Usage analytics Using Process Mining
Key Findings for the Dutch Employee Insurance Agency

Farideh Heidari , Nour Assy

Department of Mathematics and Computer Science, Eindhoven Univery of Technology,
The Netherlands
{f.heidari; n.assy}@tue.nl

Abstract. Web analytics is the process of collecting, analysing and understanding web data for the purpose of optimizing web performance and design. Different web analytics techniques exist. They differ in the type of data being analysed and the purpose of the analysis. In the context of the BPI challenge 2016, clicks data of not-logged-in and logged-in customers of a Dutch Employee Insurance Company (UWV) are provided for analysis. UWV aims to understand the usage process of its website in order to make business-driven decisions that foster user engagement and optimize customer journey. More precisely, it looks for valuable insights into usage patterns based on customers' demographics, changes in the customers' behaviour over time and the way its different communication channels are being used. In this paper, we use Process mining for Click-Data analytics. The provided data are analysed from different perspectives and actionable insights for optimizing the UWV website usage are discussed.

Keywords: Usage analytics, process mining, web analytics, click analytics.

1 Introduction

The UWV (Uitvoeringsinstituut Werknemersverzekeringen) in an employee insurance agency as an independent administrative office commissioned by the Dutch Ministry of Social Affairs Employment (SVZ). The agency provides employee insurances, labor market and data services for its customers.

At the end of 2015, DigiD was implemented as the single port (Werk.nl) to access digital environment and services providing a safe and easy gateway to UWV services.

Many services provided, forms and transactions have been digitized (e.g., 95% of the applications for benefits are digitalized). Moreover, unemployment people get only online support for the first 3 months. Through work.nl or Local Job Center (werkplein), the unemployment benefit (WW uitkering) can be applied. This includes the people who are make redundant (workloss), (older) job seekers (werkzoekende) or unable to work (arbeidsongeschikt).

Through www.werk.nl, one can regulate related matters to unemployment such as requesting for unemployment benefit, searching for vacancies, reintegration to the workforce and uploading CV's for the employers [1] . As soon as one receive the WW benefit, that person is obliged to look for a job. The "Werkmap" (workfolder) is an instrument used to track such legal obligations. Moreover, the site www.werk.nl assists the customers on such matter and collects open positions and allows customers to find suitable jobs. Customers can search for vacancies (logged-in or not-logged in).Customers get the first 3 months of unemployment online support via werk.nl. Then UWV consultant work can invite job seekers through the Werkmap for a period of e-coaching. We develop for job seekers online training and webinars that support them in their search for work.

If the first day of unemployment is known, one can can apply online by visiting werk.nl (UWV WERKberdift). For the application, one will need his/her DigiD[1] code as well as your BSN (social security number). Applicants can acquire information about the process and rules and regulations through www.werk.nl and apply for benefits or upload their CV's through www.login.werk.nl which demands for logging with DigiD.

The website also provides required information on the procedure, rules and regulations as well as answers to frequently asked questions or messages from UWV. If a customer cannot find the answer s/he is looking for on the site, s/he can contact through the call centre enabling posing questions. Besides through the website, one can send a message or submit a complaint. For posing questions and submitting complains as well sending a message, BSN number should be registered.

UWV is interested in discovering the way both sites www.werk.nl and www.digid.werk.nl are used to improve their usability and performance[2]: when customers move from one contact channel to the next and why and if there are clear customer profiles to be identified in the behavioural data. In particular, the agency is interested in getting insights into customers' journey and the usage patterns across various challenge. Such interests are reflected in the following questions:

1. Are there clear distinct usage patterns of the website to be recognized? In particular, insights into the way various customer demographics use the website and the Werkmap pages of the website are of interest (detailed in Section 4).
2. Do the usage patterns of the website by customers change over time? Do customers visit different pages when they start using the website versus when they have been using the website for some time? How does the usage change over time? (detailed in Section 5)
3. When is there a transition from the website to a more expensive channel, such as sending a Werkmap message, contacting the call centre or filing a complaint? Is there a way to predict and possibly prevent these transitions? (detailed in Section 6)

[1] DigiD (short for Digital Identification) is a form of online ID that allows you to access many services and government websites in the Netherlands. The DigiD consists of a username and password that are linked to your personal public service number (BSN).

[2] http://www.uwv.nl/overuwv/wat-is-uwv/hoe-werken-we/detail/dienstverlening

4. Does the behaviour of the customers change after they have send a Werkmap message, made a phone call or filed a complaint? Are customers more likely to use these channels again after they have used them for the first time? What is the customer behaviour on the site after customers have been in contact through the Werkmap or by phone? (detailed in Section 6)
5. Is there any specific customer behaviour that directly leads to complaints? (detailed in Section 6)

To enable answering the aforementioned questions, five sets of data are provided that will be elaborated further in Section 2. This paper aims at answering the above questions through process mining techniques. Process mining [2] aims to transform event data recorded in information systems into knowledge of an organisation's business processes.

To conduct the analysis a workflow is developed including three main phases of initialization, data preparation and process mining. In each phase different tools were used to serve the required purposes. Statistics on the data considering demographics of the customer were provided. Following a purposeful strategy, data was processed and filtered to be able deal with inconsistency and reduce its complexity to be able to answer the aforementioned questions. Analysis was conducted in three categories of "usage patterns", "usage change" and "expensive usage" and while appropriate incorporated with demographic analysis. Transitions and changes are taken into account and the usage of different channels in different stages are discovered. On the basis of the results, meaningful observations are made providing answers to questions and consequently recommendations for UWV are provided.

This paper is organized as follows: Section 2 elaborates on the workflow of the analysis and the steps taken. Section 3 provides insights into data and data preparation phase. Sections 4-6 discuss the analysis results. Finally, the paper concludes in section 7 with set of a conclusions and recommendations.

2 Methodology Workflow

Inspired by PM^2 [3] (i.e., a methodology to guide the extraction of process mining project), the workflow of the analysis is developed considering the research questions. This workflow outlines the steps taken in conducting the analysis. The workflow (Fig. 1) consisted of three main phases: *Initialisation*, *Data Preparation*, and *Process mining* followed by a discussion on the finding. These phases were conducted in iterations to answer different research questions.

In the first phase "*Initialisation*", the domain was understood and the require data is acquired. There was a need to understand the goal of UWV, targeted customers and their types, the services provided via werk.nl and login.werk.nl, the specifications of such services (e.g., e-support in the first 3 months), etc. Looking at different data sets, the terminology behind each data field was understood while the governmental rules and regulations behind it are taken into account. This was vital in our views to be able to interpret the results and recognize different customer's behaviour demanding different services. Such knowledge governed the whole analysis.

The second phase included "*Data Preparation"* with the goal of creating event logs in such a way that it is helpful in conducting analysis and answering research questions. In such phase, the overall view acquired through understanding the domain is obtained through conceptual modelling [4] and statistical analysis. This also governed the way the notion of cases are defined and later the data is prepared.

In the first step, the overall view "*Bird's eye view*" was acquired through discovering the number of cases, variety of cases, the performance related information (e.g., mean duration) etc. This led to the development of conceptual models depicting the foundation of available data and the relationship between different fields [4] . Concurrently, data logs were cleaned, irrelevant data or noises were filtered out and when necessary new data logs are created and data sets were merged for the analysis purposes. The aim of filtration was reducing complexity and focusing on a specific analysis serving answering specific questions. Two types of filtering techniques were deployed: *Slice and Dice* (for removing events or traces based on values or statistics) and *Variance based filtering* (partitioning the event logs for having relevant and simpler processes)[6].

In the *Process Mining* phase, process mining techniques were applied in iterations to answer the research questions and gain insights to the process. The focus here was process discovery to (a) analyse usage pattern, (b) analyse usage change, and (c) analyse the behaviour of the customers using the portal. On the basis of the process mining outcomes, findings were discussed.

