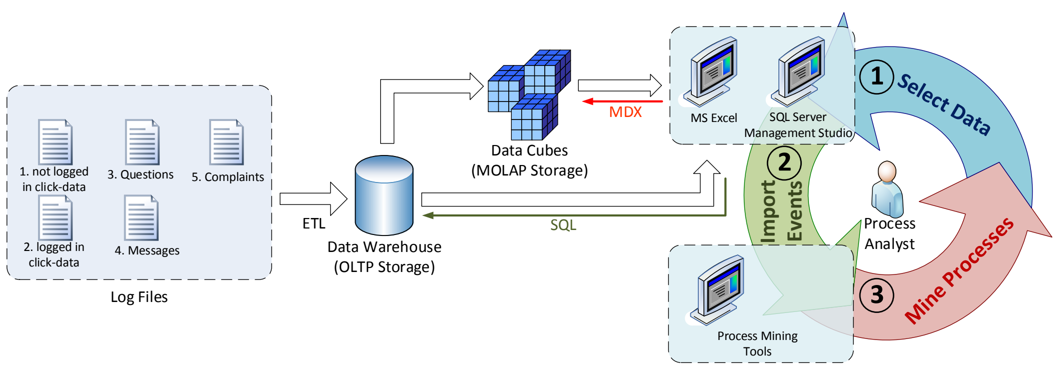


Fig. 1. The overall picture of our approach

The data can be retrieved from the data warehouse and the cubes by different client applications. We used Microsoft Excel and Microsoft SQL Server Management Studio to retrieve the data. The data warehouse is stored in the usual relational database format with the OnLine Transactional Processing (OLTP) storage mode, thus the data can be retrieved from the data warehouse using Structured Query Language (SQL). The data from the data cubes can be retrieved by Multidimensional Expressions (MDX).

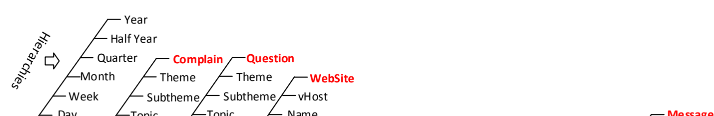
A process analyst can follow three cyclical steps when mining the processes in our approach.

1. The analyst can select a part of data (events) from the client applications.
2. The analyst can import the selected events set into a process mining tool.
3. The analyst can mine different aspects of the process based on the imported log.

The analyst can re-initate the process based on the insight that (s)he has from the mining result to re-select the data. In this way, the analyst can mine different parts of a process model with different level of abstraction in an iterative way.

2.2 The multidimensional structure of the data warehouse

Fig. 2 shows the defined fact tables, dimensions and their relations. This table is called the data warehouse *bus matrix* [3]. It also shows hierarchies that are define for some dimensions. The hierarchies can be used to drilling down or rolling up into different levels of detail in analysing data. Some dimensions are shared among different fact tables, which are called *conformed dimensions* [3]. These dimensions may enable another sort of operation, called *drilling across* [3]. By drilling across the data, an analyst can relate the data of two fact tables together for analysis.



Fact tables ⇩ / Dimensions ⇨	Date	Complain Type	Question	Page	Channel	Customer	Session	IP	Message
UnknownClick	●			●			●	●	
KnownClick	●			●		●	●	●	
Question	●		●			●			
Message	●				●	●			●
Complain	●	●			●	●			

Fig. 2. The Bus Matrix of the data warehouse

We defined the following six fact tables to measure different aspects of customers behaviour in the business process:

- UnknownClick: to measure the behaviour of unknown customers when using the website.

– KnownClick: to measure the behaviour of known customers when using the website.
– Question: to measure the behaviour of known customers when asking questions.
– Message: to measure the behaviour of known customers sending messages.
– Complain: to measure the behaviour of known customers when complaining.

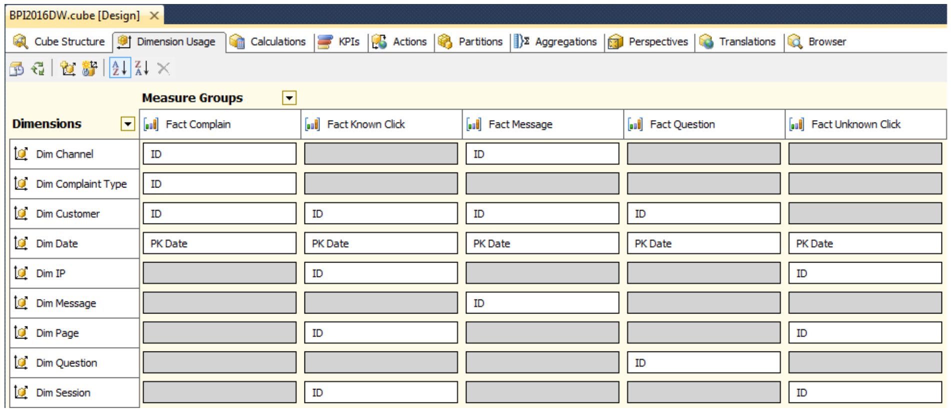
We also defined the following ten dimensions:

– Date: to capture the calendar information. This dimension has a hierarchy with the following levels scaled from high to low level from left to right: Year → Half Year → Quarter → Month → Week → Day.
– Complain Type: to capture the complain information. This dimension has a hierarchy with the following levels scaled from high to low level from left to right: Complain[1] → Theme → Subtheme → Topic.
– Question: to capture the question information. This dimension has a hierarchy with the following levels scaled from high to low level from left to right: Question[1] → Theme → Subtheme → Topic.
– Page: to capture the visited page information. This dimension has a hierarchy with the following levels scaled from high to low level from left to right: WebSite[1] → vhost (host address) → (Page) Name.
– Channel: to capture the available channels for communication.
– Customer: to capture customer information.
– Session: to capture the session information.
– IP: to capture the IP information.
– Message: to capture the Message information. This dimension has a hierarchy with only one level, i.e. Message[1].

Note that each dimension has different attributes that enables us to investigate different aspects of a business process. For example, the customer dimension has age category, Gender, and other attributes that enables us to focus on different aspects. As it can be seen in the figure, some dimensions are shared among different fact tables. These dimensions enables us to perform the cross analysis by combining different measures. It should be also noted that the lowest granular level of the date dimension is day. We capture the time of events as attributes in different fact tables. The session and ip dimensions are defined in order to enable us to combine known and unknown click measures. In this way, we have the opportunity to capture the identity of some customers who continue to surf the website after logging out or before logging in.

Fig. 3 shows the screen-shot from the implementation of the described bus matrix in Microsoft SQL Server Analysis Services (the Microsoft Data Warehouse toolkit and OLAP Server). The fact tables are listed as columns, and the dimensions are listed as rows. The order of fact tables and dimensions in the cube cannot be changed to be the same order that we defined, yet their relations is identically the same.

[1] This level is only defined for drill across feature which will be explained later.



Dimensions	Fact Complain	Fact Known Click	Fact Message	Fact Question	Fact Unknown Click
Dim Channel	ID		ID		
Dim Complaint Type	ID				
Dim Customer	ID	ID	ID	ID	
Dim Date	PK Date	PK Date	PK Date	PK Date	PK Date
Dim IP		ID			ID
Dim Message			ID		
Dim Page		ID			ID
Dim Question				ID	
Dim Session		ID			ID

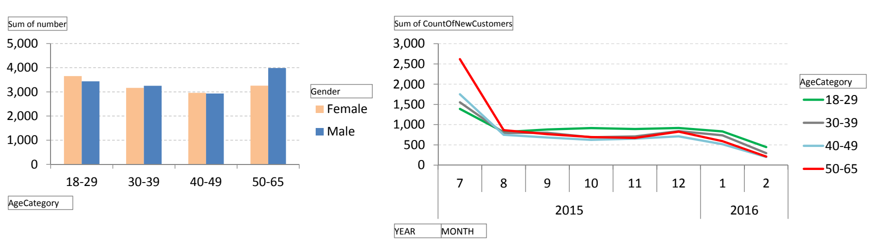
Fig. 3. The Bus Matrix of the data warehouse

3 Customers

Customers are the core of most of businesses and processes. Thus, we spend some time to investigate some insights about our customers in this section. Most of these insights are obtained by using Microsoft Excel in connection to the Data Cubes.

3.1 Customers demography

In this section, we considered the customers who has used the website as active customers. Fig. 4 shows basic insights about the active customers, where Fig. 4(a) shows the customers demography and Fig. 4(b) shows the appearance of new customers over time. We considered the first use of website by a customer as the time that customer uses the service, so we calculated how different customers come to the website over time by this assumption.



