Does werk.nl work?

BPI Challenge 2016

Gert Janssenswillen[1,2], Mathijs Creemers[1], Toon Jouck[1], Niels Martin[1], Marijke Swennen[1]

[1] Hasselt University, Agoralaan Bldg D, 3590 Diepenbeek, Belgium
[2] Research Foundation Flanders (FWO), Egmontstraat 5, 1060 Brussels, Belgium

Abstract This paper examines the interactions between clients of the Dutch Employee Insurance Agency UWV and their website werk.nl. The interactions taken into account are visits to the website, messages sent, questions asked and complaints filed. The analysis includes a characterization of clients based on their demographic aspects as well as their behaviour. The study further creates an understanding of how people use the website and points out why clients move to more expensive channels, such as messages and questions, or why clients file a complaint. The insights suggest possible ways to improve the website and thereby avoid clients switching to other channels, eventually improving both the client-experience and the cost-effectiveness of UWV's operations.

1 Introduction

This paper focuses on retrieving event log insights on the actions performed by clients of the Dutch Employee Insurance Agency UWV on their website werk.nl. The aim of the paper is to provide insigths and suggest possible ways to improve the website and thereby avoid clients switching to other channels, eventually improving both the client-experience and the cost-effectiveness of UWV's operations.

More specifically, the key contributions of this paper are threefold. Firstly, an exploratory analysis is performed to get an insight into the demographics and behaviour of the customers on the website. This behaviour can be clicks, sending and reading messages, asking questions or filing complaints. Based on this exploratory analysis, secondly, a thorough data preparation is performed. Thirdly, the data is analysed in view of the proposed research questions. Therefore, usage patterns of the customers are investigated with log-based process metrics and clustering algorithms. In addition to this, changes over time are considered by looking at flows between types of sessions and between other events.

All analysis results outlined in this paper are generated using three distinct tools, which are R, ProM and Disco. From a methodological perspective, this research has been conducted following the principles of the PM^2 methodology [3].

The paper is structured as follows. Section 2 outlines how the followed research methodology is applied to the UWV case and gives an overview of the tools used.

Next to this, the steps taken to explore and prepare the data are outlined in Section 4 and Section 5, respectively. The analyses that have been performed to answer the research questions of the UWV case are outlined in Sections 6 - 7. Finally, Section 8 concludes the paper.

2 Methodology

2.1 Analytical methodology

The results presented in this report are based on a process mining analysis, backed by a clear methodology. Three key methodological frameworks are proposed to support the execution of a process mining project: (i) the process diagnostic method [2], (ii) the L* life-cycle model [1] and (iii) the PM^2 methodology [3]. While the process diagnostic method aims to quickly retrieve event log insights in the absence of domain knowledge [2], the L* life-cycle model proposes a more profound step-wise approach to discover a control-flow model and extend it with insights from other process mining perspectives [1]. The PM^2 methodology stresses the iterative character of the process data analysis and states that both process models and analytical models can be generated from event data [3]. Given these distinguishing characteristics, the PM^2 methodology is selected to guide the project.

As visualized in Figure 1, the PM^2 methodology is composed of six stages. For a detailed overview of each of these phases, the reader is referred to [3]. The remainder of this subsection focuses on briefly outlining how the methodology has been applied to the UWV case. The first stage, the planning stage, involves delimiting the business process that needs to be analyzed and identifying the research questions at hand. The provided case study covers these activities as it indicates that the focus is on customers' journeys and through the specification of the questions summarized in Section 3. The same holds for the extraction stage, the second step, as data files are already provided. Domain knowledge, which should also be gathered in the extraction phase, could be obtained by means of the ProM forum. After the first two stages, multiple analysis iterations were conducted in which data is transformed to a format suitable for analysis, analyses are conducted and the obtained results are interpreted. In the first iteration, data processing efforts in the third methodology phase are limited as the envisioned analyses have an exploratory nature. These analyses, part of the fourth phase of the PM^2 methodology, are reported in Section 4. They aim to support event log creation in the second analysis iteration and limit the burden on process experts in the form of forum questions. In the second analysis iteration, two event logs with different levels of aggregation are created as outlined in Section 5. These are used as an input for the mining and analysis phase, generating results which are interpreted consistent with the diagnose activity in the evaluation phase, the fifth stage of the PM^2 methodology. The obtained analysis results are outlined in Section 6 and 7. Note that the verification and validation activity in the fifth stage has to be conducted in close collaboration with the process owner and is, hence, beyond the scope of this report. The same holds for the sixth stage of the methodology as it involves the implementation of improvement ideas.

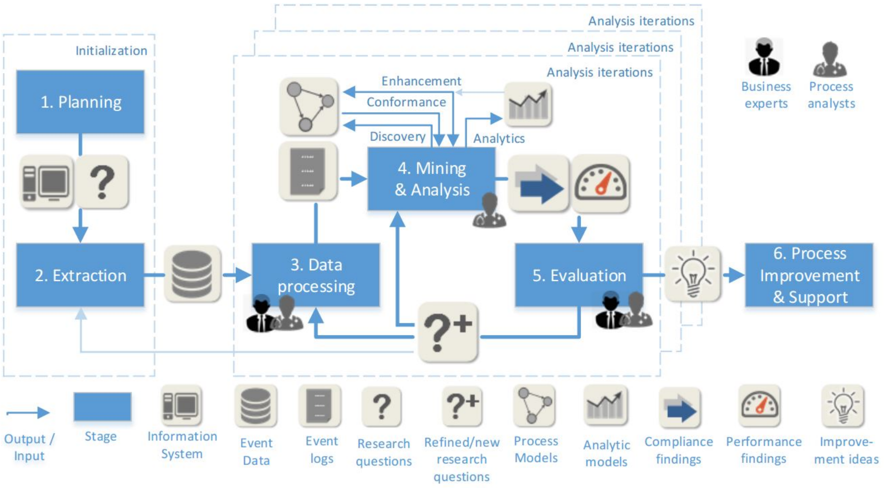

Figure 1: Overview of PM²-methodology.

2.2 Tool support

The analysis results outlined in the remainder of this paper are generated using three distinct tools: R, ProM and Disco. This subsection briefly introduces each of them.

ProM [1] is an open-source process mining platform[1], containing a series of implementations of process mining algorithms developed in academia. The platform is composed of a set of plugins, which provide the implementation of a particular algorithm. In general, ProM has a strong focus on retrieving control-flow models from an event log and, e.g., checking the degree of conformance with respect to a reference model. Although ProM was used for this analysis, it was found to be not completely suitable for the job because of several reasons. Firstly, the tool had severe problems with the amount of data that was available, thereby slowing down the data analysis significantly. Secondly, it is believed that a lot of flexibility and freedom is needed for a data analysis task such as the one at hand, fueled by the singularity and uniqueness of each data analysis project, which is not really provided by the ProM framework. Thirdly, given the fact that the process is highly unstructured, as will be elaborated upon in the remaining sections, none of the process discovery plugins did a good job at discovering a process model, even after simplifing and subsetting the event data.

[1] http://www.promtools.org

In order to retrieve understandable control-flow models from a real-life event log, preprocessing is often required. This requires insights in the data at hand. Hence, besides the versatile functionalities that ProM offers, there is a need for flexible exploratory and descriptive analyses of an event log.

In this respect, **R** is used[2]. R is an open source software language for statistical analysis of data and creating of graphics. R and the Integrated Development Environment RStudio were used as a main tool during this project, including data processining, data analysis, visualization, and report writing. The R-packages that were used for the analyses include *dplyr*, *tidyr*, *readr*, *ggplot2* and many others. The package *edeaR* was also used. EdeaR, the acronym for Exploratory and Descriptive Event-based Data Analysis in R[3], is a recently developed R-package [4]. The package enables the exploratory and descriptive analysis of an event log. This encompasses, among others, filtering the event log and calculating process metrics [6]. Besides allowing the exploration of data and its preparation for, e.g., its use in ProM, the metrics included in edeaR also convey useful event log insights. Furthermore, this paper, including most of the table and figures, was created automatically using RMarkdown, adhering to the ideology of Reproducible Research [5].

In contrast to R and ProM, **Disco** is a commercial closed-source process mining tool[4]. Compared to ProM, the functionalities that Disco offers are very limited as its focus is on user-friendliness. For instance: the control-flow models that Disco discovers from an event log are discovered with an improved version of fuzzy miner, which is only one of the control-flow discovery algorithms available in ProM. Given its closed-source character, Disco will only be used for visualisation purposes in this report.