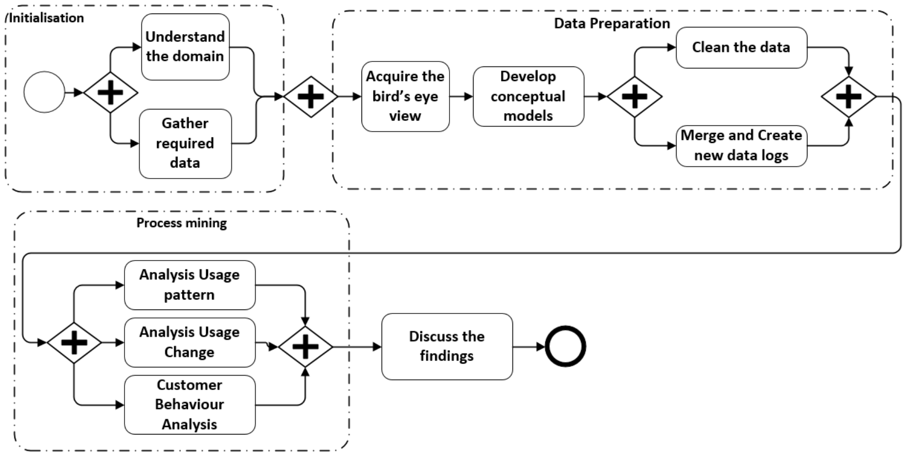


Fig. 1 Workflow of the analysis

Tools used include Minit V.2 (Process mining purposes), PROM [5] (Process mining), Disco (process mining) [6] ,MySQL [7] (data storage and preparation), SPSS IBM (Statistical analysis) [8], and CELONIS (process mining, mainly for OLAP features)[9]. The following sections describe the steps taken in details.

3 Data Preparation

This section provides a high level overview (Section 3.1) and also elaborates on the pre-processing facilitating the data for further analysis (Section 3.2).

3.1 Bird's-eye View

The data provided for the challenge are collected from visitors' clicks of UWV websites www.werk.nl and www. digid.werk.nl over a period of 8 months (from 01/07/2015 to 29/02/2016). The data is focused on customers in the WW (unemployment benefits) process. The following five different log files provided in CSV format are:

1. **"Clicks-not-logged-in"**: contains the clicks data of the customers who are not logged in to the website;
2. **"Clicks-logged-in"**: contains clicks data of the customers who are logged in to the website;
3. **"Questions"**: contains data about the questions asked by customers to UWV through call centres;
4. **"Werkmap Messages"**: contains data about the messages sent by logged-in customers to UWV through digital channels;
5. **"Complaints"**: contains data about the complaints filed by customers.

Each individual customer can be identified with a unique ID while being logged in or when making contacts through the channels. Fig. 2 depicts the overall structure governing the data available with regards to the availability of the knowledge about the identity of the customer. This categorization fosters answering the questions.

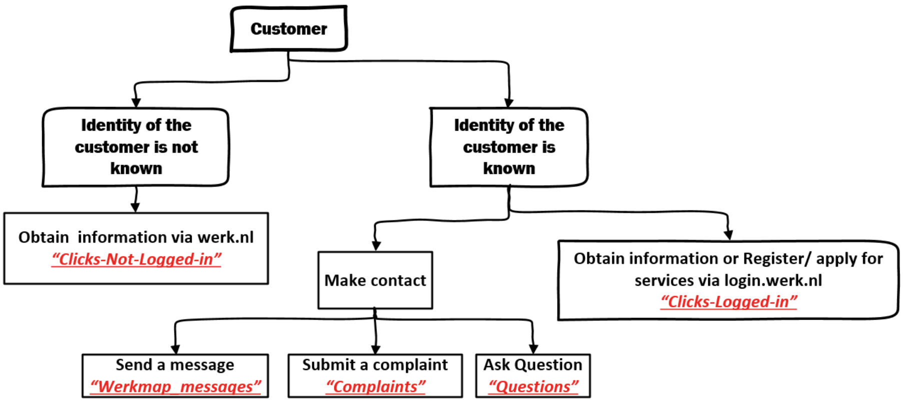

Customer
Identity of the customer is not known
Identity of the customer is known
Obtain information via werk.nl
"Clicks-Not-Logged-in"
Make contact
Obtain information or Register/ apply for services via login.werk.nl
"Clicks-Logged-in"
Send a message
"Werkmap_messages"
Submit a complaint
"Complaints"
Ask Question
"Questions"

Fig. 2 The overall structure governing the data available

On the basis of available data, a conceptual model (Fig. 3) was developed enabling us to realise Case ID as well as filtering and merging Data (Data Preparation) and ultimately serving answering the governing questions (Section 1).

As can be observed in Fig. 3, a *Customer* can be *LoggedIn* or *NotLoggedIn.* The identity of *NotLoggedIn.Customer* is not known. In the case of being *LoggedIn,* the identity of a *Customer* is known as well as the gender and the age category of the customer. A *LoggedIn,Customer* can make several contacts in terms of *Question*, *Message* and *Complaint*. Both *NotLoggedIn.Customer* and *LoggedIn,Customer* can make several *Clicks* related to visiting several *Pages* in several *Sessions*.

As discussed, a unique *ID* realises a *LoggedIn,Customer* for posing a *Question*, sending a *Message* or making *Complaint*. It also realises the identity of a *Customer* while making a *Click* for visiting a *Page* (obtaining information or requesting a service, …). Related to the unique *ID* of a customer (DigiD/BSN), certain attributes can be captured and in the available data sets, *AgeCategory* and *Gender*. On such basis and considering the questions concerning customer behaviours, *ID* of a *LoggedIn,Customer* is selected as Case ID.

Each *Click* has a *TimeStamp* as its attribute. In a *Page* visit, *Name* of the *Page*, its *VHost* and *URLFile* and where applicable its *ServiceDetails* are provided (e.g., “aanvragen-ww”/apply for the unemployment benefit). The granularity of time stamp for page includes date, hours, min and second. Several pages can be visited by a *Customer* in a *Session* that corresponds to a unique *ID* and *IPID*.

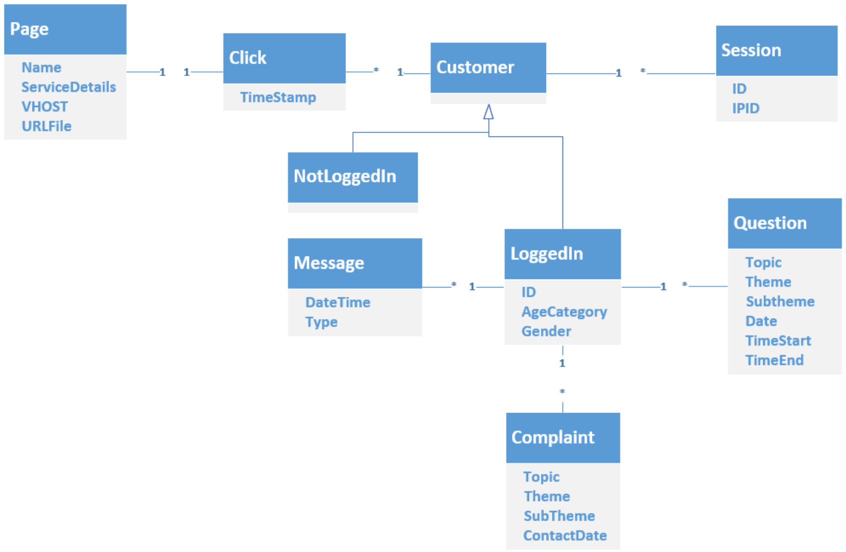


Fig. 3 Conceptual model of the data available

A *NotLoggedIn.Customer* visits www.werk.nl with the purpose of obtaining information from different *Pages* in different *Sessions*. There is an assumption that each *IPID* is used by a single *NotLoggedIn.Customer* (i.e., a *NotLoggedIn.Customer* does not share its device with another *NotLoggedIn.Customer*); therefore *IPID* can be considered as the Case ID for process mining purposes.