(a) The demography

(b) New customers over time

Fig. 4. Basic insights of Active Customers

As it can be seen from Fig. 4(a), we have almost similar number of users for different age categories and gender. Indeed, the difference is not much between different customer categories. We can see that the website has gained considerable amount of customers in the age category 50 to 65 at July 2015 (see Fig. 4(b)). However, the number of new customers in that age category that the website faced decreases a lot afterwards.

In this section, we take a look on general customer behaviour in terms of the usage of services,

3.2 Customers behaviour

In order to get an insight on how customers uses different services in the business, we performed a drill across operation on all fact tables except unknown click. *Drill across* refers to the operation that an analyst can relate different fact tables based on a conformed dimension (shared dimension). Here, we used the customer as the conformed dimension. Except the customer dimension, the following dimensions is also used in selecting the data, i.e. Date, Complain Type, Question, Message, Page. We, filtered the data in a way to use these dimensions value as the activity name. We selected data from the cube in the day level of granularity for the Date dimension. We selected the data from other dimensions based on their upper level of granularity. We consider the first occurrence of events for combination of our dimensions when selecting the events from the data warehouse. This means that, if there are multiple visits on a website by a customer in a specific date, we only consider its first occurrence event in that day.

The selected set of events is imported to Disco considering The customer is used as the case, and the day is used as the timestamp. Thus, it enables us to discover the behaviour of the customer on how she or he uses different services. The result is shown in Fig. 5. The frequency shown in this figure reflects the case frequency.

As it can be seen from the process, customers uses the website in the most of the cases. Then, the questions play very important role in the most of the cases (more than 80 percent of the cases). Message is also very important for customers (with more than 60 percent usage). Most of customers will continue to use website after sending a message; while some raises a question. There are also very small number of customers who issue a complain after sending a message.

We drilled through the cube, and dice two cubes reflecting the usage of services for two categories of customers, i.e. (bellow 50 years old and greater or equal to 50 years old). The section was made due to the difference between the number of customers in the last age category of customers which is explained before. We discovered the process in the same manner for those categories, so we skipped to demonstrate them here. It might be interested to know that there are totally 27,412 customers registered, among which 26,647 have visited the website.

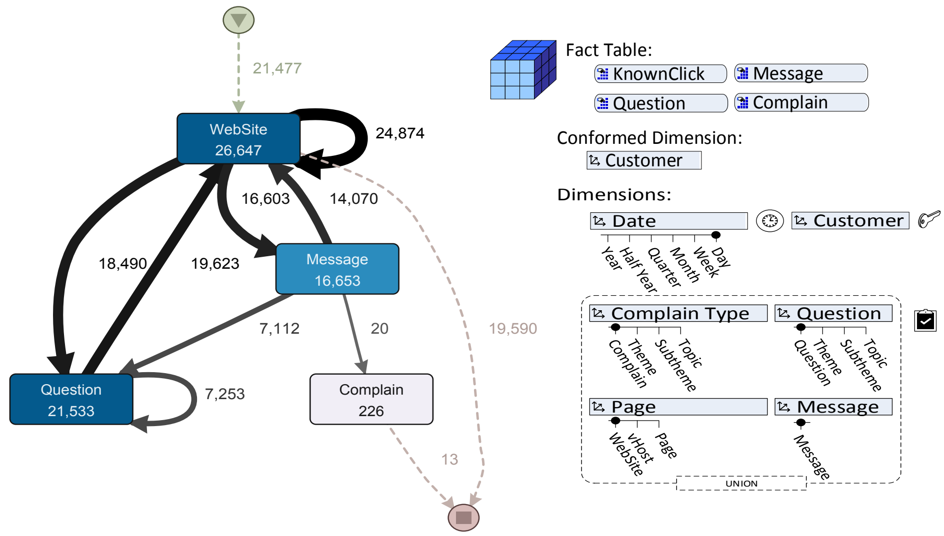


Fig. 5. A general view on the customer behaviour

3.3 Website visits

In order to explore how different customers explored different pages of the website, we used our plugin which demonstrate the social networks using the chord diagram [2]. The chord diagram is a good candidate to visualize dense social networks, and we used this diagram to visualize how different customers switched between different pages when visiting the website. We selected the top 20 pages that have been visited by users. We selected the set of events related to users visits of the website with a special setting. We consider the Website as the activity name for all events, and we set the page name as the resource information. We used the hand over of works metrics of social network analysis plug in in ProM to discover how users explore the website, and we visualized the result by using our plugin [2].

Fig. 6(a) shows the chord diagram which is mined from this log file. A chord diagram looks like a circle having its circumference divided in several fragments, i.e. arcs. Each arc can represent a resource of a social network. In this figure, each of these arcs represent a page in the website. The page which is used more has wider length. For example, we can see that the page tasks (taken) is used more than others.

In addition, arcs can be connected through chords. In our network, the chords represent moving from one page to another, e.g. a and b. In this example, the width of the chord attached to the arc a shows the amount of navigation from page a to b; while the width of the chord attached to the arc b shows the amount of navigation from page b to a. The color of the chord is set to the arc from which more navigation is performed to another.

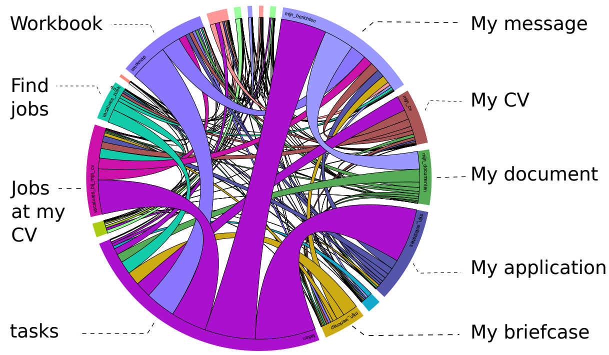
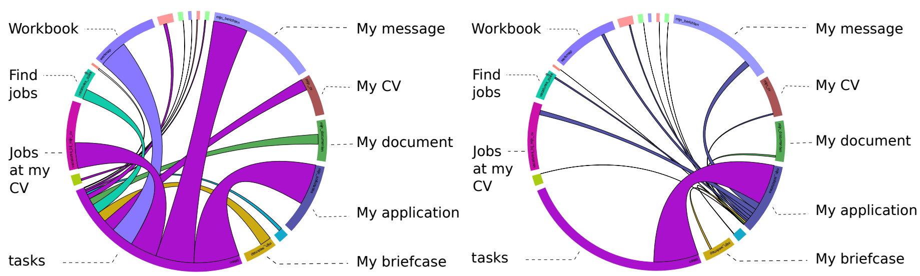


Fig. 6. A general view on the customer behaviour

For example, we can see there are navigations from page tasks (taken) to my message (mijn berichten). This chord is colored as the tasks page, so there is more navigation from tasks page to my message.

There are a lot of chords connecting arcs together, so it might be difficult to explore the navigations from one page to another. The diagram is configured in a way to only show the relevant chords when hovering the mouse over an arc. Fig. 7 shows two version of this diagram when having the mouse over tasks (taken) and my applications (mijn sollicitaties) pages.



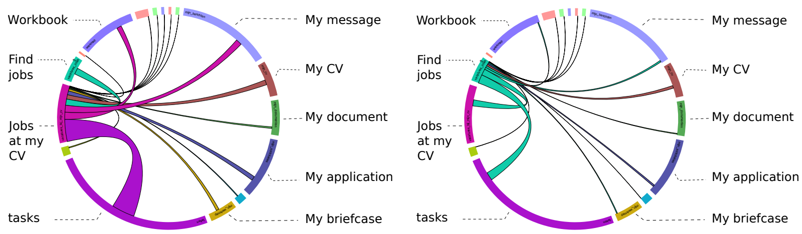
(a) from/to tasks page (b) from/to my applications page

Fig. 7. Pages navigations

As it can be seen from Fig. 7(a), users usually navigate more into pages my application, my cv, my message and jobs at my cv from the tasks page. They also mostly come to the tasks page from my briefcase, my document, workbook and find jobs pages.

As it can be seen from Fig. 7(b), users usually navigate more into pages my message, workbook and jobs at my cv pages from the my application page. They also mostly come to the my application page from tasks and my briefcase pages.

Fig. 8 shows two version of this diagram when having the mouse over Jobs at my CV (vacatures bij mijn cv) and Find Jobs (vacatures zoeken) pages.