3 Research questions

The research questions addressed in this paper are the following:

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.
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?
3. When is there a transition from the website to a more expensive channel, such as sending a Werkmap message, contacting the call center or filing a complaint? Is there a way to predict and possibly prevent these transitions?
4. Does the behavior of the customers change after they have send a Werkmap message, made a phone call or filed a complaint? Are customers more likely to

[2] https://www.r-project.org/

[3] https://cran.rstudio.com/web/packages/edeaR/

[4] https://fluxicon.com/disco

use these channels again after they have used them for the first time? What is the customer behavior on the site after customers have been in contact through the Werkmap or by phone?
5. Is there any specific customer behavior that directly leads to complaints?
6. Finally, we challenge the creative minds, to surprise UWV with new insights on the provided data to help improve the experiences of our customers when using the website.

4 Data Exploration

In this Section, we provide the reader with a first exploration of the data. In total, five different data sets are available. The first two data sets concern click data, both from logged in users[5] and not logged in users[6]. The thrid data set contains information about questions.[7] Finally, also information about messages[8] and compliants[9] is available.

4.1 Visitor Analysis

The datasets combined contain information about 27 412 different visitors. In this section, the demographic characters of the visitors will be discussed, as well as their behaviour.

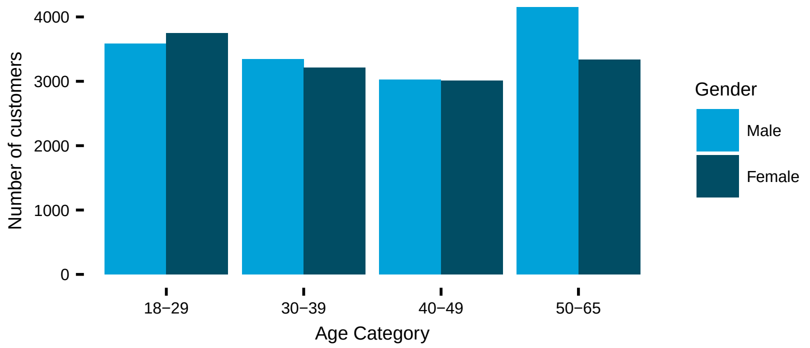

Figure 2: Number of visitors, categorized by age and gender.

[5] https://data.3tu.nl/repository/uuid:01345ac4-7d1d-426e-92b8-24933a079412
[6] https://data.3tu.nl/repository/uuid:9b99a146-51b5-48df-aa70-288a76c82ec4
[7] https://data.3tu.nl/repository/uuid:2b02709f-9a84-4538-a76a-eb002eacf8d1
[8] https://data.3tu.nl/repository/uuid:c3f3ba2d-e81e-4274-87c7-882fa1dbab0d
[9] https://data.3tu.nl/repository/uuid:e30ba0c8-0039-4835-a493-6e3aa2301d3f

Demographics Little information about the visitors is available, except for their gender and their age. Figure 2 shows the number of visitors in each of the age categories for each gender. It can be seen that there is a balance between males and females in all age categories, except for the older segment. Here, males seem to be more represented than females.

Although it seems that there are slightly more visitors in the youngest and oldest segment, this is mainly due to the larger size of these segments. Dividing the number of visitors in each segment by the number of years included in the segment would level out these differences.

Behaviour The behaviour of visitors can be analysed in terms of their interactions with UWV. It can both be examined which communication channels visitors use, and how intensive they use these channels. We distinguish four different communication channels: website, questions, messages and complaints. However, not all visitors use all four channels. The website is the most commonly used communication channel, used by 97.21% of the visitors The percentage of visitors using the different channels can be seen in Table 1.

Table 1: Presence of customers in different channels.

Channel	Frequency	Use Percentage
Website	26647	97.21
Questions	21533	78.55
Messages	16653	60.75
Complaints	226	0.82

Note that the reported 2.79% not using the website, might use it without signing in. In order to investage the intensity with which the channels are used, Table 2 shows some descriptive statistics on the amount of interations in each of the channels[10].

On average, each visitor for which website activity was recorded, spent about 25 sessions on the website, and made a total of 269 clicks. However, some outliers are present: there is a customer that spent 494 sessions and there is a visitor who made 9701 clicks (this might be the same visitor, though not necessarily).

Each visitors asked 4.5 questions on average, while the maximum number of questions asked by an individual customer is 102. Messages and complaints occured more rarely, as was already shown in Table 1. Here we see that the amount of messages by an individual customer is limited to a maximum of 21. The maximum number of complaints filed by a single customer equals 5.

[10] Note that visitors which were not related to a question, complaint or message were assumed to have zero of these. However, visitors for which no website activity was recorded, were regarded as a missing value. As a matter of fact, the clicks might be recorded while the customer was not logged in. This explains why the minimum for the related characteristic is 1 and not zero as for the other characteristics.

Table 2: Intensity of behaviour

Characteristics	Min.	Q1	Mean	Median	Q3	Max.
Number of clicks	1.00	89.00	269.26	184.00	339.00	9701.00
Number of sessions	1.00	8.00	24.79	17.00	33.00	494.00
Number of questions	0.00	1.00	4.50	3.00	6.00	102.00
Number of messages	0.00	0.00	2.41	1.00	3.00	61.00
Number of complaints	0.00	0.00	0.01	0.00	0.00	5.00

From the previous, it can be concluded that although the more expensive channels are used less than the website, they are still used by a significant amount of visitors, and these visitors use them quite often. It would therefore be interesting to compare the behaviour of these visitors to those who only use the website. This will be elaborated upon further in this paper.

4.2 Click data

In total, 16 504 352 clicks were recorded, of which 43.47% where performed by visitors logged in to the website. According to the data, 1396 different pages were visited on the website. Of these, 594 were visited by logged in users and 1368 by unknown users.

Table 3: Most commonly visited pages

Page name	Unknown	Logged visitors	Total	Relative	Cumulative
home	1518218	583545	2101763	0.13	0.13
aanvragen_ww	1575937	251063	1827000	0.11	0.24
taken	2324	1823175	1825499	0.11	0.35
cvs_zoeken	1366091	804	1366895	0.08	0.43
vacatures	1035388	78837	1114225	0.07	0.50
inschrijven	812829	148872	961701	0.06	0.56
vacatures_bij_mijn_cv	1004	953969	954973	0.06	0.62
mijn_cv	1256	880597	881853	0.05	0.67
vacatures_zoeken	342	582645	582987	0.04	0.70
zoekaantalindicatief	505339	37274	542613	0.03	0.74
mijn_berichten	551	529311	529862	0.03	0.77
werkmap	339808	181865	521673	0.03	0.80
zoekberoep	279623	28135	307758	0.02	0.82
aanvragen_bijstand	270905	12020	282925	0.02	0.84
mijn_werkmap	2538	207776	210314	0.01	0.85

Table 3 shows the 15 most commonly visited pages, which together account for 84.9% of all the clicks. The two most frequently visited pages are *home* and *aanvrange ww*. Almost all the pages in the table are visited by both logged in users as by unknown users. However, some are much more frequently accessed by users who are logged in, such as *vacatures bij mijn cv* and *vacatures zoeken*. Others are more accessed by unknown visitors, such as *cvs zoeken*. This page name, translated as *looking for curriculum vitae*, suggests that the activity of employers looking for employees on the website is included in the clicks not logged in.

4.3 Questions

The questions dataset contains information on 123 403 questions asked by visitors. The degree to which this channel is used does not depend strongly on the age category of the visitor, as the proportion of visitor asking questions ranges from 75.45% for age category 18-29 to 80.12% for age category 50-65. For customers using this channel, the number of questions asked is similar between age categories. The latter is shown in Table 4, indicating that the median number of questions asked equals four for all age categories.

Table 4: Number of questions by age

Age Category	Min	Q1	Mean	Median	Q3	max	Std. Dev.
18-29	1	2.00	5.38	4.00	7.00	46	5.02
30-39	1	2.00	5.77	4.00	8.00	71	5.46
40-49	1	2.00	5.84	4.00	8.00	102	5.81
50-65	1	2.00	5.94	4.00	8.00	88	5.63

On the content level, 94.21% of the questions are related to the theme *WN WW*, followed by theme *WN WW Faillissementen / Betalingsonmacht* covering only 1.50% of the recorded questions. Within the theme *WN WW*, the most prevalent subthemes are *Formulier Inkomstenopgave (WWZ 1-7-2015)* and *Betaling (WWZ 1-7-2015)*, accounting for respectively 21.41% and 13.29% of all questions. On a more detailed level, the ten most frequently asked questions are included in Table 5, together with some statistics on their prevalence.