As depicted in Fig. 3, each *Customer* can visit several *Pages* in different *Sessions.* We looked at the data and there is one unique *SessionID* for each *Session* for each *IPID* (i.e,, *NotLoggedIn.Customer*). This means that the website does not allocate a particular *SessionID* to two *Customers.*

3.2 Global Statistics

To facilitate data manipulation and querying, we imported the provided CSV files into a MySQL database. **Table 1** shows global statistics on the created tables. The number of click events (#click-events) refers to the number of rows in a table. The number of customers (#customers) refers to the IPID in the “not-logged-in” table and to the customer ID in the other tables.

Table 1 Global statistics on the imported database tables

	Not-logged-in	Logged-in	complaints	questions	Werkmap_messages
#click-events	9,329,418	7,174,934	289	123,403	66,058
#customers	58,833	26,647	226	21,533	16,653

We found out that all of the 58,833 IPID of the “Not-logged-in” appear in the “logged-in” table. Moreover, each logged-in customer used several IPID (i.e., devises) while connecting to the website.

In order to have a global understanding of the data from a process perspective, we collected some statistics on the provided log files using Disco. The selection of the case notion is very important in process mining as it determines the scope of analysis. As discussed before, we focused on the customer “journey” analysis. We considered the IPID as the case notion in the “not-logged-in” log file and the customer ID as the case notion in the “logged-in” loge file. Moreover, we considered the attribute “page_name” which refers to the name of the visited page as the activity attribute (Fig. 2).

The same Customer ID was deployed for identification of a particular customer when logged-in in order to request/apply for some services, obtaining information and/or making a contact. This enabled us to merge the logged-in data and the data related to making a contact (complaint, question and message) in a meaningful way to provide a proper realization of customer behavior. The existing attributes for click data and contact data (Fig. 3) were considered for the process of merging the data. To be able to incorporate the contact data into the logged-in data, each line of data related to a “Question”, “Complaint” or “Message” was considered as an event line. To be able to conduct process mining, a field of “activity” was added to the event lines. Clearly the activities “Question”, “Complaint” or “Message” were added into relevant event lines. More details on the merging process will be explained in Section 6.3.

Table 2 provides general statistics on the “not-logged-in”, “logged-in” and “merged logged-in” log files. The number of variants (#variants) corresponds to the number of distinct cases in the log.

Table 2 General statistics of the log files "not-logged-in" and "logged-in"

	#cases	#activities	#variants	Median throughput time	Start date	End date
Not logged-in	58,833	1,381	47,433	~4 months	01-07-2015	29-02-2016
Logged-in	26,647	600	26,427	~3 months	01-07-2015	29-02-2016
merged logged-in	27,412	603	26,611	~4 month	01-07-2015	29-02-2016

The statistics show that the number of cases in “merged logged-in” log file is bigger than the number of cases in “logged-in” log file. This means that the data about some customers in “Complaint”, “Question” and “Werkmap_messages” do not appear in the “logged-in” data. There exist two possible explanations for this. First, it might be that these customers never used the online services to apply for the unemployment benefit. Second, it might be that these customers used the online services, however their journey dates are out of the range of dates for which data have been collected for this challenge (i.e. not between 01/07/2015 and 29/02/2016). In our analysis, we assumed that these customers did not use the werkmap website and therefore we excluded them from our analysis.

3.3 Data Pre-processing

We were faced with the challenge of dealing with high amount of data (millions of records for each log file). Therefore, it was very important to select a good filtering approach and generate “simpler” log files with which we can start our analysis. Fig. 4 illustrates our filtering strategy. Red branches correspond to the generated sub-logs that will be used for analysis purpose.

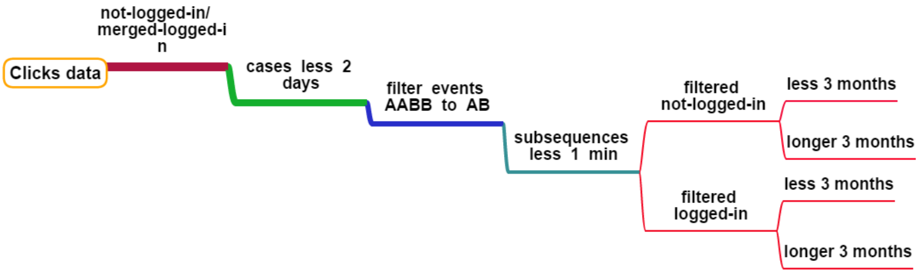


Fig. 4 Filtering strategy

As mentioned earlier, since the data about logged-in customer is spread over four different files, we merged them into one “merged-logged-in” log file. This was done through the *union of MySQL tables.* Before merging and filtering, we solved some inconsistencies in the attributes format as explained in the following subsections. In addition, we excluded the customers who did not use the website.

Dealing with Inconsistent attributes (Merged-logged-in)

It was observed that there are differences in the level of details provided in some types of attributes. We find that the time in the “complaints” log has two issues. First, the granularity of the time is only limited to the date of a complaint while the time attributes for other logs files include lower granularity (Date+time). Therefore, we added the artificial time “00:00:00” to the “complaints”events. Second, the time is defined using the pattern “mm-dd-yyyy’ while for other logs, it is defined using the pattern “yyyy-mm-dd”. Therefore, we preprocessed the data to convert the complaints’ events time to the pattern “yyyy-mm-dd”.

The “questions” log file has a start and end timestamp while all other files have only one complete timestamp. By inspecting the “questions” table, we found that most of the start and end timestamps of one event are equal. Therefore, we simply discarded the start timestamp and tagged the end timestamp as the completed timestamp of an event.

Filtering out the Cases that take less than two days

Looking at the result of process mining in DISCO, we found out that for the not-logged-in 16% and for the merged logged-in data set only 2% of the cases take less than two days. Via using Disco, we filtered out the cases that take less than two days. We filtered out these specific cases as they will not provide the insight required into customers’ journey using the services provided by UWV.

Filtering out subsequent events (AAABB → AB)

We observed that the data contains many subsequent clicks of the same page name. For instance, a visitor may refresh or reload the same page many times. This does not provide any valuable information for analysis and eventually adds to the complexity of the high number of events. Therefore, we decided to filter these events using an available plugin in ProM.This resulted in having 51% of the events remained in the new logged-in file and 17% of the events remained in the new logged-out file. Note that the percentage was calculated incrementally regarding the result of the earlier filtration as the basis.

Filtering out the events separated by a very small time (less than one minute)

As we were interested in knowing the customer behavior through click data, the way the customer conducted the clicking was taken into account. It is possible that a customer had a targeted goal "Applying for a benefit" and therefore made clicks on several pages and looked for the right page till he/she actually conducted the application. Thus, the time one spent on certain pages could indicate whether that page was of actual interest or whether it was just a path serving another goal.

On such basis, we filtered out the pages that a customer took less than 60 seconds on it by using PROM (a new plug-in is developed specifically serving such functionality). This resulted in having 40% of the original events remained in the new logged-in file and 46% of the original events remained in the new logged-out file.

Please note that this filter and the next one (i.e. filtering out subsequent events) were only applied for the activities referring to website pages. The activities "question", "message" and "complaint" referring to contact channels were excluded from the filtering.