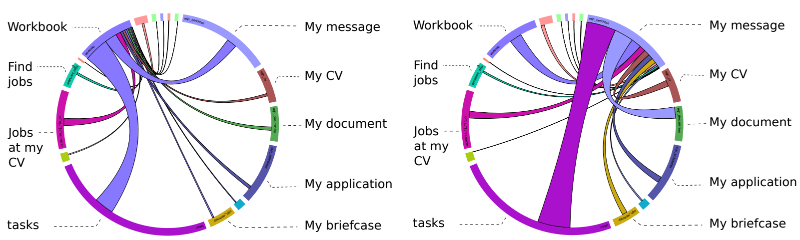
(a) from/to Jobs At my CV page (b) from/to Find Jobs page

Fig. 8. Pages navigations

As it can be seen from Fig. 8(a), users usually come to the jobs at my cv page, from the taken and my application pages. They also mostly navigate to the my message and workbook pages from this page.

As it can be seen from Fig. 8(b), users usually come to the find jobs page, from the my cv and my application pages. They also mostly navigate to the tasks, jobs at my cv and my message pages from this page.

Fig. 9 shows two version of this diagram when having the mouse over workbook (werkmap) and my message (mijn berichten) pages.



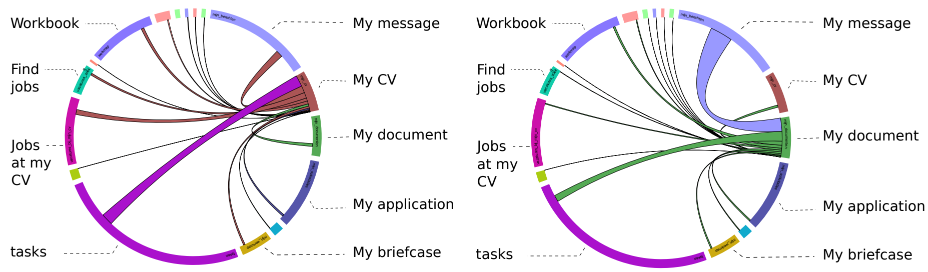
(a) from/to workbook page (b) from/to my message page

Fig. 9. Pages navigations

As it can be seen from Fig. 9(a), users usually come to the workbook page, from the jobs at my cv, my cv, and my document pages. They also mostly navigate to the tasks and my mesage pages from this page.

As it can be seen from Fig. 9(b), users usually come to the my message page, from the tasks, jobs at my cv and my cv pages. They also mostly navigate to the my document and workbook pages from this page.

Fig. 10 shows two version of this diagram when having the mouse over my cv (mijn cv) and my document (mijn documen) pages.



(a) from/to my cv page (b) from/to my document page

Fig. 10. Pages navigations

As it can be seen from Fig. 10(a), users usually come to the my cv page, from the tasks abd my document pages. They also mostly navigate to the my message, jobs at my cv, workbook and my briefcase pages from this page.

As it can be seen from Fig. 10(b), users usually come to the my document page, from the my message page. They also mostly navigate to the tasks page from this page.

Finally, Fig. 11 shows a version of this diagram when having the mouse over my briefcase (mijn werkmap) page.

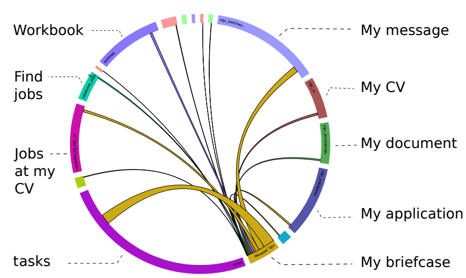


Fig. 11. Pages navigations from/to my briefcase page

As it can be seen from this figure, users usually come to the my briefcase page, from the my workbook page. They also mostly navigate to the tasks, my message and jobs at my cv pages from this page.

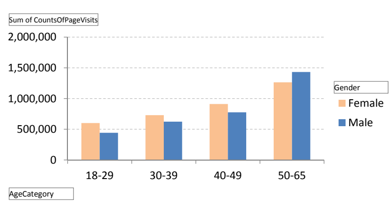
So far, we have seen how the users navigates through different pages in the website. However, it is difficult to have a holistic view based on these statistics. In next section, we will see their behaviours patterns in a more comprehensive way.

4 Website Visits

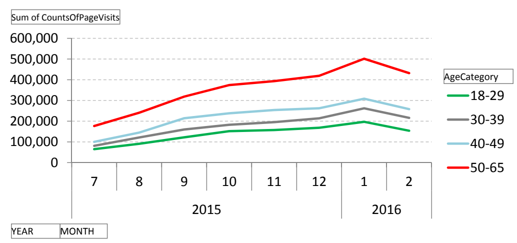
In this section, we focus on how different customers use the website. We start by considering total page visits statistics,

4.1 Page visits

Considering the demography of the customers, we can focus on how customers have visited the website. Fig. 12 shows how customers have visited the website, where Fig. 12(a) shows how much different customer demographies used the website, and Fig. 12(b) shows the usage of website by different customer demographies over time.



(a) The demography



(b) Changes over time

Fig. 12. Page Visits of Active Customers

As it can be seen from Fig. 12(a), as the age of customers increases, the number of their visits of the website is also increases. It also shows that female customer bellow age 50 uses the website more than male customers in the same age category, yet this relation is diverse for the customer above 50 years old. Fig. 12(b) also shows some interesting pattern in the use of website by different customers. As it can be seen from this figure, the use of website is generally increased over time for different age categories until January 2016. The previous trend that we saw from Fig. 12(a) is also valid here, i.e. as the customer age increases the number of visits increases. It is interesting that the number of visits is decreased after January 2016 for all customers demography.

4.2 Usage of the website

We explored different settings based on which we discovered processes. One important aspects that we found is about how the website is used in general by customers based on the number of times that they visit the website. In particular, we looked at the number of sessions that customers visited the website based on the session. Fig. 13 shows the relation between the number of customers and the number of sessions that customers visited the website, where the x-axis shows the number of sessions and the y-axis shows the number of customers. The blue line shows the number of customers that has visited the website based on the number of sessions.

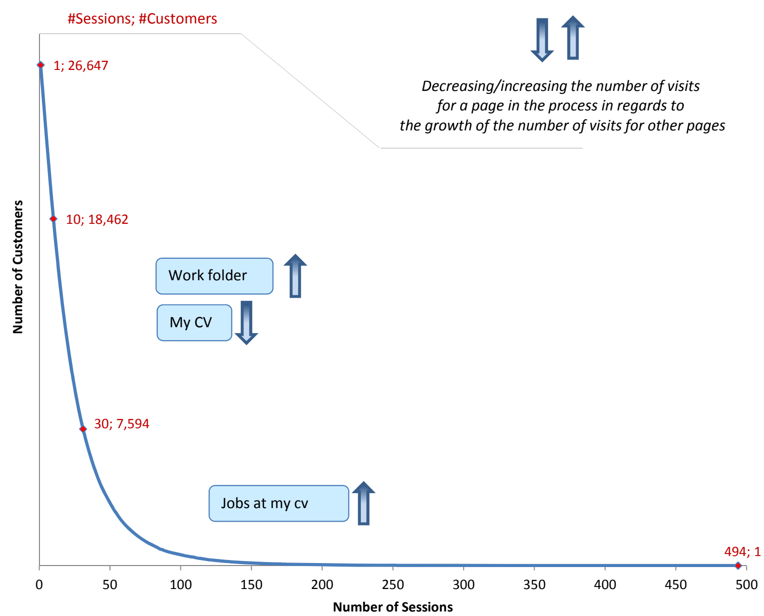


Fig. 13. Customer Sessions

We realized some change in how people uses the website based on the if they continue using the website. The change is summarized in the figure by demonstrating the page that will be used more or less in regards to the number of sessions. In particular, the use of the page called "My CV" is decreased after the tenth session, while the use of the page called "Work folder" is increased after the tenth session. The use of the page called "Jobs at my CV" is increased after the thirtieth session. Thus, we filtered the events based on the session numbers into three categories, i.e.:

- Category I: The page visits for the session numbers between 1 and 10.
- Category II: The page visits for the session numbers between 11 and 30.
- Category III: The page visits for the session numbers greater than 30

Different sets of events are selected to discover process models for each category. Fig. 14 shows the filtering setting of the fact and dimensions that are used in retrieving these data. Although the same setting is used to discover the processes for each category, the filtering on the session number is different for each category, which is corresponds to the session numbers that each category includes.

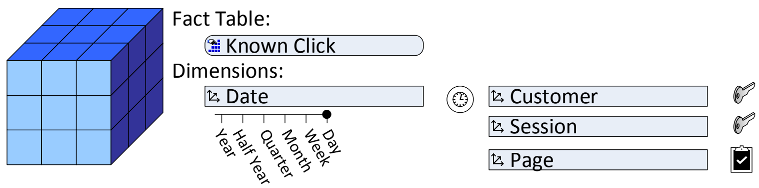


Fig. 14. Filtering options for selecting evens sets for each category

Note that we used both customer and session as the key for mining processes, so the case frequency will not show the number of customers. Instead, it represents the number of sessions by customers. We show the process and explain the change as follow. We filtered the discovered process to show only five activities, so we can present it in this paper.