4.4 Messages

The messages dataset records 66 058 werkmap messages sent by customers. In absolute terms, the degree of usage of this communication channel differs accross age categories as shown in Figure 3. Older customers tend to use the message functionality more frequently than younger customers. A similar conclusion holds when the proportion of customers within an age category that makes use of werkmap messages is determined. While 49.97% of the clients of age 18-29 uses messages, this proportion equals 67.39% for age category 50-65. For the 30-39 and 40-49 categories, these values are 61.21% and 66.33%, respectively. For customers using this functionality, the number of transferred messages also tends to be higher for older customers. As shown in Table 6, the mean number of messages ranges from 3.03 for age category 18-29 to 4.57 for category 50-65.

Table 5: Most frequenctly asked questions

Topic	Frequency	Relative frequency	Cumulative frequency
Wanneer is/wordt mijn WW-uitkering overgemaakt?	13167	10.67	10.67
Algemeen: Wanneer moet ik het formulier Inkomstenopgave versturen?	6072	4.92	15.59
Wat is de status van mijn WW-aanvraag?	4923	3.99	19.58
Ik wil een wijziging doorgeven	3593	2.91	22.49
Algemeen: Waar vind ik het formulier Inkomstenopgave?	3378	2.74	25.23
Specifieke vraag	3156	2.56	27.79
Aanvraag/Inschrijving WW	2970	2.41	30.19
Ik ga weer werken. Hoe wordt dit verrekend met mijn WW-uitkering?	2903	2.35	32.55
Probleem: Ik heb geen Inkomstenopgave ontvangen. Wat nu?	2469	2.00	34.55
Waarom is de hoogte van mijn betaling gewijzigd?	2182	1.77	36.31

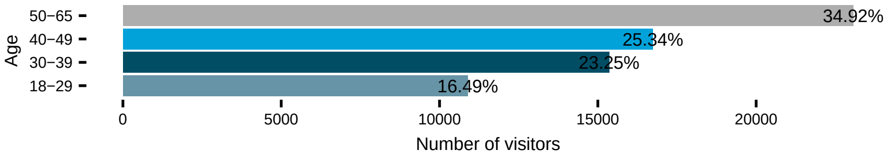

Figure 3: Number of messages per age category

4.5 Complaints

Finally, we will look into the complaints. In total, 289 complaints were filed. Table 7 displays that the number of complaints increases with age, although there is a small dip at 40-49. However, since the limited amout of complaints, assertions should be made with care.

The most commont topics of complaints are displayed in Figure 4. This figure shows that there seem to be three major causes for complaints: (1) limited or wrong information and (2) bad service provision and (3) problems with *ikf*. Since these topics cover a substantive part of all complaints, they already contain useful insights towards avoiding them.

Table 6: Number of messages by age

Age Category	Min	Q1	Mean	Median	Q3	max	Std. Dev.
18-29	1	1.00	3.03	2.00	4.00	29	2.96
30-39	1	1.00	3.83	2.00	5.00	53	3.99
40-49	1	1.00	4.18	3.00	5.00	51	4.46
50-65	1	2.00	4.57	3.00	6.00	61	4.76

Table 7: Number of complaints by age

Age Category	Min	Q1	Mean	Median	Q3	max	Std. Dev.
18-29	1	1.00	1.12	1.00	1.00	2	0.33
30-39	1	1.00	1.31	1.00	1.00	5	0.71
40-49	1	1.00	1.25	1.00	1.00	4	0.64
50-65	1	1.00	1.36	1.00	1.25	4	0.72

5 Data Preparation

5.1 Event log creation

Each of the different datasets were transformed into one overall raw event log, which can thereafter be aggregated or subsetted as desired for the analysis at hand. From the click data, only the logged in clicks were used, since only these can be connected to a specific client. These data was transformed such that for each page visited, a start and end timestamp is available. In order to do this, subsequent clicks on the same page were aggregated. Furthermore, it is assumed that a client leaves a page when another page is accessed. This results in page visits with a specific duration, and instantaneous transitions from one page to the next. As an exception to this, the last page visit in a session is atomic, as it is unrecorded how much time the client spent on this page.

For questions, the *contacttimestart* and *contacttimeend* were used as the start and end timestamp. On the other hand, only one timestamp is available for messages. As a result, these are interpreted as atomic events. Finally, only one date (without timestamps) is available for complaints. As such, assertions towards precedence relations of these events have to be made with care.

Note that this event log can be analysed from different viewpoints. For instance, one activity instance can be a visit to a specific page, but it can also be a complete visit to the website, depending on the desired granularity for a specific analysis.

5.2 Data Selection

Concerning the click data, the *page_name* attribute serves as a meaningful activity name. However, as there are many different pages, only the most frequently visited

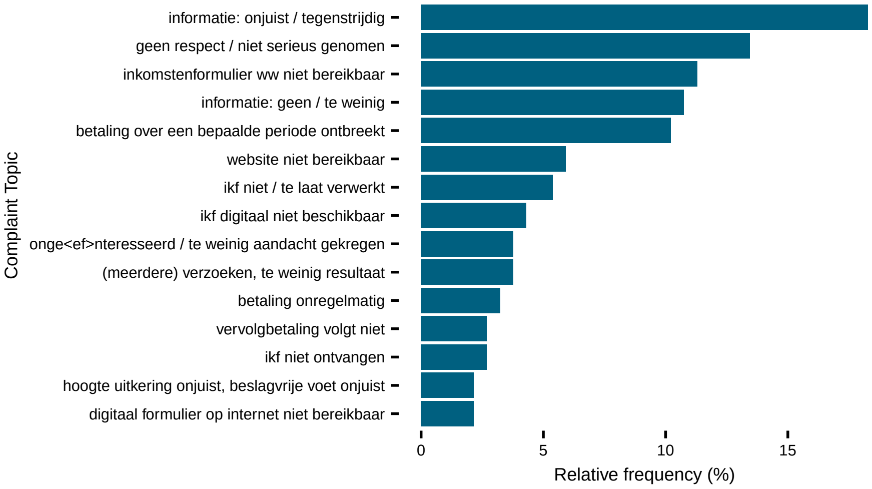

Figure 4: Number of complaints per topic

pages will be selected in the analysis. In fact, it can be stated that the distribution of frequencies for the webpages follows a power law. For instance, the 10 most visited pages already represent 83% of all page visits. It was decided to place a threshold at 95%, thereby retaining a manageable amount of only 28 different pages, while still explaining a large majority of the click activity. As for the questions, messages and complaints, all the data was retained.

6 Usage Patterns

In this Section, the analyses are targeted at answering the following research question:

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.

In the first paragraph, the structuredness of the event data will be analysed, in order to examine whether there is some typical, i.e. mainstream usage of the website. Thereafter, a cluster analysis will be performed to find different types of websessions. In the last paragraph of this section, demographic attributes will be used to find patterns related to specific user characteristics.

6.1 Log-based metrics

A first interesting fact to point out is that the behaviour in the data seems very diverse and unstructured. Even after the aggregations of equal subsequent clicks and the data subsetting as outlined in the previous section, no clear patterns emerge from the data. A total number of 170 176 different sequences of pages are recorded within the total of 658 928 sessions. Figure 5 visualizes the 10 most common sequences of page visits on the website. Firstly, one should note that due to the unstructuredness of the data, these sequences only represent 23% of the data. Secondly, it can be observed that these sequences tend to be very short, pointing out that only a few pages are visited during each session. In order to generalise the observations, the boxplot in Figure 6 shows the distribution of the number of page visits in each session. It can thus be observed that in general, sessions tend to be short in terms of page visits, although some very large outliers do exist. Nonetheless, it was found that 95% of the sessions contain at most 18 page visits.