Less vs Longer than three months

As mentioned in the introduction, customers receiving the unemployment benefit receive only e-services during the first three months of their unemployment as job seekers. This was a good reason to filter and make a comparison between the behaviour of the customers in the cases that took less than three months (did not used the services/obtain information after 3 months) and in the cases that took more than three months. Such filtering is conducted in DISCO.

As a result, it was observed that in the Not-logged data log 32% of cases took less than three months and 68% of the cases took longer than three months. For the merged logged-in, 36% took less than 3 months, and 64% lasts longer than 3 months.

Handling noise in "Not-logged-in" data log

By looking into the "not-logged-in" log file, we found out that the two virtual hosts www.werk.nl and www.digid.werk.nl were visited by not logged-in customers. Since www.digid.werk.nl contains basically features for logged-in customers, we inspected the data more in details to search for possible noise, i.e. click data related to logged-in customers. From a first high level inspection, we discovered that the activity "aanvragen-ww (~17%)" was indeed in the top level of the most frequent activities in the dataset and is followed by the activity "home (~16%)" in the second level. We tried to visit the page "aanvragen-ww" by following its URL path recorded in "url_file" attribute. Interestingly, we were redirected to the "login" page which means that the "aanvragen-ww" page cannot be visited without being logged-in. This confirms our hypothesis for the presence of noise, i.e. the date about logged-in customers are logged in the "not-logged-on" file.

In order to find and filter all possible noise coming from “logged-in” data clicks, we decided to filter out the entire session whenever we find it, associated with the same IP address, in the “logged-in” table.

4 Usage Patterns analysis

This section provides an answer to the question of “*Are there clear distinct usage patterns of the website to be recognized?* In this regard, insights into the way various customer demographics used the websites and web-services are provided. Section 4.1 is dedicated to investigating the customers that used the website less than three months vs the customers that visited the website longer than 3 months. Section 4.2 provides demographic oriented statistics.

4.1 Less vs Longer than three months

We extracted a diagram on distribution of the active cases that took less than 3 months over the period of six months in DISCO (Fig. 5).

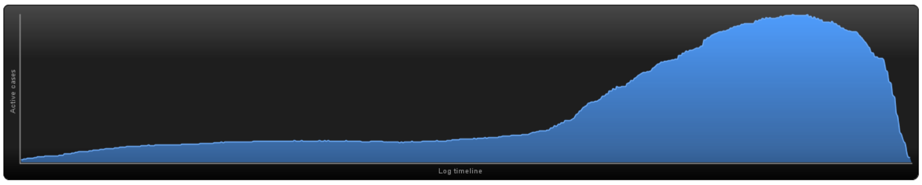


Fig. 5 Active cases that took less than 3 months

As can be observed, almost at start of November the number of active cases increased. Basically, by start of winter more customers were prone to use unemployment services or contact UWV for acquiring services. This might be related to the seasonally of available jobs in the Netherlands (e.g., tourism, agriculture, etc.). A suggestion for UWV can be to consider such workload in terms of number and allocation of resources as well as maintaining a stable website and e-services provided in the desired level.

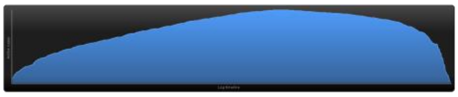


Fig. 6 Active cases that took more than 3 months

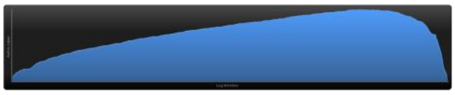


Fig. 7 Active cases during the 8 months period

As depicted in Fig. 6, there is a slight increase in number of active cases that took more than 3 months when we reach the end of the year (not as noticeable as cases that took less than 3 months). Fig. 7 shows that for all the cases no matter what were their durations, by end of the year there was a pick in number of active cases. Taking into account the behaviour of job market and the new decisions made by the end of the year, there can be some correlations in this regard.

4.2 Demographic-oriented statistics

Fig. 8 depicts the number of logged-in customers per gender and per age category. The provided Statistics were driven using SPSS V2.3. It is clear that the demographics of customers logging-in to the website and probably applying for an unemployment benefit are roughly equally distributed. One slight noticeable difference is between customers in the 50-65 age category. The number of male customers is more than the number of female in this category. With the assumption that there is no difference between percentage of the female and male customers in making use of e-services provide there can be two possible explanations for such difference: Either male were more prone to unemployment in such age category or in general male were more in job market in that age category.

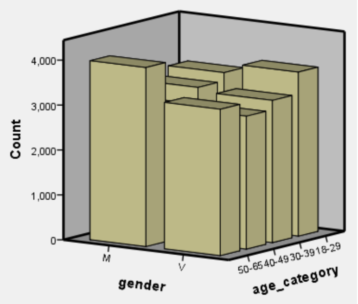


Fig. 8 Customer demographics statistics in "Logged-in" table

Driven by statistics provided by European Commission[3], the employment rates between age of 25 and 54 is higher than the age category of 15-24 and age 55-64. The rage of employment between age categories of 15-24 is almost the same as age 55-64. Given the facts that most of young people are busy with studying between age of 15-24, this explains that people older than 55 are more prone to be out of job in comparison to other age categories. For Netherlands for all age categories, females are less employed in comparison to males (%68 vs. %78). Aggregating all these facts and the observation justify the hypostasis that male customers of UWV in the age group of 50-64 were more in job market and therefore using more such services due to the fact that this category is prone to being unemployed.

[3] http://ec.europa.eu/eurostat/statistics-explained/index.php/Employment_statistics

5 Usage Change Analysis

This section provides and answer to the following questions: *Do the usage patterns of the website by customers change over time*? *Do customers visit different pages when they start using the website versus when they have been using the website for some time*? *How does the usage change over time*?

We split the “filtered-logged-in” log into four sublogs by trimming the cases on a basis of two months. The four sublogs have been compared and matched against each other to study how the behavior of customers change over time.

Fig. 9 shows the top 10 frequent activities in the different logs. It is clear that there is a change in the frequency level of visited pages over time. For example, the page “mijn-cv” is the second most visited page in the log of the first two months and becomes the seventh most frequent visited page in the log of the fourth two months. This indicates that customers applying for an unemployment benefit start searching for a job and subsequently uploading their CVs in the earliest steps of their journey. Therefore, it is recommended to UWV to ensure the availability of needed resources and services for helping customers searching and finding a job as the journey of a customer starts. The activity “question” appears as a frequent activity in all the logs while the activity “message” only starts to appear in the third and fourth two months. It seems that customers are using the call center channel excessively throughout their journey. On the other hand, the werkmap message channel starts to be helpful as customers have been visiting the website for a while and have been engaged in the unemployment process.

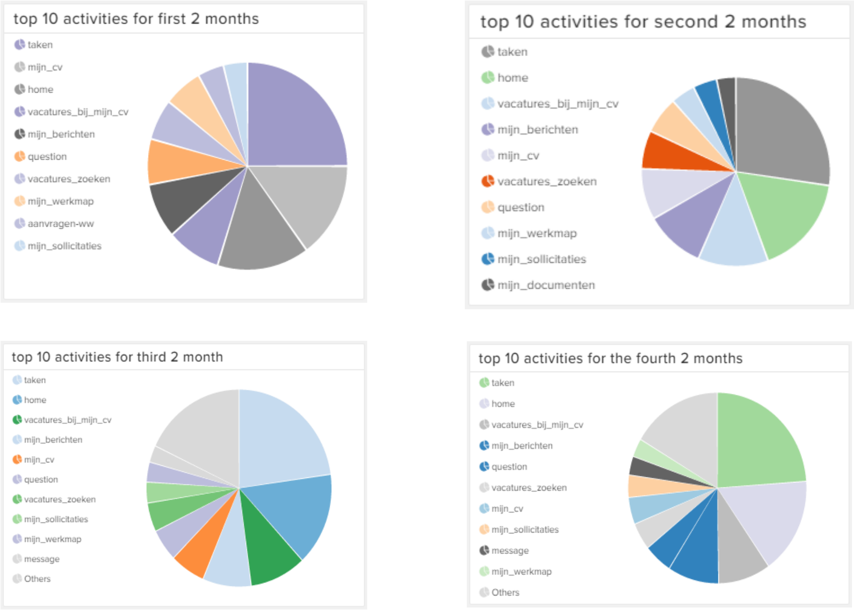


Fig. 9 Top 10 frequent activities of the 4 sub-logs over the whole period

We also compared and matched the process maps generated by Minit against each other using the process variant comparison feature. **Fig. 10** to **Fig. 12** show the process maps of 1) first two months vs second two months, 2) second two months vs third two months and 3) third two months vs fourth two months respectively. The activities (and edges) colored in orange are those that are present in both maps and therefore are matched. The activities (and edges) colored in blue are those that belong to the first variant. The activities (and edges) colored in green are those that belong to the second variant. The maps are configured to show 10% of the most frequent activities and 5% of the most frequent edges.