Category I

The first set of events represent the pages that customers visit during their first ten sessions. This can represent the behaviour of the new customer who

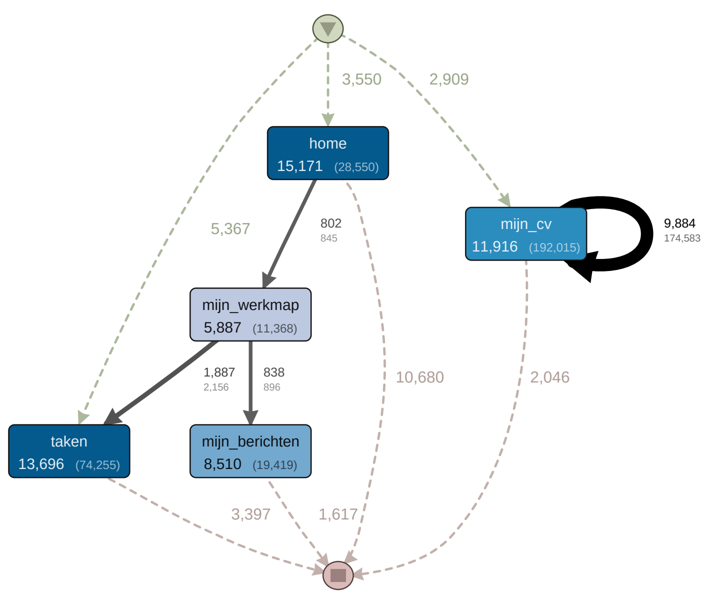


Fig. 15. Pages visited by customers during their first 10 sessions

needs to feed their information into the website. We assume each session of a customer as a case, so different sessions in which a customer visited the website is considered as different cases. Fig. 15 shows the process that is mined using Disco based on this set of events.

As it can be seen from this process, there are two pages except the home page that customer visits more than others. In most of the cases, customers visited both "taken" and "mijn_cv" pages - representing the "tasks" and "my cv" pages. The frequency number for visiting the "my cv" page is much more, which is 192,015. This shows that customers develop their CVs when they are newbies. It is also interesting that some customers use to follow a specific path when visiting these pages, i.e. from home page to mijn_werkmap (my work folder), and from mijn_werkmap to either taken or mijn_berichten (my messages) [2].

Category II

The second set of events represent the pages that customers visit during the sessions with number 11 till 30 . This can represent the behaviour of the customer who already is developed her or his CV. We applied the same settings

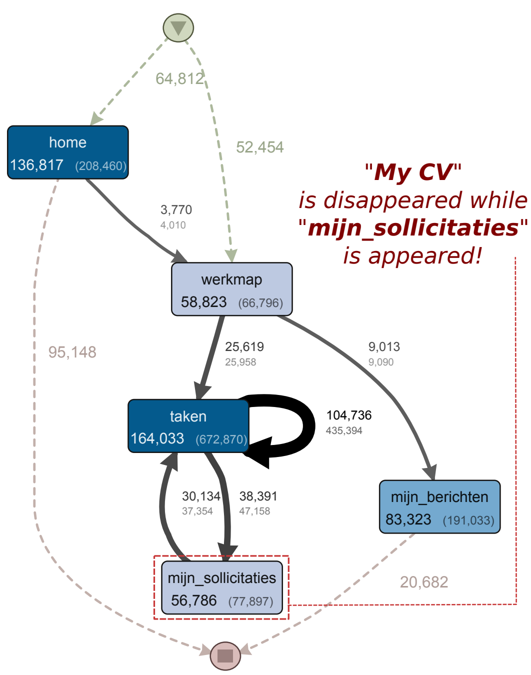


Fig. 16. Pages visited by customers during sessions 11 until 30

[2] We used google translation for translating some of pages names

as previous category yet applied different filter option for the session number. Fig. 16 shows the process that is mined using Disco based on this set of events.

As it can be seen from the mined process model, the overall process is still the same. However, the use of “My CV” page is decreased, which can be an indication for that the customers has already developed their CV well enough. It should be noted that this page still exist in the process model, but it’s frequency is not as high to be included in the most five top pages. On the other hand, we can see that the page “mijn_sollicitaties” (my application) is appeared in the five top used pages. This might be due the interest of customers in pursuing their applications that they made. This page is accessed mainly after the page tasks, which can indicate that customers are redirected from that page.

Category III

The last set of events represent the pages that customers visit after the session 30 (session number>30). This can represent the behaviour of the customer who has long been in the process. We applied the same settings as previous category yet applied different filter option for the session number. Fig. 17 shows the process that is mined using Disco based on this set of events.

As it can be seen from the mined process model, the overall process is still the same. However, the use of “mijn_sollicitaties” (my application) page is decreased, which can be an indication that the customers have already applied for

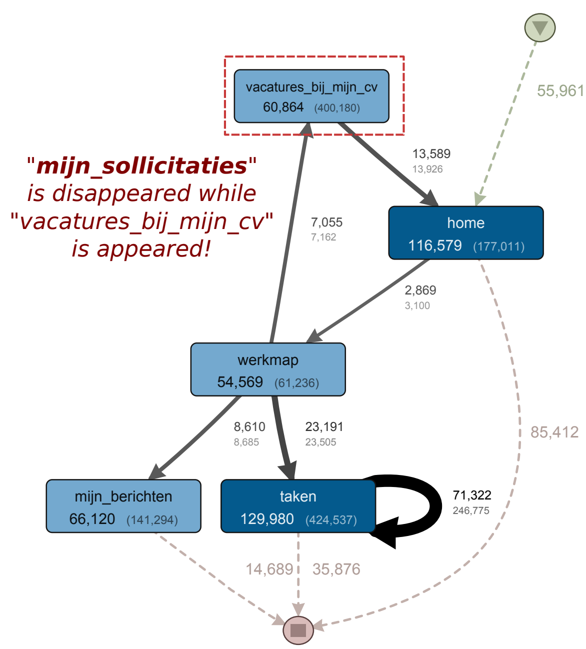


Fig. 17. Pages visited by customers during sessions 31 till 494

enough jobs, and they do not apply as before. It should be noted that this page still exist in the process model, but it's frequency is not as high to be included in the most five top pages. On the other hand, we can see that the page "vacatures_bij_mijn_cv" (jobs at my CV) is appeared in the five top used pages. This might be due the follow up activities that these customers perform.

There are many other differences that we skip to describes, since there are more interesting results to present.

4.3 The behaviour of customers before logging in

There are two fact tables recording the visits of users who are logged in the system (Known click fact table) and those who surf the website anonymously (unknown click fact table). We are interested to investigate the behaviour of customers before they have logged in to the system. To perform such an analysis, we need to drill across these two fact tables to retrieve the events of known customers before logging into the system. Thus, we used the session dimension

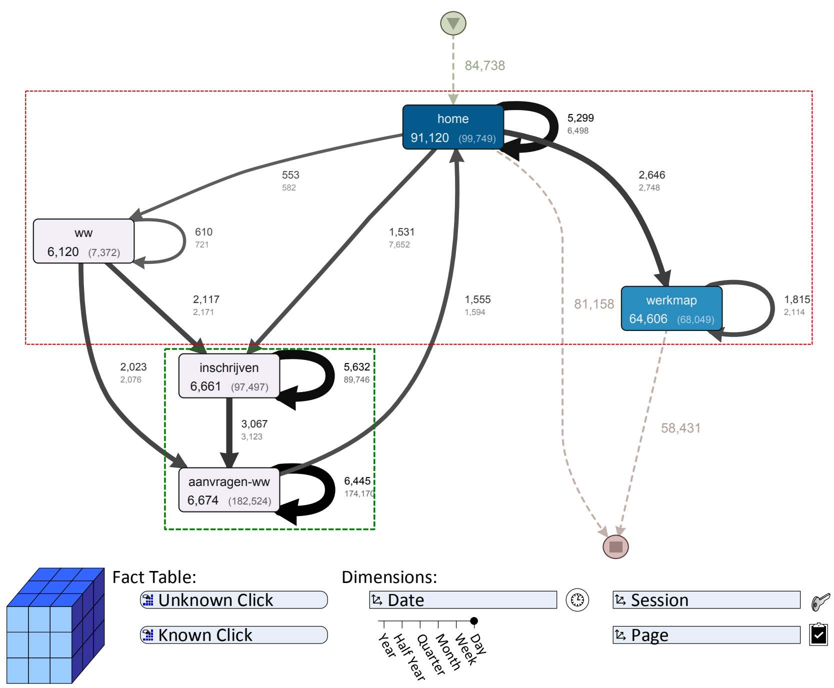


Fig. 18. A general view on the customer behaviour by drilling across known and unknown click fact tables

as a conformed dimension to perform such filter. We selected the events from the unknown click fact tables for sessions for which the events are also recorded in the known fact table. We selected all events from the unknown fact table before the first event that happens in the known fact table. Fig. 18 shows the rest of the configuration and the process which is mined in Disco.