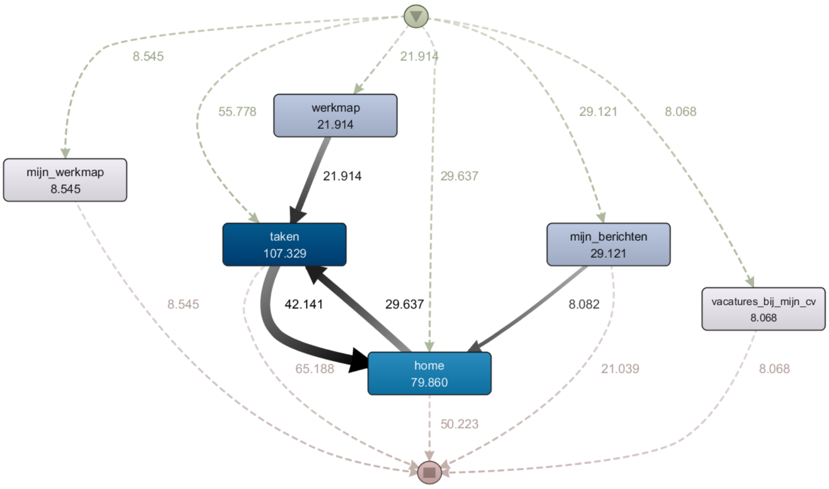

Figure 5: 10 most frequent traces

The unstructuredness of event data is is a well known phenomenon and can have several causes: (i) a large set of activities, (ii) long sequences of activities within each case and (iii) a relatively high amount of parallellism between activities. Since the sessions are typcially small and only a limited number of pages is considered, it can be stated that the unstructuredness originates mostly from the parallellism that exists between different activities, i.e. different pages. This is intuitive since a website lacks the notion of a *workflow* and all pages are typically well connected. Even though the website might be designed following a tree structure, the large number of connections between pages makes it a highly connected network. Visitors do not navigate up and down the different branches, they rather hop from one branch to another.

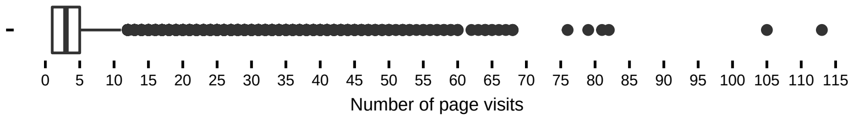

Figure 6: Distribution of number of page visits per session.

Furthermore, the unstructuredness is facilitated by the fact that, due to the lack of a *work flow* notion, there is not even one page through which each of the process instances must go. Instead visitors only look at those pages they actually need, which are limited in number and might be very different. As shown in Figure 7, only the pages *taken* and *home* occur in more than half of the sessions. Most of the most common pages are only present in less than 10% of sessions, which partly explains why sessions are short yet so diverse.

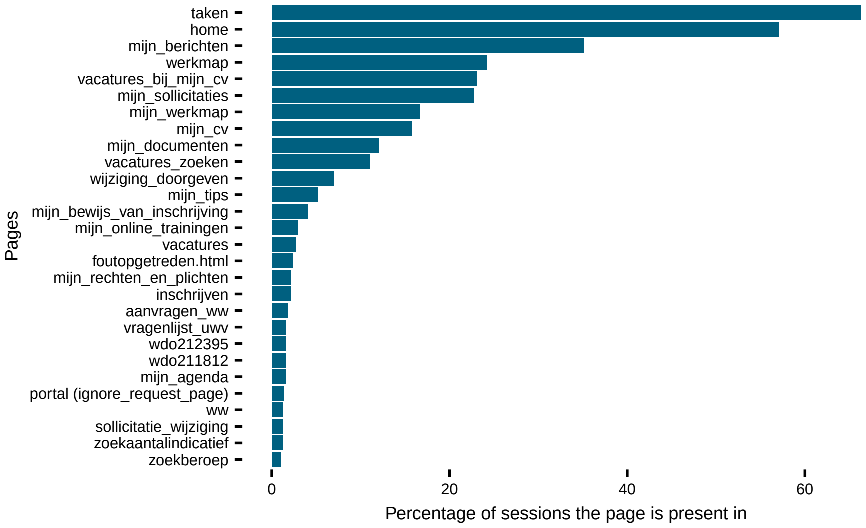

Figure 7: Presence of pages in sessions.

Additionaly, there is no clear starting point of a session. Figure 8 shows for each starting page the percentage of sessions that started on that page. While a quarter of all sessions start, logically, on the home page, there is a lot of diversity in the starting page. Clearly visitors do not start on the home page because of several reasons: they bookmarked other pages of the website in their browsers, they are directed there through search engine, or they clicked on a link somewhere. This again adds to the unstructuredness of the behaviour.

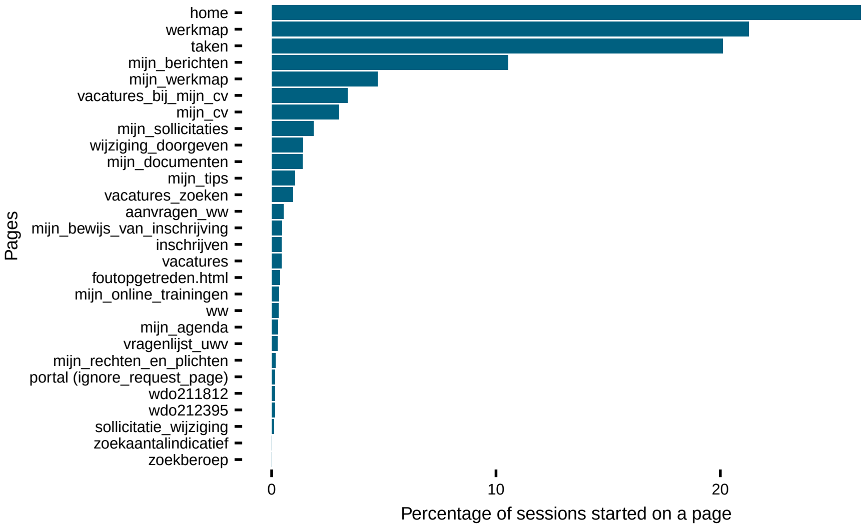

Figure 8: Start point of the sessions.

6.2 Cluster analysis

In order to find different ways in which the website was used, this section categorizes each of the sessions in a limited number of specific groups. Although algorithms for sequence clusering are available in both Prom 5.2 and Prom 6.6, these proved to be not suited for the task at hand, due to the large amount of data. As an alternative, traditional k-means clustering was chosen, based on the presence (absence) of pages in each session.

Since the number of page visits in each sessions in mostly small, it is justified to look at presence of pages, rather than precedence relations between page visits which would be used by the algorithms in ProM. However, since there are some outliers with very long sessions, these might prevent the discovery of good quality clusters. For this reasons, the clusters are defined for sessions with 18 or less page visits, which is the 99% percentile. The sessions which have more page visits are placed in an additional cluster and will be refered to as *Exceptional behaviour*.

Next to the cluster with exceptional behaviour, 7 clusters were mined using k-means clustering. The number of clusters was set at 7 since additional clusters could not improve the overall quality of the clustering. The presence/absence of pages in a specific session was used as input for the clustering. The matrix in

Figure 9 characterizes each of the seven clusters by displaying the percentage of sessions in a cluster in which a specific page was visited one or multiple times. Note that for the sake of simplicity, only the 10 most common pages are shown. The graph on the right shows the overal presence of the activities, as a baseline.

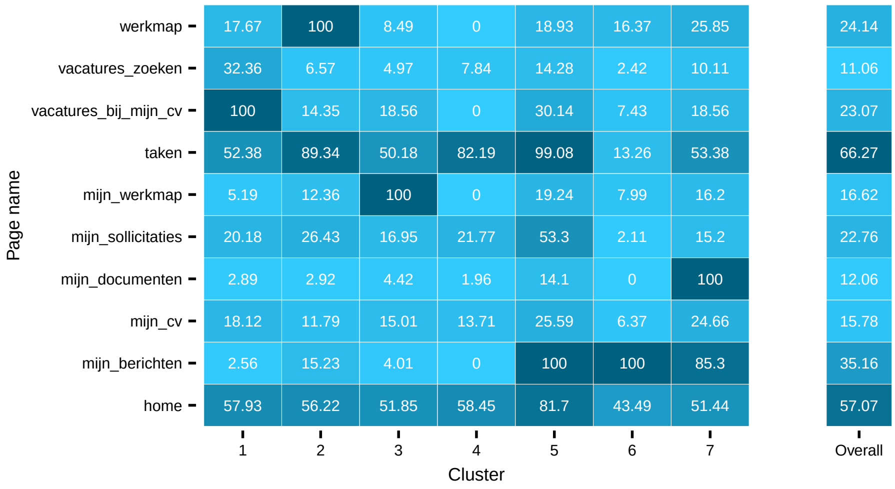

Figure 9: Characterization of clusters based on the presence of the most common activities.