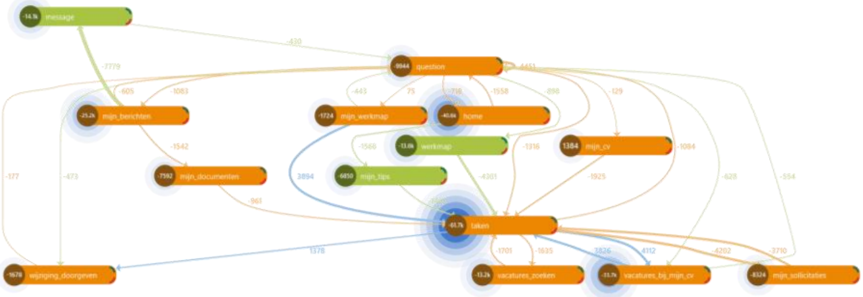


Fig. 10 Sub-log of the 1st two months period vs. sub-log of the 2nd two months period

The first figure shows that the most frequent pages visited by the customers during the first two months are continued to be visited during the second two months. However, in the second two months the activities “message”, “werkmap” and “mijn-tips” start to appear in the most frequent ones. The explanation can be made with regards to the legal obligations that they must fulfil within a certain timeline. It is also observed that when a customer became engaged, he/she can start using the message services.

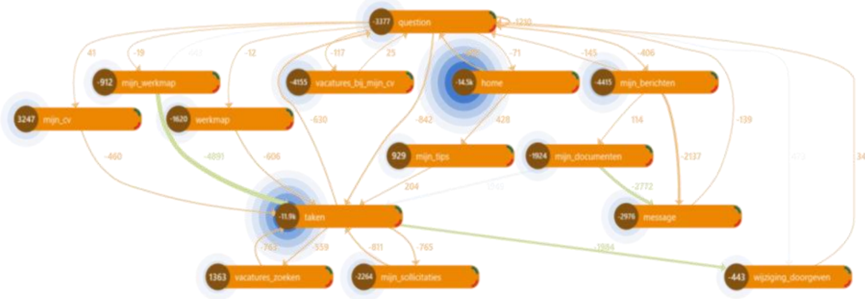


Fig. 11 Sub-log of the 2nd two months period vs. sub-log of the 3rd two months period

In the last period, an observation is made toward the use of “Vragenlijst-uwv” than can indicates the obligation toward filling a questionnaire.

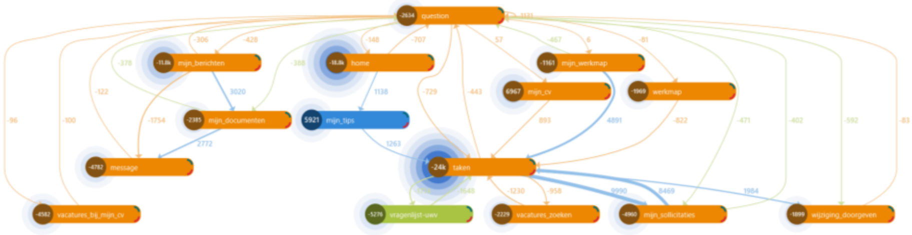


Fig. 12 Sub-log of the 3rd two months period vs. sub-log of the 4th two months period

6 Expensive Usage Analysis

In this section, an analysis on the behaviour of customers who contacted UWV through their channels is provided. We refer to this behaviour as "expensive". To do so, we used the "filtered-logged-in" log file. In Section 6.1, we provide some global statistics on the use of channels by demographics and the top contacted channels by customers. In Section 6.2, an analysis is provided on when and why a transition from cheap to expensive channel occurs. This provides and answer to the questions of "*When is there a transition from the website to a more expensive channel, such as sending a Werkmap message, contacting the call centre or filing a complaint?* " and "*Is there a way to predict and possibly prevent these transitions?* "

In Section 6.3, an analysis is offered on the change in the behaviour of customers after their first contact with UWV. This provides an answer to the questions of "*Does the behaviour of the customers change after they have send a Werkmap message, made a phone call or filed a complaint?", " Are customers more likely to use these channels again after they have used them for the first time?","What is the customer behaviour on the site after customers have been in contact through the Werkmap or by phone?"*

In Section 6.4, we infer the behaviour of customers that directly lead to complaints. This provides an answer to the question of "*Is there any specific customer behaviour that directly leads to complaints?"*

6.1 Global Contact channel usage statistics

In order to have a global understanding of the expensive channels' usage, we collected some statistics related to their usage by customers' demographics. Moreover, an statistic is provided on the top 10 asked questions and filed complaints accompanied by some recommendations

Fig. 13 shows the number of logged-in customers who contacted UWV through all available channels (i.e., via the call center to ask a question, or via a digital channel for sending a werkmap message or making a complaint). The following abbreviations were used: "C" for customers who made a Complaint, "Q" for customers who posed a question and “M” for customers who sent a werkmap message. The “!” sign before a letter (e.g. “Q,M,!C”) means that the customer did not use the corresponding channel. Each class shows the statistics based on customers’ demographics (i.,e, *Gender* and *Age Catergory).*

The first inspection of the results allowed to conclude that most of the customers use simultaneously the call center (i.e. to ask questions) and the digital channel (i.e. to send werkmap messages). Few of them used the three channels together and none of the customers used only one channel. This raises the first question whether the digital channel is optimized or not (e.g., usability, usefulness, etc.) and what makes customers to prefer pose a question via the call center or see the urge to use both channels to get a help/result. We try to analyze and answer this question in the next sections.

The diagram also shows that the use of both channels to ask questions and send messages have a correlation with the age category. Old customers in the 50-65 age category use both channels more than the younger customers in the 18-29 age category. The diagram also shows that the contact channels are more or less equally used by customers of different genders. Finally, we can observe that the number of filed complaints is very negligible. This means that the majority of the customers were satisfied with the UWV unemployment benefit services.

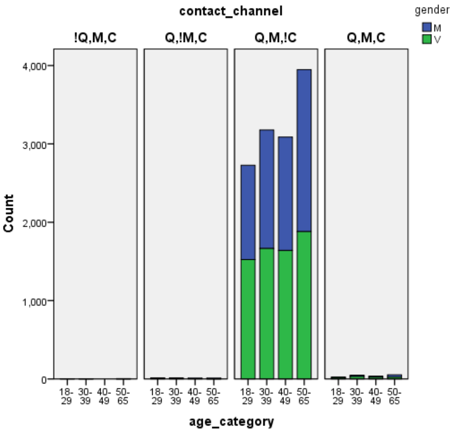


Fig. 13 Statistics of using different contact channels based on customers' demographics

To provide more insights on the distribution of the expensive channel usage in time, we filtered the traces on the events containing one of the three contact channels as activity. Then, the traces’ events were plotted against time using ProM (Fig. 14).