As it can be seen in this figure, we can recognize two categories of pages when these customers have not yet logged in. These categorises are differentiated based on the difference between two measures, i.e. absolute frequency of the page and the case frequency. In this figure, we only see few pages since we limit the number of pages for the sake of the presentation. Among these pages, we can see that *home*, *werkmap* and *ww* pages are belong to the first category (annotated with the red area), where these two measures are close. *inschrijven* (register) and *aanvragen-ww* (request-ww) belong to the second category (annotated with the green area), where there is a considerable difference between these measures.

The second category of pages are very interesting. For example, the register pages is expected to be filled by customers once, but it is frequently has been visited. As it can be seen, it has been widely revisited during a session, which can have different reasons. It might happens because the users need to enter much data when registering or search for some data that they know. This is very tentative assumption, and more in depth analysis can be performed in case of having access to technical information about the pages and architecture of the system.

5 Complains

In this section, we explore users behaviour which might lead them to issue a complain, and the potential change in their behaviour before and after complaining.

5.1 The customer behavior

We have seen the overall picture of the users behaviour previously in Fig. 5. We explore all users and all activities that can happen in the business in that analysis. Here, we like drill down into the data warehouse to explore the customers who have at least issued a complain once.

Fig. 19 shows the overall picture of how we perform the drill through to explore different possibilities using our data warehouse. First, we dice our cube to filter only the data that includes the users events who has issued a complain once, i.e. Fig. 19(A). Then, we drill down into the theme level at the complain type dimension (Fig. 19(B)), and we further continue to drill down into the subtheme level at the same dimension (Fig. 19(C)). We also drill down from our diced cube which was drilled into the theme level of the complain dimension into the Page level of the Page dimension (Fig. 19(D)). We refer to each of these scenario by the letter mentioned in the figure.

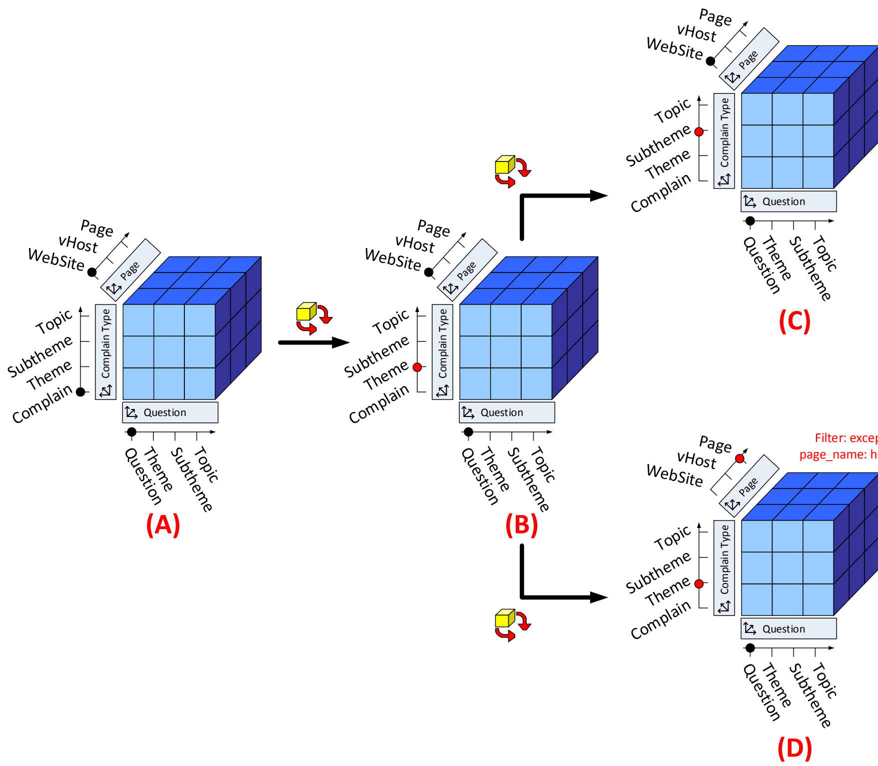


Fig. 19. Drilling Scenario

Scenario A

First we diced our cube to have a sub-cube that includes all events of customers who have issues a complain at least once. For detail about the configuration and the structure of the main cube, you can look at Fig. 5. We have four dimensions involved, i.e. Page, Question, Complain Types and Message. First, we rolled up into the most coarse-grained level of granularity in each dimension. Fig. 20 shows the process which is mined based on this cube.

There are totally 289 complains made from 226 customers. 180 customers only complained once, and 46 customers complained more than once. As it can be seen from the figure, customers used different services actively. We can see that once of the 46 customers, who complained more than once, issued the complain after the previous complain. 177 of these customers has used message, and they used it frequently since the absolute frequency is much higher. 214 customers asked questions, and they have asked many questions since the absolute frequency is high.

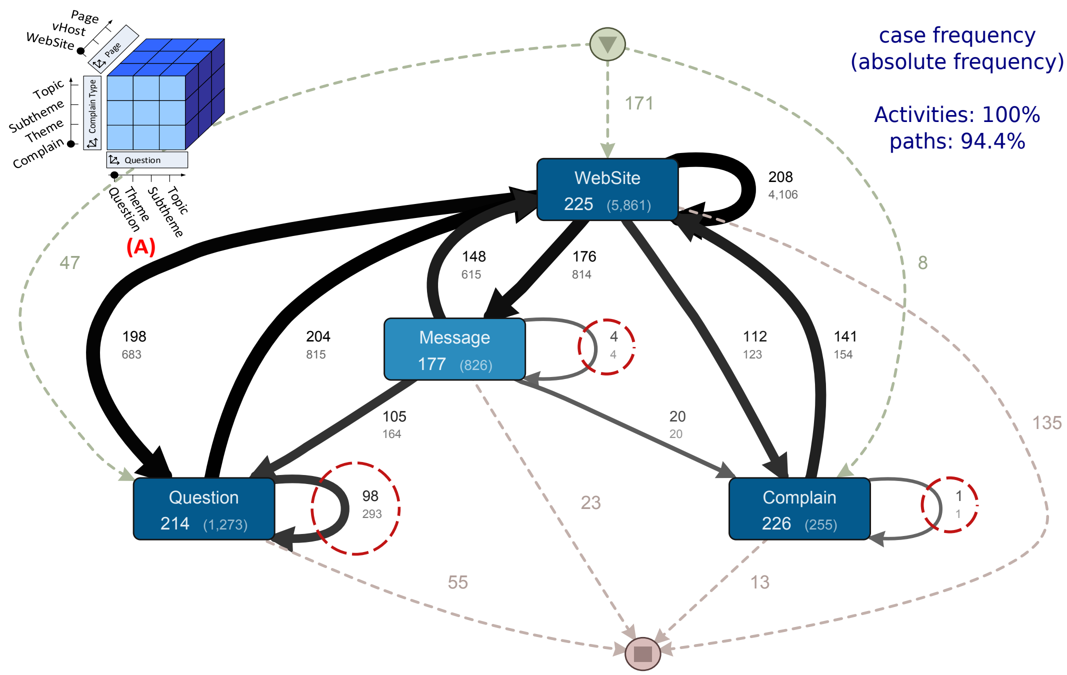


Fig. 20. Drilling Scenario A

Scenario B

In previous scenario, we explored the overall behaviour of people in the business in a very hight level of abstraction. We are interested to drill down into the complain dimension into the theme level. Fig. 21 shows the process which is discovered by this operation. We set the settings in disco for Activities to 100% and paths to 45.8%.

As it can be seen in this figure, the complain activity is disappeared, and three more activities are appeared as a result of this drilling down, i.e. Complain:Service, Complain: duration of treatment, and Complain: treatment (attitude/behavior). The other difference is on the arrows and their frequencies in the process model. For example, it was a relation from message to question in previous process, yet it disappeared in the new mined model. Moreover, the frequencies numbers for different arrows are changed, e.g. the self loop for message and Question activities. This change happens since when we drill down into a dimension, the order of events in the log is changed because of changing the level of granularity.

From this model, we can see that the major complains are made due to services, which is 203 cases. It is also interesting that 33 complains is about treatment - which can be further investigated.