When comparing the presence of pages in clusters with the baseline, it can be seen that the quality of clustering is quite good. For instance, while *werkmap* overall only occured in 24% of the sessions, cluster 2 was able to distinguish those sessions that always go to werkmap. Also, cluster 7 can be seen as a group of sessions in which one checks his documents (*mijn documenten*). Also note that the visits to the *home* page do not have a signficiant impact, they seem to be more or less equiprobable in all clusters. Furthermore, sessions in which messages are checked (*mijn berichten*) take into account 35% of all sessions. These were divided into cluster 5 to 7, depending on the occurence of other pages in the same session, i.e. *taken* (cluster 5), *mijn documenten* (cluster 7), or no specific other pages (cluster 6).

Table 8 shows the amount of sessions in each cluster and the description given to them based on the most frequent activities they represent. These will be the building blocks for different types of usages patterns analysed in this paper. Note that, despite some exceptions, the clusters appear to be relative equal in size.

Table 8: Clusters

Cluster	Number of sessions	Description
1	83547	Vacatures, Taken
2	93555	Werkmap, Taken
3	58798	Mijn werkmap, Taken
4	187751	Taken
5	80993	Berichten, Taken
6	81971	Berichten
7	53465	Berichten, documenten
8	18848	Exceptional behaviour

6.3 Differences along demographics

In the next paragraphs, we will return to the log-based metrics and the cluster analysis to see whether there are differences in sessions among persons with a different gender or among age categories. Firstly, Figure 10 shows that older visitors on average visit more pages during a session. Note that in this graph, the 5% *longest* sessions were removed in order to focus on the boxes of the boxplots. Moreover, the mean is annoted with the asterisks. Someone from the youngest category visits on average 3.12 pages, while someone from the oldest segment visits on average 3.83.

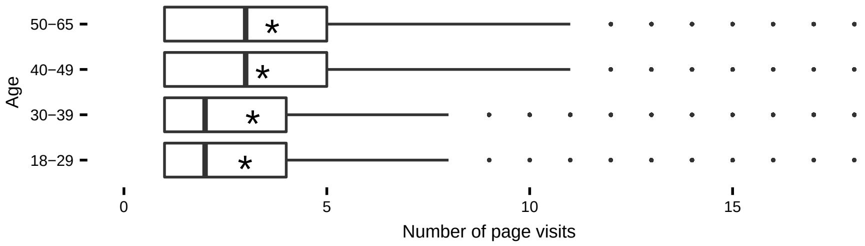

Figure 10: Distribution of the number of page visits per session, by age.

Although no differences where found concerning the most frequent sequences or the presence of activities along different age categories or between male and female users, some differences were found concerning the start page of the sessions among different age cateogries. As shown in Figure 11, it can be seen that the page *werkmap* occurs far more often as the first page in a session when users get older. For users older than 50, this page is even the most frequent start page of a session, superseding both *home* and *taken*. Furthermore, this trend seems to be much more apparent for female users than for male users.

On the part of clusters, no real impact was found for gender, i.e. men and women are equally likely to have a session of a specific clusters. Differences were

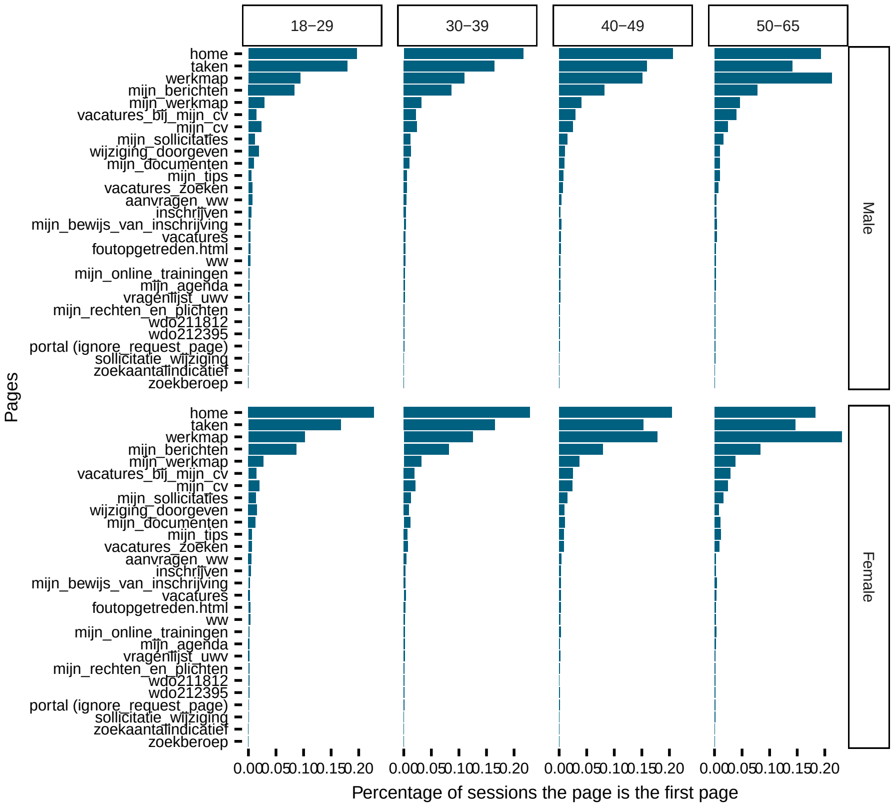

Figure 11: Start point of the sessions, by age and gender.

found, however, concerning the age category. Figure 12 shows the distribution of clusters for the different age categories. Here we can see that older people tend to have far less sessions in which they visit *taken*. While this cluster represents more than 35% of the sessions in the youngest group, it drops to about 23% in the older group. Instead, older people will look more at *taken* together with *vacatures* or *werkmap*. Other minor trends can be observed in the figure.

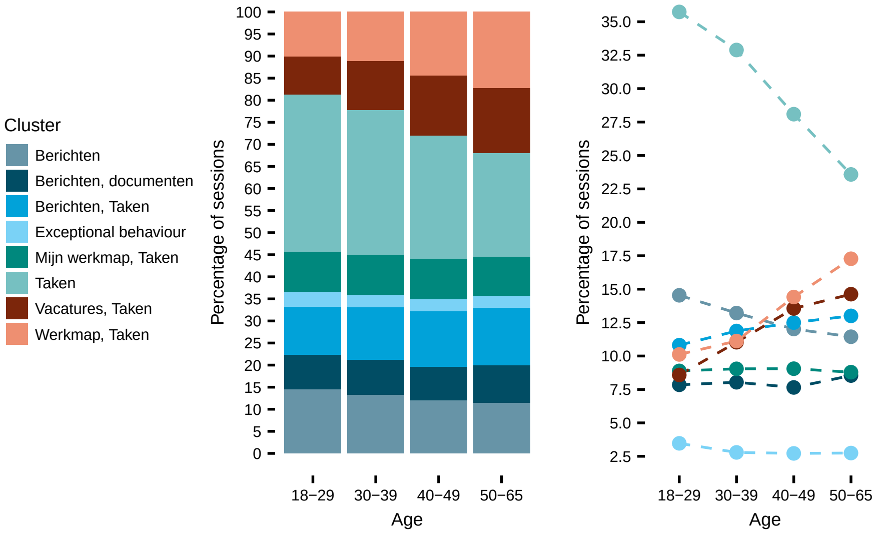

Figure 12: Distribution of clusters, by age.

6.4 Summary

In this section, click data were analysed at the level of sessions. It was found that there exists little structuredness in the data, i.e. sessions are very different from each other. This was found the be natural for the context of the application. Nonetheless, using traditional clustering techniques, it was possible to distinguish different *types* of visits, which were targetted at different pages. Furthermore it was found that older people access the website differently from younger people and, moreover, seem to have different routines while visiting: they start their visit on the website typically on other pages, they tend to visit more pages in a session, and they are more represented in specific clusters of sessions than younger people.

7 Changes in usage over time

While the previous section focussed on the webusage in a session, this section will zoom out and cast a slightly broader picture. Instead of looking at sessions, the focus will now be at visitors. Since sessions itself were already found to be highly diverse, visitors will not be described in terms of the lower level clicks. Instead, it was decided to regard each session as an *event*. The type of these events will refer to the cluster it was assigned to in the previous section. As such, 8 different activities can be found. Furthermore, the questions, messages and complaints will come into play at this point. As described in Section 5.1, these were added to the event log, and will thus be placed at the same level as the sessions. This brings the amount of different activities at 11.

In the following paragraphs, the trajectory of the visitors will be analysed in order to find common patterns and evolutions along the trajectory of the lifetime. In general, the following research questions will be adressed:

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?
3. When is there a transition from the website to a more expensive channel, such as sending a Werkmap message, contacting the call center or filing a complaint? Is there a way to predict and possibly prevent these transitions?
4. Does the behavior of the customers change after they have sent 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 behavior on the site after customers have been in contact through the Werkmap or by phone?
5. Is there any specific customer behavior that directly leads to complaints?