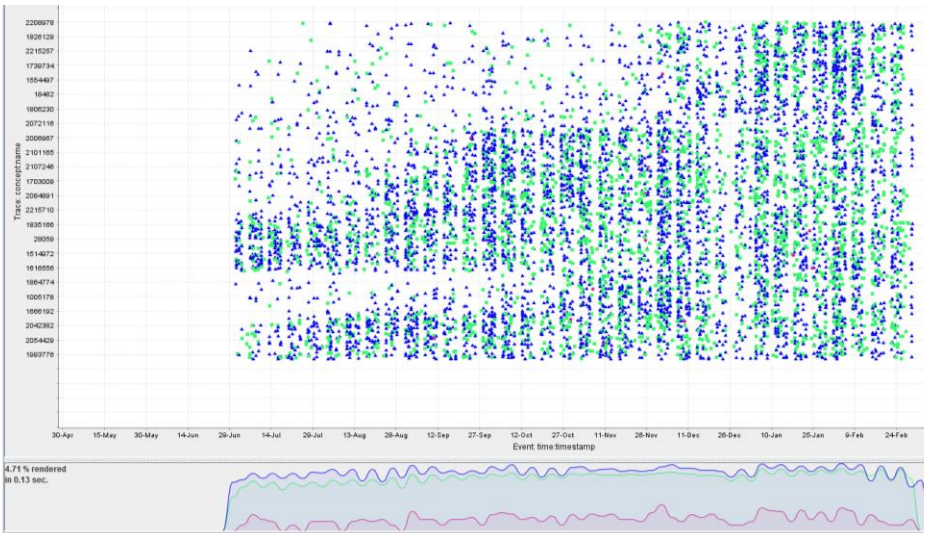
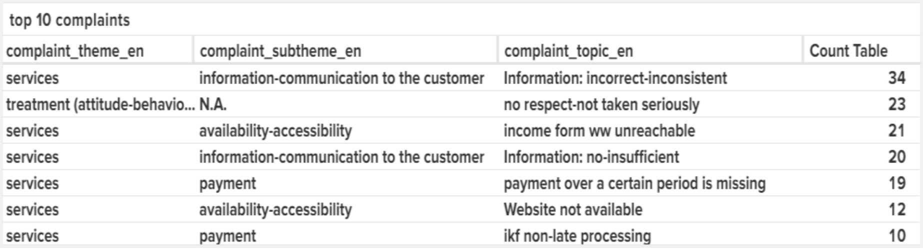


Fig. 14 Dotted chart visualization of the events related to complaints (in pink), messages (in green) and question (in blue)

It is clear from Fig. 14 that the contact channels “message” and “question” were simultaneously used by customers (this is already shown in Fig. 13). Interestingly, the customers started using these channels in the very early steps of their journey. The chart also shows that the expensive usage was very frequent and in most of the cases was separated with very low delays. To follow up on the previous analysis, we computed the most frequent questions asked by customers as well as the most filed complaints. Regarding the message channel, the data do not contain any further information about the content.

Fig. 15 and Fig. 16 show the top 10 filed complaints and asked questions by theme, subtheme and topic. The results are computed using OLAP tables in Celonis. As shown in Fig. 15, most of the complaints (54 counts) are related to the communication: incorrect-inconsistent information. This indicates that there is room for improvement in terms of providing information and the integrity of the information provided in different sources. The next category of complaints is related to the website and its accessibility (33). Given the fact that this was also observed during the peak load of cases (Fig. 5), UWV can work on availability and the uptime of its website and e-services to be able to handle a huge amount of cases it encounters. Complaints about payment (late or missing) count 29 complaints. Thus, improvement in payment procedures and its reliability can be taken into account. Finally, 23 complaints are received with regards to the way they are treated which shows there are rooms for improvement in this category.



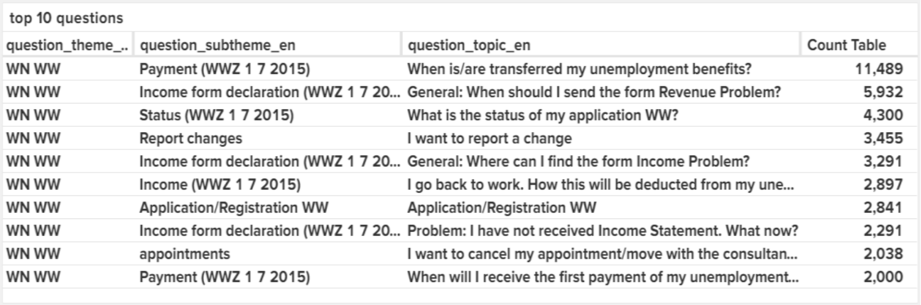
top 10 complaints

complaint_theme_en	complaint_subtheme_en	complaint_topic_en	Count Table
services	information-communication to the customer	Information: incorrect-inconsistent	34
treatment (attitude-behavio...	N.A.	no respect-not taken seriously	23
services	availability-accessibility	income form ww unreachable	21
services	information-communication to the customer	Information: no-insufficient	20
services	payment	payment over a certain period is missing	19
services	availability-accessibility	Website not available	12
services	payment	ikf non-late processing	10

Fig. 15 Top 10 complaints by theme/subtheme/topic

As depicted in Fig. 16, majority of the questions falls in the category of "Payment". This indicates that there is a room for improving the communication methods in this regards (e.g., via a SMS service) as the questions concerns the time of the payment. The next category of questions concerns the "income form declaration". This again concerns the communication (i.e., clarity of procedure) in this category. The way communication is conducted perhaps can improve (e.g., via providing explanatory clips).

Improvements can be made in terms of having the status of an application updated and reliable on the website while clarifying the expected timeline and procedure. The amount of questions with regards to "application/registration" again justify the need for looking for another methods of communicating the procedures.



top 10 questions

question_theme_..	question_subtheme_en	question_topic_en	Count Table
WN WW	Payment (WWZ 1 7 2015)	When is/are transferred my unemployment benefits?	11,489
WN WW	Income form declaration (WWZ 1 7 20...	General: When should I send the form Revenue Problem?	5,932
WN WW	Status (WWZ 1 7 2015)	What is the status of my application WW?	4,300
WN WW	Report changes	I want to report a change	3,455
WN WW	Income form declaration (WWZ 1 7 20...	General: Where can I find the form Income Problem?	3,291
WN WW	Income (WWZ 1 7 2015)	I go back to work. How this will be deducted from my une...	2,897
WN WW	Application/Registration WW	Application/Registration WW	2,841
WN WW	Income form declaration (WWZ 1 7 20...	Problem: I have not received Income Statement. What now?	2,291
WN WW	appointments	I want to cancel my appointment/move with the consultan...	2,038
WN WW	Payment (WWZ 1 7 2015)	When will I receive the first payment of my unemployment...	2,000

Fig. 16 Top 10 questions by theme/subtheme/topic

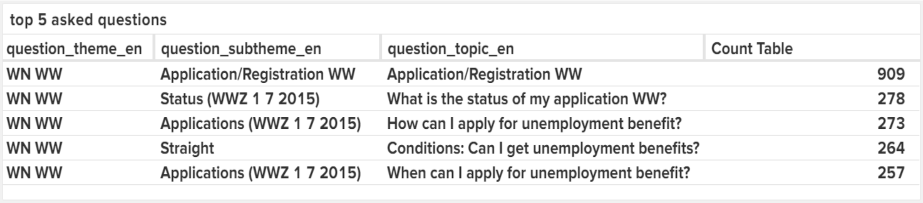
Considering this fact that this channel is the most expensive channel of communication, such considerable amount of questions justifies improvement plans with regards to communication, payment, etc.