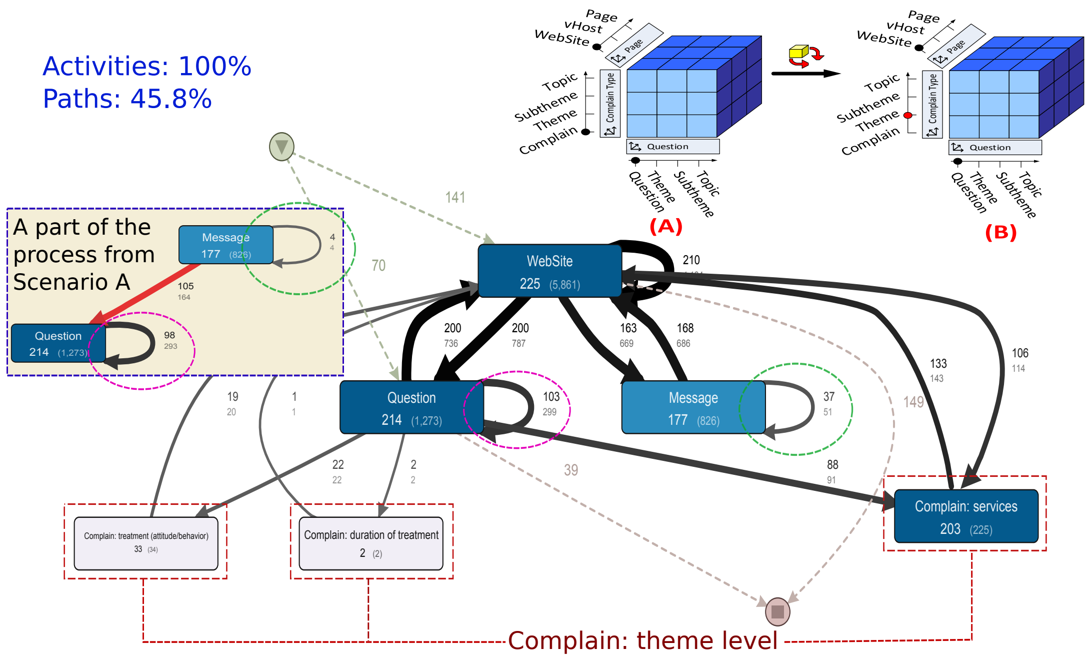


Fig. 21. Drilling Scenario B

Scenario C

We could see that the most of complains are related to the services category from the previous mined process. Thus, it might be interesting to drill down further into the subtheme level in the complain dimension to see how the process look like. We diced the cube also based on all complains which are in the services theme. The result is shown in Fig. 22.

As it can be seen from the figure, four activities are appeared in the mined model, i.e. “information/communication to the customer”, “payment”, “availability/accessibility” and “support/handling” - mentioned from the most happened to less. We can see that from the customer and question, some users ended up to complain about the “support/handling”. Most of users complained about “information/communication to the customer” after sending a message to the company. Some of them even issued another complain about the payment afterwards, and it is interesting that the mean duration between the times of these activities is instant - showing the potential relation between these activities. The complains about “availability/accessibility” usually come from the website, yet sometimes after customers send messages. It seems that it might be potential points for the company to improve the process to communicate to customers more effectively in order to reduce the number of complains significantly.

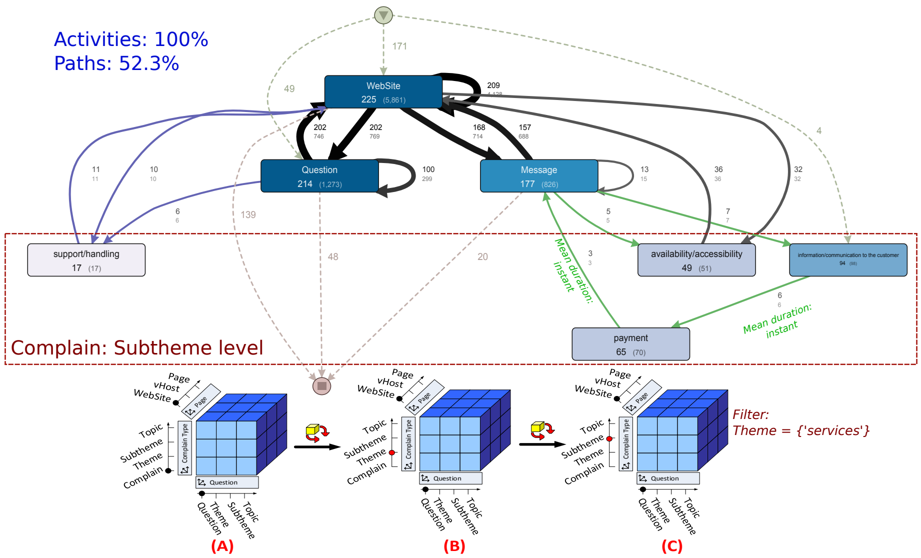


Fig. 22. Drilling Scenario C

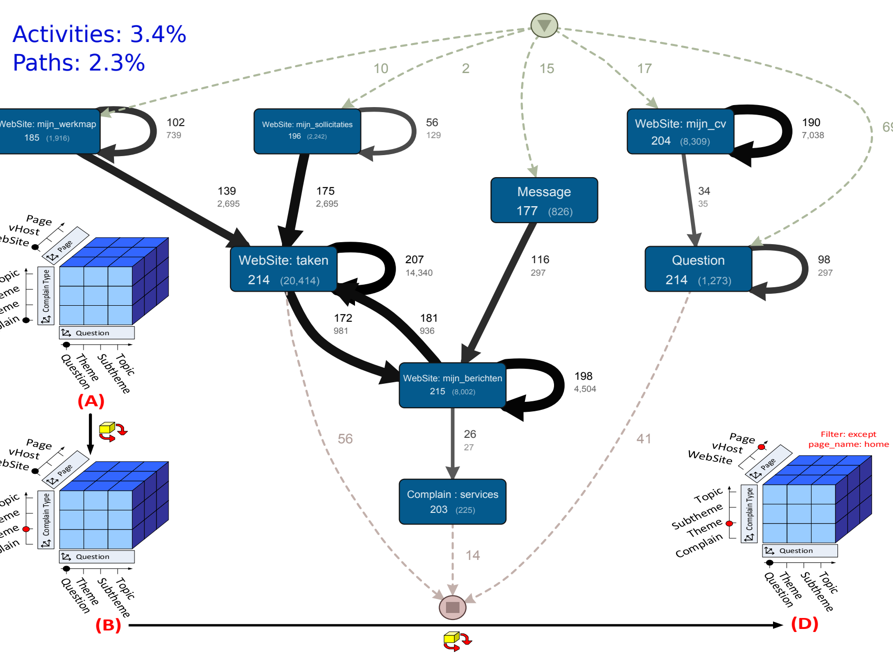


Fig. 23. Drilling Scenario D

Scenario D

In order to see the relation among different pages which are browsed in the website and issued complains, we drilled into the page level of the page dimension from the cube which was discussed in scenario B. We also exclude the home page, since it can be frequently used by users for navigation purpose. Fig. 23 shows the process which is mined from the diced cube.

As it can be seen from this figure, most of the users who complained browsed the my message page in the website after contacting to the company through message. Afterwards, some of them complained about the services.

5.2 Complains and channels

We have explored different possibilities of the general user behaviour using different dimensions and fact tables so far. In this section, we will explore the potential relation between different channels that users have used in sending messages and the complains. Therefore, we sliced our cube based on the name property in the Channel hierarchy. In this analysis, we only consider two fact tables, i.e. Message and Complain. We explore the relations by drilling into three levels in the complain type dimension, i.e. complain, theme and sub-theme.

Drilling into the complain level

Fig. 24 shows the process which is mined from the selected cube, where our focus is on the highest level of granularity at the complain type dimension. Both

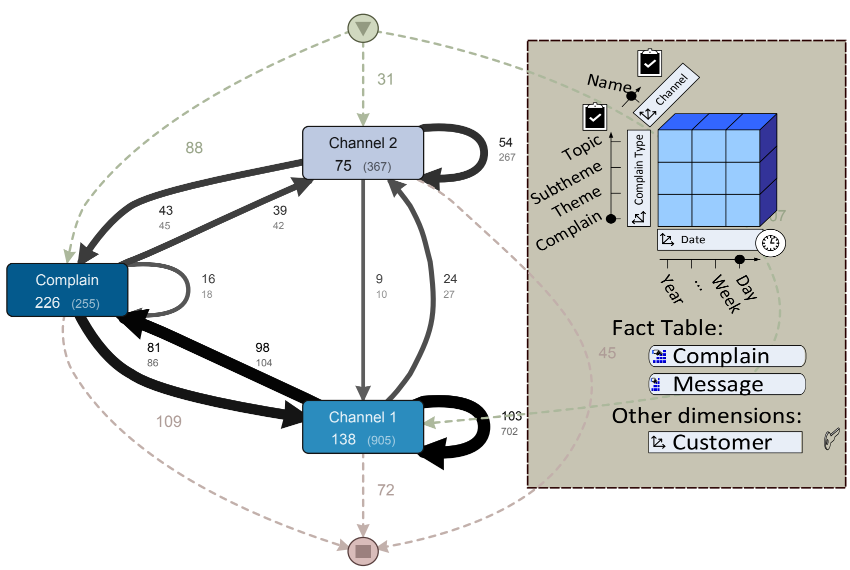


Fig. 24. Complains and Channels

Activities and Paths are set to 100% in Disco when discovering the process model.