7.1 General metrics

Table 9 displays summary statistics describing the number of events - i.e. sessions, questions, messages and complaints - per visitor. The majority of the visitors has between 11 and 41 events. The minimum number of events is 1, which is the lowest possible number, while the maximum number is 512 events. Concerning the sequences of activities, a total amount of 25 581 different sequences can be observed describing the trajectories of 27 411 visitors. This means that a lot of visitors have a *unique* trajectory, i.e. sequence of events. It is therefore useless to look at the sequences which occur the most.

Table 9: Number of events per visitor - summary statistics

Min	Q1	Mean	Median	Q3	Max
1	11.00	30.97	22	41.00	512

7.2 Evolution of usage

The first interaction A first important fact to understand the changes in usage over time, is to see where a visitor's journey start. Figure 13 shows that more than a quarter of visitors start of with a visit to the website where they visit mainly *taken.* Remarkably, another 24% of visitors' first interaction is by way of a question. At the other extreme, only a very small group of visitors start with filing a complaint. In general, it should be remarked that these percentages differ widely from the overall distribution of clusters.

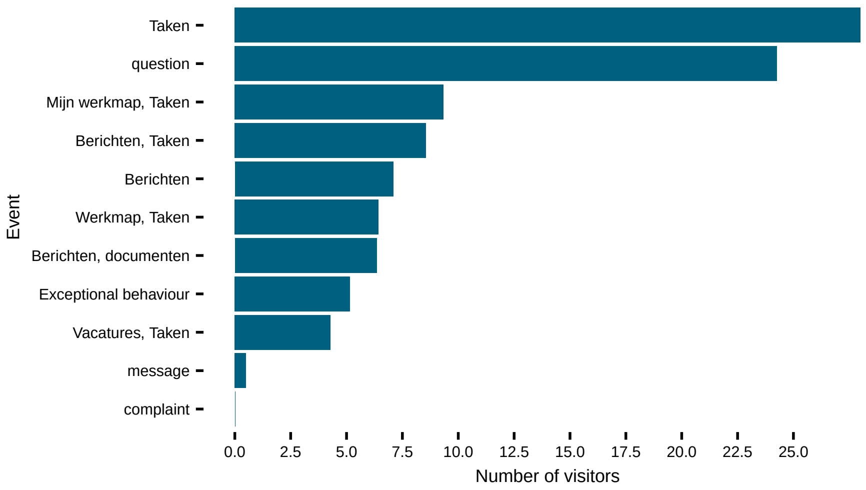

Figure 13: Starting points in a visitor's trajectory.

Stickyness of behaviour When we focus solely on the clusters, an interesting question to ask is whether a specific visitor stays with the same behaviour throughout its use of the website, i.e. the same clusters. In other words, how *sticky* are the clusters? Table 10 shows summary statistics on the number of clusters that typically occur in the trajectory of a visitor. It can be seen that the complete range from 1 till 8 is covered, and the average is quite high, i.e. five clusters.

Table 10: Number of clusters per visitor - summary statistics

Min	Q1	Median	Mean	Q3	Max
1	4.00	5.00	5.04	7.00	8

This information raises some additional questions. When only one cluster is present, which is it? Do visitors switch back and forth between clusters, or do they stick with the same cluster for some time? If a visitor's trajectory contains more clusters, are they equally spread, or is there a dominating usage? These questions will be adressed next.

Firstly, Figure shows that of the visitors who only use one of the clusters, roughly 40% used cluster *Taken*. Cluster *Berichten* follows with about 20%. The other clusters occur far less in isolation.

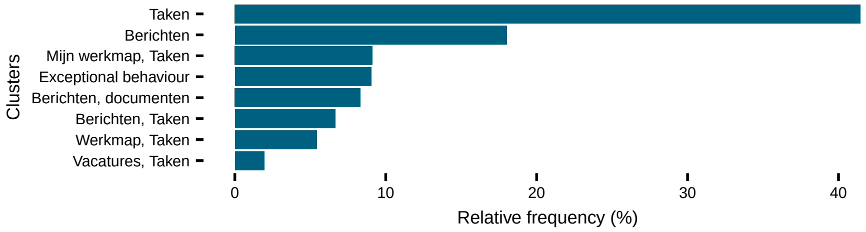

Figure 14: Frequency of clusters in trajectories with a single cluster.

Secondly, the number of *switches* between clusters per visitor were computed. It was found that, on average, a visitor switches after 2 events. In Figure 15, the number of switches per visitor are plotted in relation to the number of events per visitor. Visitors who are close to the diagonal switch almost at each new event. However, it seems that these are only a minority, and the general trend is much weaker. Overall, it can be concluded that visitors do not switch back and forth between clusters at each session, but at the same time do not stay in one specific cluster for a large number of sessions.

Thirdly, to check the existence of dominant clusters, the relative frequency of the clusters were calculated per visitor. Then, the most frequent cluster for each visitor was selected. Visitors were then divided according to the number of different clusters. For each group, the distribution of the relative frequency of the most dominant cluster was plotted in Figure 16. The larger dots point out the largest frequency if there would be no dominante clusters. For instance, when visitors use only 2 clusters, the expected relative frequency of the largest cluster would be 50% if there was no dominant usage. Note that this number is also a lowerbound. It can be observed that dominant clusters exist at any time. Even for visitors from which the usage comes from all clusters, the relative frequent of the most frequency cluster they use is about 35%. Figure 17 shows which clusters are dominant in particular.

In conclusion it can be stated that, when visitors combine usage from different clusters over their lifetime, they will not switch at every session, but only change their usage on average every 2 sessions. Furthermore, they will not equally use patterns from different clusters, but tend to have one dominant cluster from

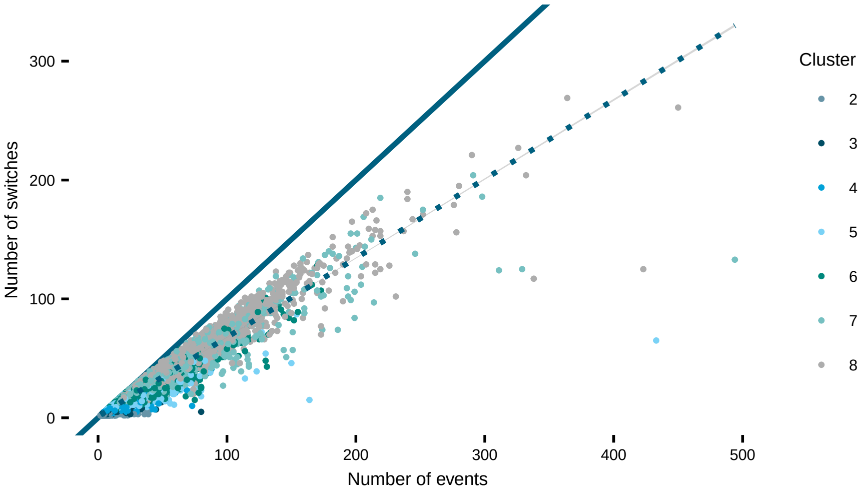

Figure 15: Switches between clusters in relation to number of events.

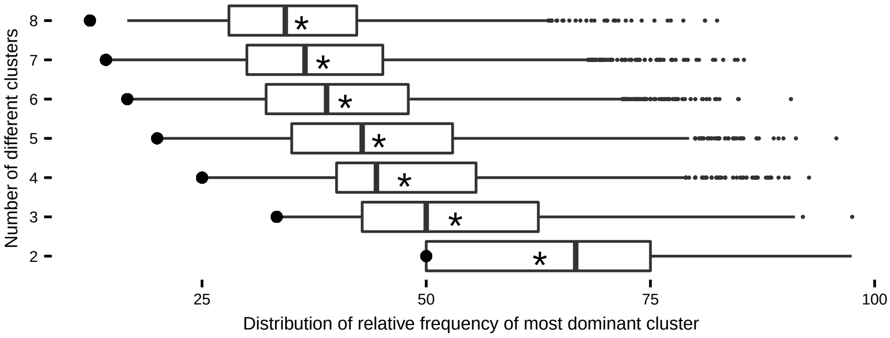

Figure 16: Distribution of frequency of the most frequent cluster per visitor, conditioned on the number of different cluster per visitor.