6.2 Transition change analysis

In order to analyze the first transition of customers to an expensive channel, we trimmed the log cases to the first occurrence of one of the activities related to a contact channel (i.e. question, message and complaint) using a Disco filter. We found out that the median time of cases is **11.7 days** which indicates that the transition occurs in a very short time after the customer starts using the website (this is also related to the dotted chart in Fig. 14. Moreover, we discovered that the average number of events per case is **19.8** while the average number of events per case for the complete cases is **87.67.** This indicates that the customers go to an expensive channel after very few clicks.

By inspecting the different variants in Disco, we realized that the most frequent variant (**24.79% of cases**) contains one event having the activity "question", **144 cases** contain only the activity "message" and **8 cases** contain only the activity "complaint"[4]. This means that these customers started their journey by contacting one of the three available channels. In order to have a deeper understanding of this behavior, we computed the statistics related to the asked questions (Fig. 17) and filed complaints (Fig. 18) using OLAP tables in Celonis. Statistics related to the werkmap messages are not available since there do not exist any information about the content.

As can be seen in Fig. 17, the most frequent questions regard application (procedure and timeline). This again highlights the need for improving the way the application/registration process is communicated. The status of the applications were also concerns of the customers. Perhaps another ways of communicating the status combined with clear timeline on the expected timeline could be beneficial in this regards. The questions on the sub-scheme of straight can also be reduced while the procedures, rules and regulations are communicated in a more effective way.

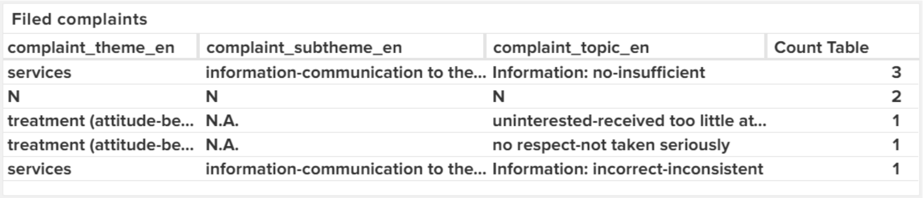


top 5 asked questions

question_theme_en	question_subtheme_en	question_topic_en	Count Table
WN WW	Application/Registration WW	Application/Registration WW	909
WN WW	Status (WWZ 1 7 2015)	What is the status of my application WW?	278
WN WW	Applications (WWZ 1 7 2015)	How can I apply for unemployment benefit?	273
WN WW	Straight	Conditions: Can I get unemployment benefits?	264
WN WW	Applications (WWZ 1 7 2015)	When can I apply for unemployment benefit?	257

Fig. 17 Top 5 questions that start the journey of customers

Fig. 17 elaborates on the counts of complaints in the start of the journey by the customer. While the numbers are negligible, the figure also indicates the need for improving communication.



Filed complaints

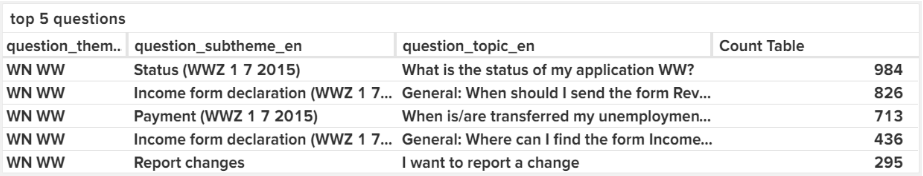
complaint_theme_en	complaint_subtheme_en	complaint_topic_en	Count Table
services	information-communication to the...	Information: no-insufficient	3
N	N	N	2
treatment (attitude-be...	N.A.	uninterested-received too little at...	1
treatment (attitude-be...	N.A.	no respect-not taken seriously	1
services	information-communication to the...	Information: incorrect-inconsistent	1

Fig. 18 Complaints that start the journey of customers

The OLAP tables of the remaining cases (i.e. cases that contain more the one activity) are separately generated (Fig. 19 and Fig. 20). Fig. 19 depicts the first

[4] It is worth noting that, because of incompatibility in the time granularity of the complaints events and the insertion of artificial times, the statistics related to complaints may not be reliable.

questions posed by the customers after using the website for a while. As can be seen, this again indicates the need for improving the communication of the matters to customers such as updating on the status of the application, the timeline, the procedures and regulations.

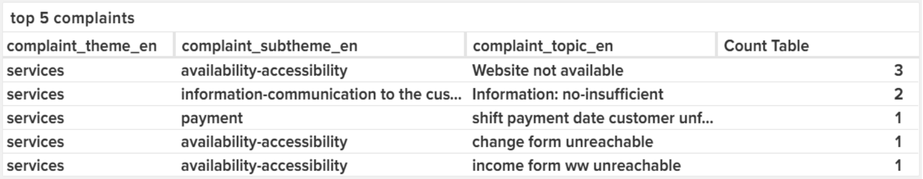


top 5 questions

question_them..	question_subtheme_en	question_topic_en	Count Table
WN WW	Status (WWZ 1 7 2015)	What is the status of my application WW?	984
WN WW	Income form declaration (WWZ 1 7...	General: When should I send the form Rev...	826
WN WW	Payment (WWZ 1 7 2015)	When is/are transferred my unemploymen...	713
WN WW	Income form declaration (WWZ 1 7...	General: Where can I find the form Income...	436
WN WW	Report changes	I want to report a change	295

Fig. 19 First questions asked by customers after using the website for a while

Fig. 20 indicates the first complaints filed by the customers after using the website for a while, while the numbers are negligible, the need for improving the availability of the website and the e-services provided is also highlighted here.



top 5 complaints

complaint_theme_en	complaint_subtheme_en	complaint_topic_en	Count Table
services	availability-accessibility	Website not available	3
services	information-communication to the cus...	Information: no-insufficient	2
services	payment	shift payment date customer unf...	1
services	availability-accessibility	change form unreachable	1
services	availability-accessibility	income form ww unreachable	1

Fig. 20 First complaints filed by customers after using the website for a while

In order to predict the behaviour that leads to a transition to a more expensive channel, we developed a business process model (Fig. 21) by DISCO leading to "complaint", "message" and "question" transitions (25% paths and 100% of activities).

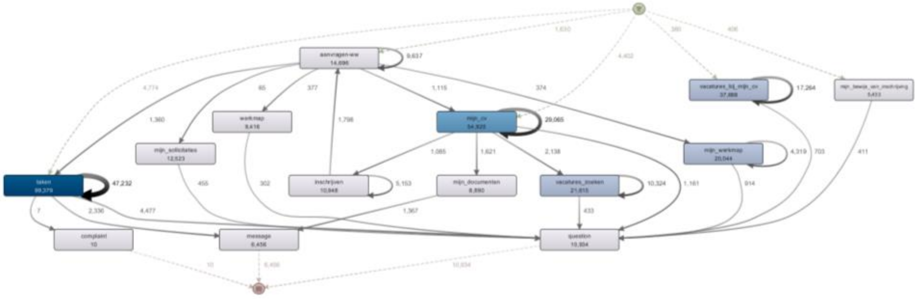


Fig. 21 Activities lead to "Complaints", "Questions", and "Message" transitions

As can be seen, there are distinct paths leading to these transitions. For example, path "*request the benefit*, *my CV* and *searching vacancies* leads to a question. In another example, (*request for a benefit* and) *tasks*, lead to a complaint (total of 7 counts). One example of the paths leads to a message is *request for a benefit*, *my CV* and *my documents.* Majority of paths led to posing a question which is an expensive channel of communication. This shows that there might be several matters which are not clear or are not communicated in an effective way with the customers. This also has been demonstrated earlier. Therefore, there is a need for communicating the processes, procedures, status, etc. in a more clear and effective way through the website.