The channel has no hierarchy, so the members are appeared in the mined model, i.e. channel 1 and channel 2. We explore the Complain Type hierarchy at its highest level of granularity, so only one activity named complain is appeared in the process model. In this process, we can see users have used channel 1 and 2 to issue a complain. The channel 1 is used more than the other one. There are also some cases, that a user switched between these channels.

Drilling into the theme level

It might be interesting to drill into the complain type dimension to see the relation between different channels and the theme of the complains. Thus, we drilled into the theme level of the complain dimension and mined the process model for the new set of events. The Activities setting is set to 100% while the Paths setting is set to 50% when discovering this model.

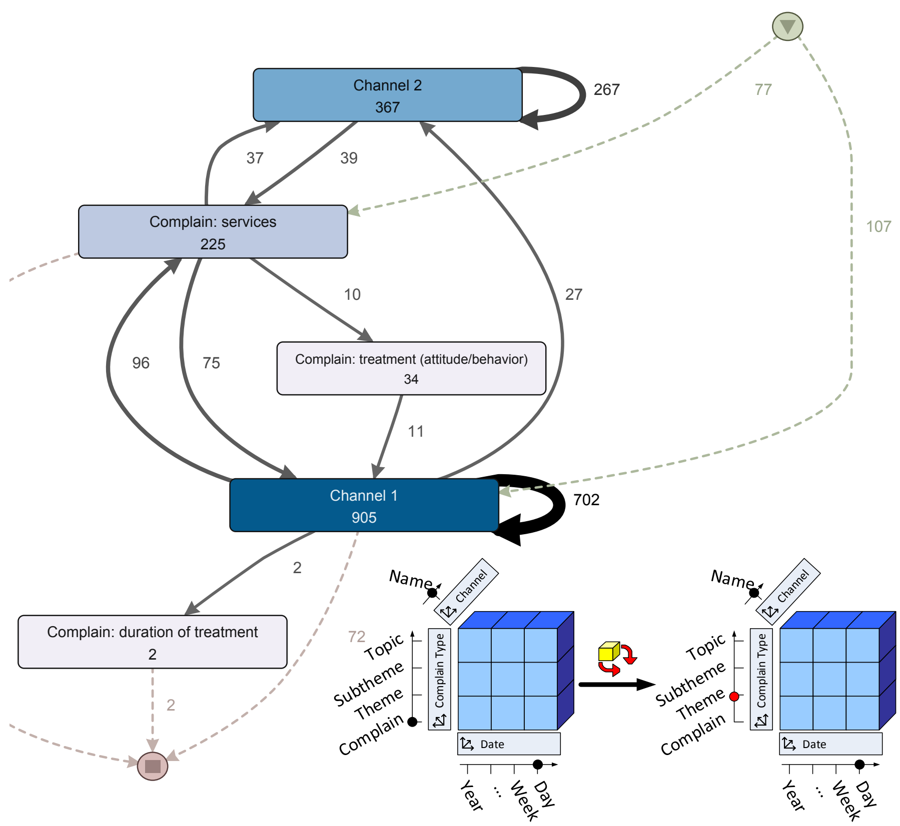


Fig. 25. Complain and Channels Drilled into the Complain Theme level

Fig. 25 shows the process which is mined by such filtering. As it can be seen from this figure, channels one and two are mostly used for complaining about the services. There are some cases (10 cases) in which the user complained about the treatment after complaining about the services.

Channel one is also mostly used for complaining about the duration of the treatment. Since this channel is used mostly by users, it might need more resources to reduce the time of treating a case. We also considered the performance analysis in Disco for this process. The outgoing flow from channel two has 17.5 days as median duration. All outgoing flows from channel one has more or equal value compared to channel two, so it might be a good point to consider if the performance of this activity can be improved.

Drilling into the sub-theme level

We continue to drill into the subtheme level of the complain dimension. Fig. 26 shows the mined process. We set the Activities setting to 50% and Paths setting to 31% for this analysis in Disco.

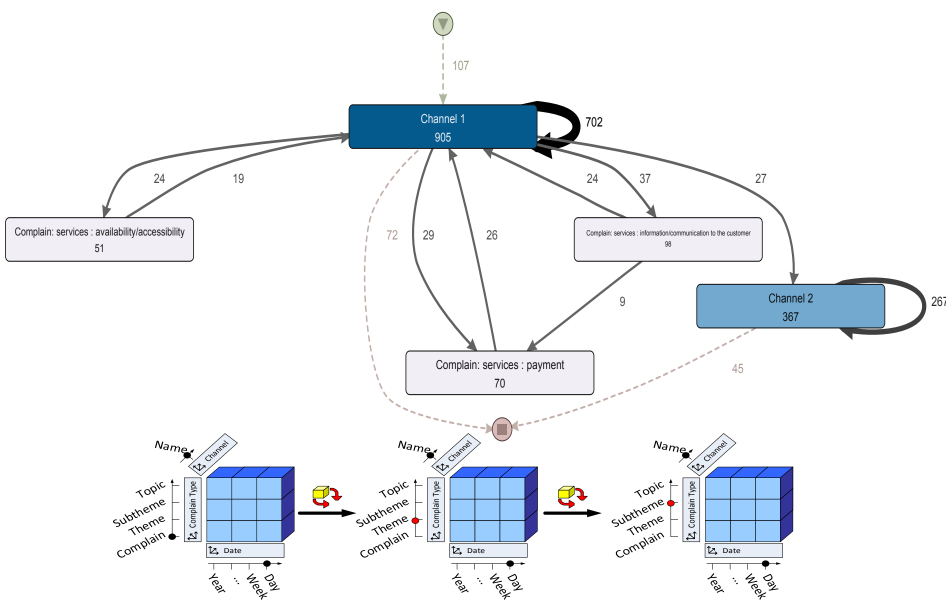


Fig. 26. Complain and Channels Drilled into Complain Subtheme level

As it can be seen from this model, the first channel is mostly used for three sort of complains, i.e. availability/accessibility, payment and information/communication to customer. There are also nine customers who complained about the payment after complaining about the information/communication to customer. This point can be considered as a potential point to improve the process in order to reduce the number of unnecessary complains.

We showed the potential of identifying different aspects of a business process by applying different cube operations so far. Now, we will try to see if we can explore the changes in user behaviour before and after making a complain.

5.3 The Change in customer behaviour

We have explored different scenarios in which users issued a complain. In this section, we will explore how the behaviour of users changes before and after making a complain. To explore this behaviour, we divide each trace in which a user has complained into two parts, i.e. before and after making a complain (see Fig. 27).

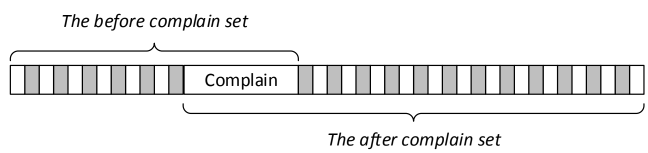


Fig. 27. Dividing the events that contains complain event into two sets

The set of events from the beginning of each trace until the first complain is considered as the trace that represent the behaviour of users before making a complain. The set of events from the first complain until the end of the trace is considered as the trace that represent the behaviour of users after making a complain. We will see how these processes are different as follow.

5.4 Before the complain

To discover the behaviour of the customer before making a complain, we exclude the website clicks from our list of fact tables and dimensions. We used a new cube, using complain, message and question fact data. We used the customer, question, date, complain type and channel dimensions. The settings and are annotated in Fig. 28, where the mined process is also represented.

As we can see from this figure, users asked questions before complaining. However, some of the messages sent from channel 1 resulted in complaining about the service. This process is not very helpful to understand more aspects and relations among different activities. To perform more in depth analysis, we used MINERful miner [1] in ProM to discover the declarative model for the same set of events. Fig. 29 shows the declarative process. We set the support feature of this plug-in to 0.9.