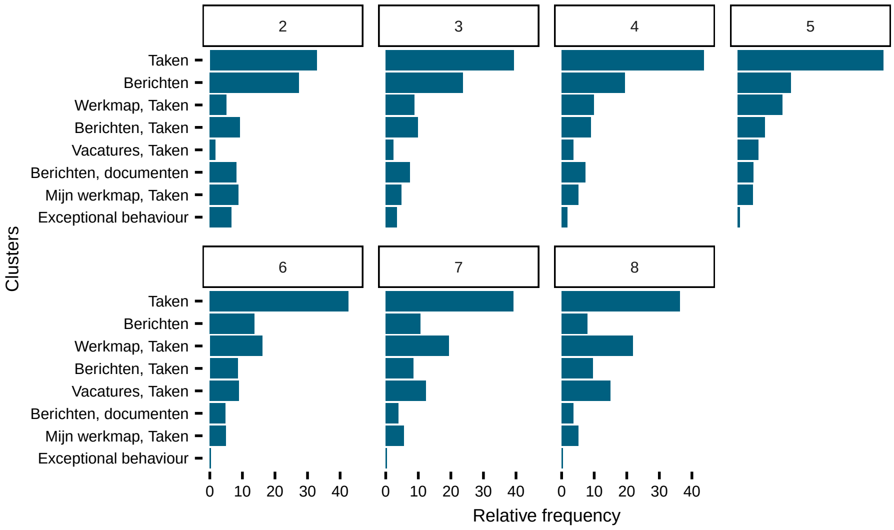

Figure 17: Dominant clusters, conditioned on the number of different clusters per visitor.

which they use the patterns the most. One specific dominant cluster emerges, i.e. *taken*, which is in line with the overall frequencies of the clusters.

7.2.1 Common switches Finally, we will look into which switches between clusters are most common. To this end, Figure 18 visualizes the flows between the different clusters. This figure is slightly simplified in the sense that it shows only the most common switches, though it represents 95% of all switches. Among the most common flows between clusters are

- *Taken* → *Berichten, Taken*
- *Taken* → *Berichten*
- *Taken* → *Mijn werkpmap, Taken*

Finally, we point out that no differences were found among different genders and age categories concerning the number of clusters in a visitor's lifetime, the number of switches, or the most typical type of flows.

7.3 Transitions to expensive channels

In the followong paragraph, the flows between clusters and questions, messages and complaints will be analysed in order to understand why these expensive channels are used by visitors. Specifically, is there a typical type of use that precedes there special events. Attempts were done to discover process models using various process discovery algorithms availiable in ProM, but these all lead

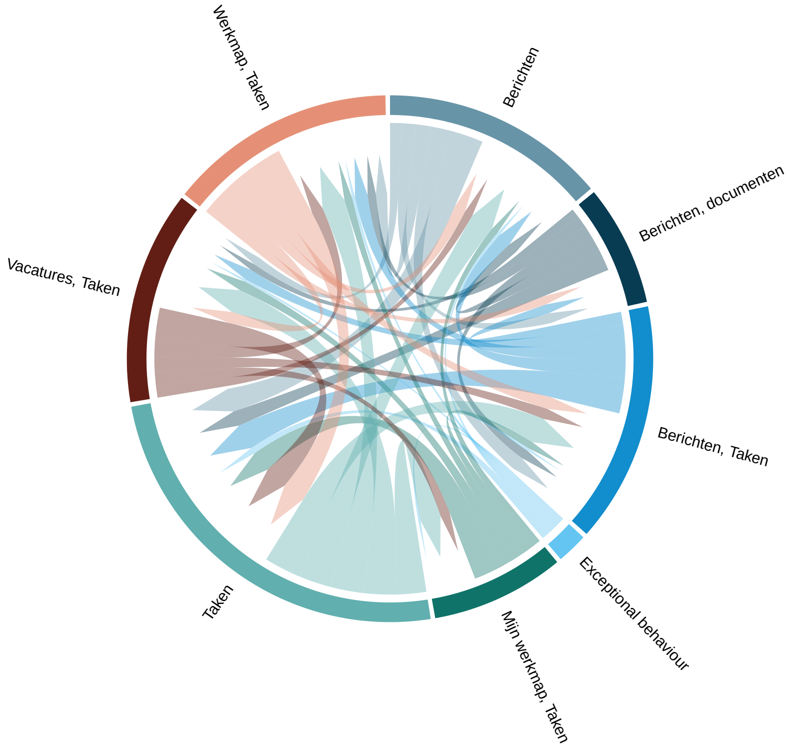

Figure 18: Flows between clusters.

to spaghetti models or flower models. As was already mentions, each visitor tend to have a unique trajectory of sessions, complaints, messages and questions, which makes it non-trivial. For this reason, alternative methods have to be used.

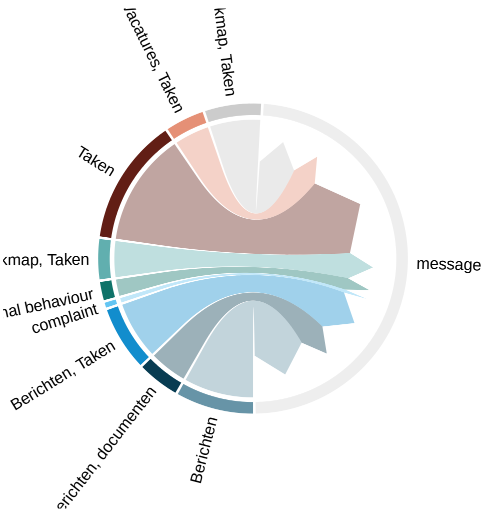

Figure 19: Precedence of messages.

Figures 19, 20 and 21 shows which type of session or which event preceded a message, a question or a complaint. Although a few observations can be made, for example, questions are preceded by messages in the majority of cases, most flow diagrams confirm to the overall distribution of the eventlog. In other words, we expect that *Taken* is more common that *Exceptional Behaviour*. Nonetheless, the flow-diagrams convey more information that the typical models which result from the process discovery efforts.

In the flow diagram for each of the *special events*, the other events and the clusters where considered as precedence activity. However, in a significant amount of cases, these events are also preceded by themselfs. For instance, 49 405 of the messages were preceded by another message, which is approximately 75% of the time. For questions, this percentage is much smaller, i.e. 14.5%. For complaints, it is 21.8%.

Another way to look at the occurences of these events is by using the observation that they often happen together over the lifetime of a visitors. This is to say, in 50% of the cases 2 different type of events happen together, and in 1% of the cases, all three events happen together. This could suggest that one of these events might increase the probability that also the other events occur.

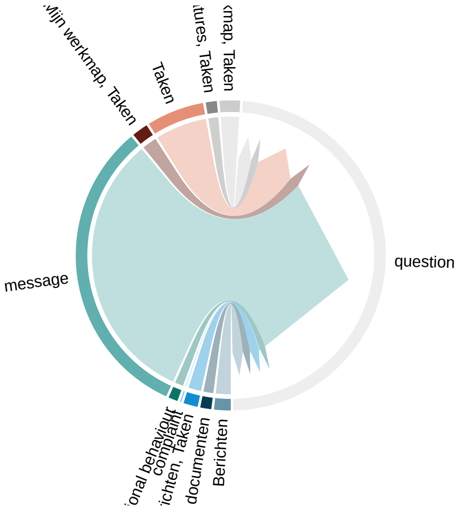

Figure 20: Precedences of questions.

Based on the data, the probability that a visitor sends a message is P(M) = 0.608 . The probability that a visitors asks a question is P(Q) = 0.786 and the probability that he files a complaint is P(C) = 0.008. Moreover the probability that a visitors asks a questions and sends a messages is P(Q ∪ M) = 0.501. Analogously, P(Q ∪ C) = 0.008 and P(C ∪ M) = 0.006.

We can then use Bayes' Theorem to compute the probability for one event given that also the other occured. The probability that a message is send given that also a question was asked is then $P(M|Q) = \frac{P(M \cup Q)}{P(Q)}$, which is equal to 0.638. Compared to the baseline probability for messages, there is only a slight increase. The probability that a message was sent if also a complaint was made is equal to $P(M|C) = 0.783$. This means that visitors who file complaints are more likely to have also have sent a message. Reversely, given the fact that a message was sent, the probability that a complaint was filed is $P(C|M) = 0.011$.Note that the presence of messages in a visitor's actvitity slightly increase the probability that he has sent a complaint. The probability of complaint, given that a question was aksed is equal to $P(C|Q) = 0.01$. Also here the probability is slightly higher than the baseline.