6.3 Expensive usage change behavior

The behavior of customers before and after the first contact was compared using the process variant comparison feature in MINIT. To do so, two variants of event logs were created: 1) the event log generated from the previous step which is labeled as "expensive-usage-before-transition" and 2) the event log representing the complement of the first one and is labeled as "expensive-usage-after-transition". This log was generated by trimming the traces after the first occurrence of a contact channel activity.

Fig. 22 shows the process maps generated by MINIT that are matched and compared. The maps were configured to show 30% of most frequent activities and 5% of most frequent edges. The orange color means that the activities are matched, the green color means that the activities appear in the second variant, i.e. the variant corresponding to the behavior after transition.

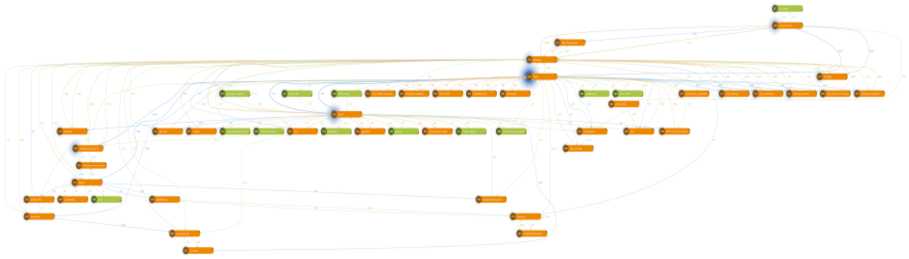


Fig. 22 Difference in page visits by customers (green) after making a contact

As the figure is not visible, the major differences in pages visited occurred after making a contact are as following: online job application, job application, job search, online training, support request and instruction.

Taking the second log referring to the customers' behavior after their first contact with UWV, we found that in average a customer contacts one or more of the UWV contact channels **7.7 times**. One outlier customer having the identifier **2027753** has the maximal number of contacts (**107 contacts**) after his first contact and spent a journey of **7.7 months**.

7 Conclusion

In the context of BPI Challenge 2016, this paper conducted an analysis on the click data provided by the Dutch Employee Insurance company UWV. The organisation is responsible for providing employee insurances and data services for its customer. By applying process mining techniques the aim of the analysis is improving efficiency and effectiveness of the services provided via discovering customer journey and the patterns associated. A set of questions are provided by UWV in this regard.

Five data set are provided in terms of click data for not-logged in customers and not logged-in customers as well as the contacts made where analysed. A methodology was provided elaborating steps taken to conduct the analysis. The methodology consist in three major phases of "initialisation"," Data Preparation" and "Process Mining" followed by discussion of the findings. On the Bird's-eye-view a general overview of the data was provided elaborating on the overall structure governing the data. A set of generic statistics provides insights on the number of cases, activities, throughput time, etc.

To provide answers to the question, there was a need to pre-process the data. A strategy is provided in this regards. This was operationalised through merging, filtering out the data on the basis of the length of the cases/events and clustering the data taking into account different patterns and where applicable rules of UWV. Concerning the merge, there was a need to handle inconstant attribute in the preparation process.

The analysis of the result is categorized into three parts: Usage patterns analysis, Usage change analysis, and Expensive Usage analysis. In the usage pattern analysis, length of the cases are considered into account. As known, in the first three months that a customer is receiving a service only e-services are provided. Therefore, it was necessary to investigate if there is a change in the behaviour of the customers (i.e., cases) if the length of a case is less or more than 3 months. Moreover, a demographic oriented statistics are provided taking into account different age categories and gender of the customers accompanied by some interpretations.

The usage change analysis indicates the change of the behaviour of the customer during their journey. This can be correlated with the rules, regulations as well as obligations and duties of the customers. In the expensive usage analysis, the behaviour of the customers contacted UWV is investigated. Different clusters of contacts is developed and behaviour of the customers considering their age and their gender is taken into account. While considering evolving the behaviour of the customer during its usage (e.g., beginning of the usage), statistics on the basis of different sub-themes of contacts are provided. On such basis, recommendation of improvement is drawn. Moreover, the behaviours (paths) leading to a transition (complaint, message or a questions) are also discovered. Finally, the change in customer behaviour after the first contact is analysed.

Based on the result, following conclusions can be driven:

1. A seasonality element was observed that affect that number of active cases and therefore affects the workload of UWV and the website visits
2. Most of the customers used at least two of the channels to make contact with UWV, (c) older customers tended to use two channels in comparison to the younger customers using just one channel of communication
3. The number of complaints are negligible in comparison to messages and more importantly in comparison to questions (the highest percentage)

4. Contact channels of “message” and “question” were mostly used simultaneity and at the very early stage of customer journey
5. The expensive channel were used very frequently and in a very short time intervals
6. The top three complaints were related to (a) communications and incorrect and inconsistent information, (b) availability of the website and e-services and (c) payment.
7. The top three questions were related to (a) payment, (b) income form declaration, and (c) status of application and procedures.
8. Transitions occurred in a very short time after the first visit to the website. More importantly, the customers uses expensive channels after very few clicks. Marjory of these transitions includes with a question.
9. On average a customer contact different channels of UWV, 7.7 times during its journey.
10. Home” was the most prominent activity was followed by all other activities without any intermediate activities in between.

On the basis of the above conclusions, following recommendations can be made:

1. Availability and uptime of the website and e-services: Given the fact that this was also observed during the peak load of cases (**Fig. 5**), UWV can work on availability and the uptime of its website and e-services to be able to handle a huge amount of cases it encounters.
2. Payment procedure: considering the number of complaints, improvement in payment procedures and its reliability can be taken into account.
3. Communication: Improvements can be made in terms of effective and clear communication of
    a. Status of an application updated and reliable on the website while clarifying the expected timeline and procedure.
    b. Procedures, rules, timelines and regulations
4. User-friendliness of “home”: As a most frequent and intermediate activity, user-friendliness of the design can be specifically considered for this page.

For future work, more data from different sources could be included. The analysis can become more complete if data from the other sources rather than click-data can be also included such as the payment information (e.g., when the payment is made and when it is received) or the information with regards to communicating the status of an application to a customer. In such a way, causality of a certain activity or the path through it could be figured in a more accurate way.

8 References

1. Gathier, M.: Welkom in Nederland, Kennis van de Nederlandse Maatschappij voor het inburgeringsexamen (2016)
2. van der Aalst, W.M.P.: Process Mining: Discovering, Conformance and Enhancement of Biasness Processes. Springer (2011)
3. van Eck, M.L., Lu, X., Leemans, S.J. , an der Aalst, W.M.: PM^ 2: A Process Mining Project Methodology. In: International Conference on Advanced Information Systems Engineering, pp. 297-313. Springer International Publishing, (Year)
4. Brodie, M.L., Mylopoulos, J., Schmidt, J.W.: On conceptual modelling: Perspectives from artificial intelligence, databases, and programming languages. Springer Science & Business Media (2012)
5. Verbeek, H.M.W., Buijs, J.C.A.M., van Dongen, B.F., van der Aalst, W.M.P.: Prom 6: The process mining toolkit. In: BPM Demonstration Track, pp. 34-39. 615, (Year)
6. Günther, C.W., Rozinat, A.: Disco: Discover Your Processes. BPM (Demos) 940, 40-44 (2012)
7. Greenspan, J., Bulger, B.: MySQL/PHP database applications. John Wiley & Sons, Inc. (2001)
8. Field, A.: Discovering statistics using IBM SPSS statistics. Sage (2013)
9. van der Aalst, W.M.: Extracting event data from databases to unleash process mining. BPM-Driving innovation in a digital world, pp. 105-128. Springer (2015)