The declarative models try to explain the process model based on constraint exist in the events. As it can be seen from this process, the message from both

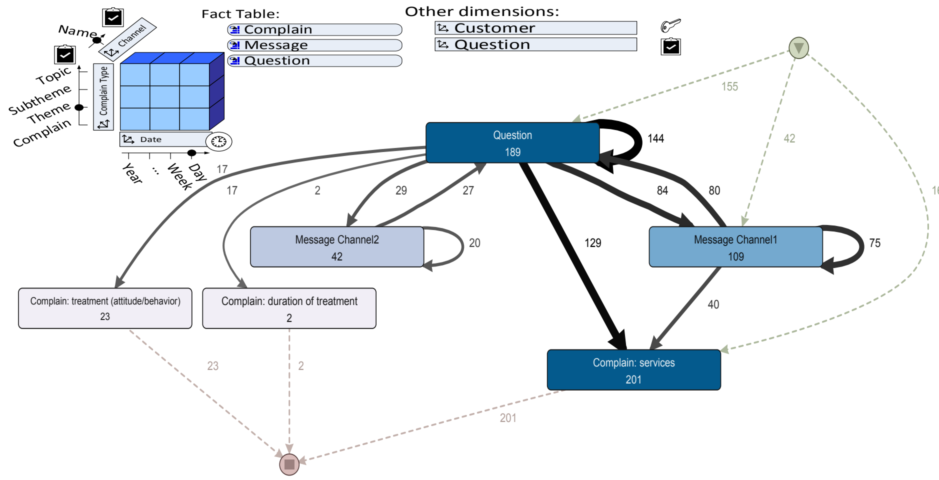


Fig. 28. A way to Complain

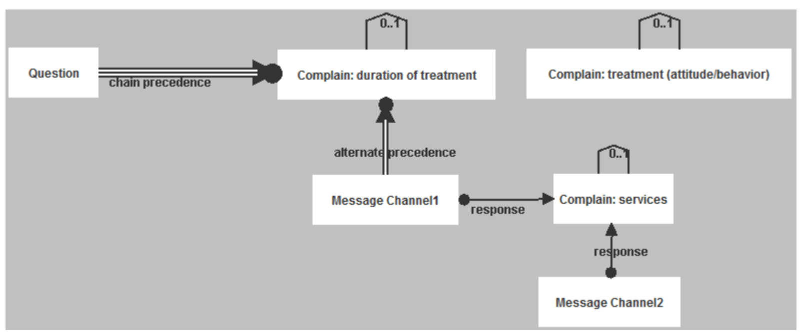


Fig. 29. A way to Complain - Declarative version

channels can result in complaining about services. However, if we see the cases in which a user complained, we can see that users send a message through channel 1 before complaining about the duration of treatment (alternate precedence relation). We can also see that users asked a question before making the complain about the treatment. We can assume that users asked a question, and since they have not been happy they send a message and finally made a complain about how they have been treated. Note that we set the support to 0.9, so this rules is not always true in the log file but it is in the most of the cases.

5.5 After the complain

We can explore the if there is any difference between the behaviour of customers before and after complaining by slicing the cube. Fig. 30 shows the overall process which is mined by filtering events.

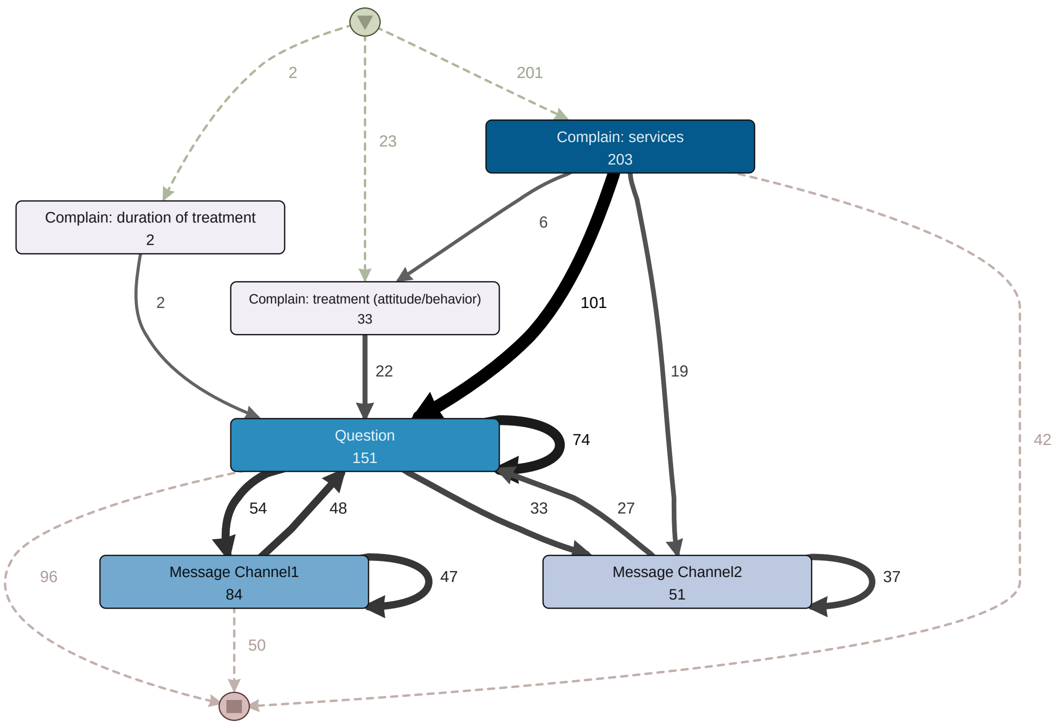


Fig. 30. A way from complain

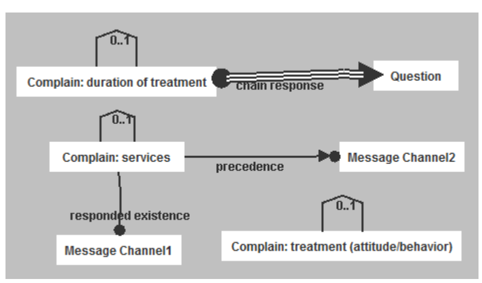


Fig. 31. A way from complain - Declarative version

As it can be seen from this process, the overall trend still is the same. We also mined the declarative version of the process with the same setting as previous declarative model, which can be seen in Fig. 31. The semantics that is described by the declarative and imperative process models actually describe the same issue with different process modelling language.

6 Conclusion

In this report, we explored how users have used different services in the Dutch autonomous administrative authority based on information exist in the event log. We have created a date warehouse, which is used as a foundation for our analysis. We explored different possibilities by filtering the sets of events, and applied different process mining techniques to discover how different users use different services. The result shows potential points which can be used to improve the business process.

This report also shows how structuring and denormalizing events based on multi-dimensional data warehouse design principle can help us to apply more effective process mining analysis. In this way, events can be organized into a process cube, which enables us to apply advanced filtering based on cube operations. We have the possibility to apply drilling up and down into different level of granularities. We also can drill across several fact table to discover more advanced process mining tasks.

References

1. Claudia Di Ciccio and Massimo Mecella. Minerful, a mining algorithm for declarative process constraints in mailofmine. *Department of Computer and System Sciences Antonio Ruberti Technical Reports*, 4(3), 2012.
2. Amin Jalali. Supporting social network analysis using chord diagram in process mining. In *International Conference on Perspectives in Business Informatics Research*. Springer, 2016.
3. Ralph Kimball and Margy Ross. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.
4. Wil MP van der Aalst. Process cubes: Slicing, dicing, rolling up and drilling down event data for process mining. In *Asia-Pacific Conference on Business Process Management*, pages 1–22. Springer, 2013.
5. M. Dees; B.F. van Dongen;. Bpi challenge 2016: Clicks logged in, 2016. Available at `http://dx.doi.org/10.4121/uuid:01345ac4-7d1d-426e-92b8-24933a079412`.
6. M. Dees; B.F. van Dongen;. Bpi challenge 2016: Clicks not logged in, 2016. Available at `http://dx.doi.org/10.4121/uuid:9b99a146-51b5-48df-aa70-288a76c82ec4`.
7. M. Dees; B.F. van Dongen;. Bpi challenge 2016: Questions, 2016. Available at `http://dx.doi.org/10.4121/uuid:2b02709f-9a84-4538-a76a-eb002eacf8d1`.
8. M. Dees; B.F. van Dongen;. Bpi challenge 2016: Werkmap messages, 2016. Available at `http://dx.doi.org/10.4121/uuid:c3f3ba2d-e81e-4274-87c7-882fa1dbab0d`.
9. M. Dees; B.F. van Dongen;. Bpi challenge 2016: Werkmap messages, 2016. Available at `http://dx.doi.org/10.4121/uuid:e30ba0c8-0039-4835-a493-6e3aa2301d3f`.