It is important to note, however, that these Bayesian probabilities do not tell us anything about the causal relationship. In order to created a basic notion of causality, Figure 22 shows the most frequenty flows between the three events, when all three have happened. On the left, it can be seen that if a complaint is filed, is is typical that questions and/or messages were already sent. But also after that complaints are send, these event tend to happen a lot. Although the

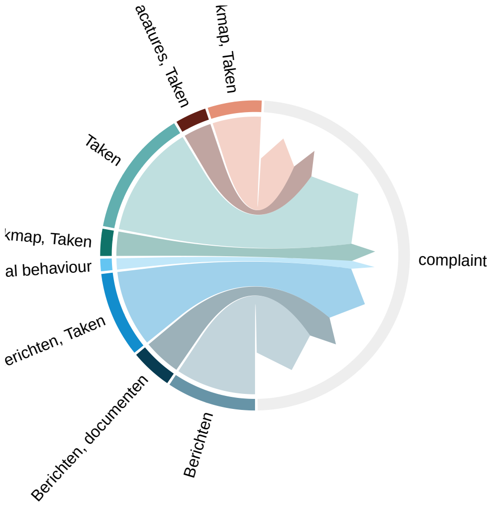

Figure 21: Precedence of complaints

situation is unclear for messages, it is clear that questions are more probable to happen before the complaints than after.

In summary, it is clear that messages and question can act as *warning signals* for complaints. It is therefore advised to thoroughly review why questions or messages are send and how they are handled, both in terms of content and in terms of service (recall that a large part of the complaints relate to the service aspect.)

7.4 The impact of questions, messages or complaints

In order to see whether behaviour differences when the more expensive channels are used, Figure 23 related the number of these event to the number of different usage types present in the visitor's lifetime, as characterized by the clusters. It can be seen that for each of the events, the more it happens, the more clusters are present and thus, the more diverse the behaviour is.

Furthermore, the usages types in terms of clusters were analyzed in relation to the different events. While no relation was found concerning the questions and complaints, some slight effects were found with respect to messages. There have been displayed in Figure . Note that only the result for less than 10 messages are shown, in order to have a significant amount of observations in each category. It can be seen that people with more messages will look more to *Berichten*, which is trivial. However, this seems to be specifically lowering the amout of time they look into *Taken*, while the other types of visits remain stable.

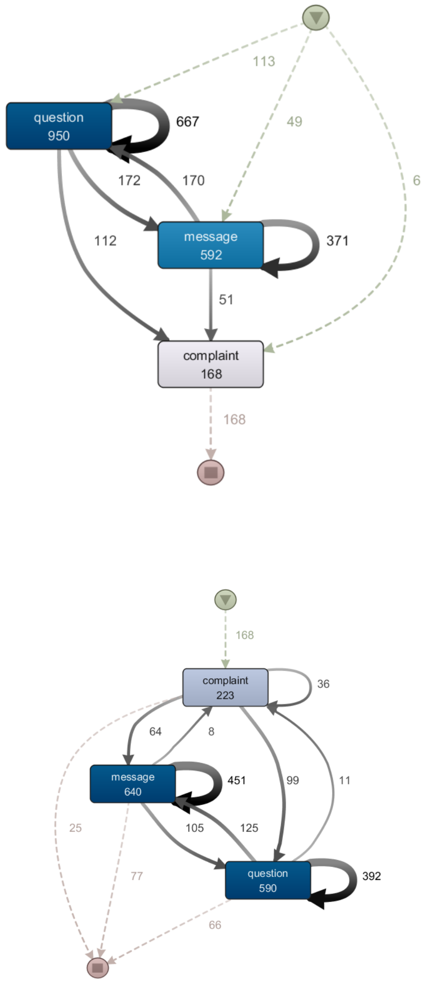

Figure 22: Flows before (above) and after (below) that a complaint is filed.

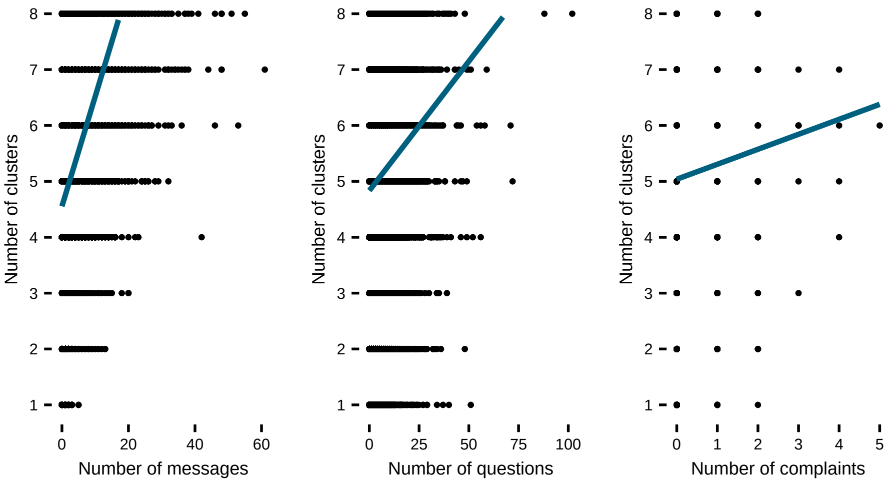

Figure 23: Number of different clusters in relation to messages, questions, and complaints.

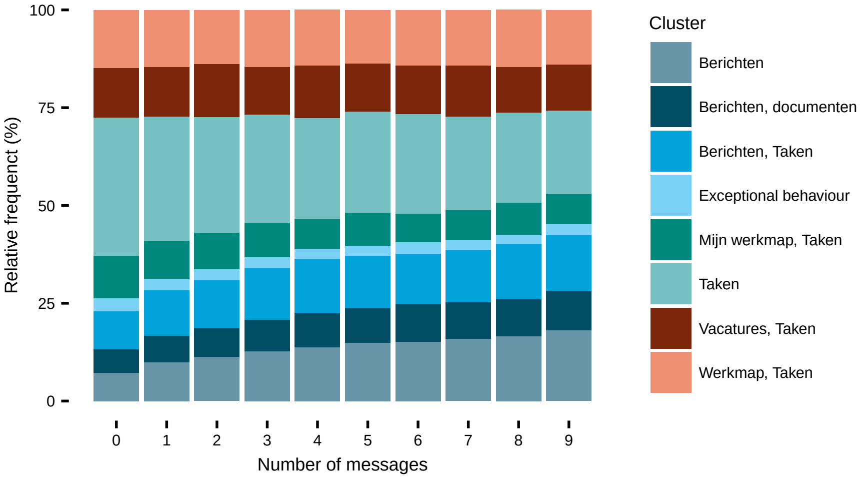

Figure 24: Distribution of clusters, by frequency of messages.

8 Conclusion

This report analysed the interactions between users of the webite werk.nl and the UWV in terms of website visits, questions, messages and complaints. The overall goal of the analysis was to understand how different users use the website and why visitors transfer to the more expensive channels or file complaints. Although it was observed that the behaviour was extremely diverse among visitors, some interesting results were found.

Firstly, single visits to the website were analysed and different types of visits were found. With respect to age and gender, it was also discovered that older people tended to start their visit on the website on other pages and had slightly other types of visits.

Secondly, it was noticed that over the *lifetime* of a visitor, their visits to the website were of different types, although on average a single *dominant* type was present. Furthermore, switches from one type to another were estimated to happen at very 2 session, on average. Which means that the average visitors does not use the website in the same way for a long time.

Thirdly, the transfers to more expensive channels were found to facilitate each other. Half of the users had used 2 of them, which is more compared to users which used only one. Also the filing of complaints was found to be correlated with the presence of message and questions. We therefore advise that the handling of these should be further analysed and improved as a means to avoid complaints.

This report will thus provide the process owner with factual insights into the interactions with the users, which it will be able to use in order to improve the experience of visitors using the website and interacting with UWV.

References

1. Aalst, W.M.P. van der: Process mining: Data science in action. Springer, Heidelberg (2016).

2. Bozkaya, M. et al.: Process diagnostics: A method based on process mining. In: Information, Process, and Knowledge Management, 2009. eKNOW'09. International Conference on. pp. 22–27 IEEE (2009).

3. Eck, M.L. van et al.: PM^2: A Process Mining Project Methodology. Lecture Notes in Computer Science. 9097, 297–313 (2015).

4. Janssenswillen, G. et al.: Enabling event-data analysis in r: Demonstration. (2015).

5. Peng, R.D.: Reproducible research in computational science. Science. 334, 6060, 1226–1227 (2011).

6. Swennen, M. et al.: Capturing process behavior with log-based process metrics. (2015).