Multidirectional Knowledge Extraction Process for Creating Behavioral Personas

Andrey A. Masiero\textsuperscript{1}, Mayara G. Leite\textsuperscript{1}, Lucia Vilela Leite Filgueiras\textsuperscript{2}, Plinio Thomaz Aquino Jr.\textsuperscript{1}

\textsuperscript{1}Centro Universitário da FEI - Fundação Educacional Inaciana Pe. Sabóia de Medeiros
Av. Humberto A. Castelo Branco, 3972
09850-901 – S. Bernardo Campo – SP – Brasil
+55 11 4353 2900
{andreymasiero, mayara.gleite}@gmail.com, plinio.aquino@fei.edu.br

\textsuperscript{2}Escola Politécnica da Universidade de São Paulo
Av. Prof. Luciano Gualberto, trav. 3, 158
05508-900 – São Paulo – SP – Brasil
+55 11 3091 5200
lucia.filgueiras@poli.usp.br

**ABSTRACT**

The consideration of user aspects in the design of computer systems is vital to maintaining the quality of these products. The search for user information as to their interests, needs, and behaviors, requires actions that can be costly. However, the use of such information in an inefficient, inaccurate manner, without the formation of knowledge discourages the process of user-centered development. Interface designers need to collect information and consume it efficiently. One option for documenting and efficiently consuming the information from user research is to apply the technique of personas. The extraction of knowledge in databases is the selection and processing of data with the purpose of identifying new patterns, provide greater accuracy on known patterns, and model the real world. Thus, this work presents a multidirectional process of extracting knowledge for creating behavioral personas. The processes of data mining are performed and compared with variations allowing a real knowledge extraction in the creation of personas. The process is multidirectional being able to be applied in both directions for the composition of any kind of personas. The results of the work demonstrate the extraction of knowledge originating from demographic information to the construction of personas focusing on behavioral aspects.

**Keywords**

Behavioral Personas, User Modeling, User Profile, Knowledge Extraction, Data Mining

**INTRODUCTION**

Digital evolution has changed the concepts of space, time and the diversity of contemporary society. Because of this companies do not need to occupy much space and can even be virtual. Cyberspace has led companies to an era when buying and selling is automated. All of these changes in behavior and needs alter the ability to provide value to companies, increasing competition and reducing the differentiation between them.

This evolutionary scenario has been slightly slower in Brazil compared with developed countries, however, many Brazilian citizens have had access to information technology to carry out many activities.

Thus, the decision to invest in creation processes based on Usability and HCI for the development of interfaces has been very helpful in the attendance of people's needs. However, in highly diverse populations there are people that are excluded by the organizations that operate in the electronic market due to the difficulty in developing systems that adapt to all sorts of human interaction difficulties.

This is a very delicate topic when it comes to government services that should focus on social welfare, inclusion and accessibility, rather than profit. Companies or government departments that provide public services have many challenges in this direction.

In the case of companies providing electronic government services, investments in usability are essential to ensure access for the diversity of users that characterizes a certain society to the services provided by the government. Therefore, the e-government system designers face challenges of access and accessibility whenever there is the need to design an interface.

Research on user profiles applied to the electronic government motivated the development of this paper. User information should be collected and consumed efficiently. In this way, inside the field of technology there is a very large range of studies that shows effective methods to understand the behavior to target audiences increasingly heterogeneous. Creating personas is an acknowledged method for modeling user profiles, which help the designer to know the user profiles that are the target of a particular product. Often, persona-based profiles only express demographic information that characterize "who" is the
user of a system and lack information that more directly influences interfaces real usage (and consequently, its design) e.g., user behavior.

Thus, this paper presents a knowledge extraction process for creating multidirectional behavioral personas. Personas are said to be behavioral because these personas were created by behavior information, which are more useful in a real interface project than only demographic information, normally used. The process is said to be multidirectional, as it gets behavioral information from demographic personas and vice-versa. Data mining procedures are performed and compared with variations allowing the extraction of knowledge in the creation of personas.

For this purpose, this paper presents the application context that motivated the research, followed by the method used to document user profiles. Next section presents the complete process used to identify behavioral personas, and its results.

CONTEXT AND MOTIVATION

In recent decades, the presence of technology has been common in almost all transactions. The same has occurred in the citizen service conducted by public enterprises, which have been strengthening the relationship between the government and people through technology. Nevertheless, the inclusion of technology in responding to citizen's needs takes into regard the high degree of diversification from the target audience, given that the attendance of public services is conducted for the whole population, without distinction between citizens.

The development of projects for this audience requires knowledge of the user observed from field instruments, data collection and analysis. Filgueiras et al. [6] developed a study on the profile of users from e-Poupatempo, in Sao Paulo. E-Poupatempo is a public place for the realization of services provided by Poupatempo through computers. According to the cited study, e-Poupatempo attendants helped at that time, on average, 30 people per day, which has justified the need to improve usability of e-gov websites and thus reduce the problems encountered by the citizens. The study was based on the personas method for modeling user profiles, developing user archetypes which proved to be representative of a large public like the Poupatempo users.

Through the process of audience segmentation, 10 personas were determined to represent the users of e-Poupatempo [6]. Each of the personas was described in some detail relevant to the usability project, providing good knowledge about the users. The main advantage of using personas is the possibility of designing the system for more than one user profile, which makes possible the realization of a project focused on users considering fictional, although representative, portrayals of the citizen. In that study, fictional portrayals of citizens presented demographical information obtained from field research and behavioral data inferred from the observant experience.

In this paper, we present the method of data mining to build personas through clusters of behavioral information and consistency checked with demographic information and expectations. The objective is to document the process of building personas based on real data.

PERSONAS

The first formal description of the method of personas in HCI occurred in 1995, when there was a need to communicate different perspectives of users on a consulting project conducted by Alan Cooper at the technological class [3, 4, 5]. Cooper defined personas: "persons are not real people ... are hypothetical archetypes of actual users ... defined with significant rigor and precision" [3].

Personas is a method developed to assist product development focused on the desires and needs of customers. Used inside the methodology of User-Centered Design (UCD), personas were considered initially as stereotypes or the result of a preconceived opinion and disseminated among the members of society. Nowadays, they are understood as archetypes that represent real beings [10].

Personas are thus compositions of real data and representative characteristics that define and detail fictitious users. Its description can be composed of information envisioned by demographic and personality characteristics and biographical fictitious beings.

Although fictional, personas represent a specific user and support to decision-making becomes more driven to real customer needs [3, 4].

Technology product developers, such as Web designers, use the concepts of audience segmentation and are not so successful because, unlike personas, they do not take into consideration the behavior, perception and context in which interaction occurs. Thus, it is possible to say that the method of personas is a qualitative study assistance in understanding the needs and objectives of clients [10]. Therefore, a personas method addresses the representation needs of diverse customer profiles leading into the most influential factors of human behavior.

For Pruitt and Adlin [10], the use of fictional characters in the development of audience and segmentation can serve a company in the following ways: it increases the usability of product and its usefulness; it improves the streamlining processes of team work; it improves a company's ability to attend customer needs while serving the company; and it increases profits.

DATA CLUSTERING

Data Clustering is a procedure for the division or merging of data collections that follow recognized standards such as clusters that are mutually exclusive. Thus, data clustering identifies data with the same pattern into a cluster. The partitions show distinct data patterns [7].

Basically, in the clustering process, groups are formed in order to minimize the distance between elements of a group and maximizing the distance between clusters.
Data Clustering is applied in this paper to group relevant features characterizing the user profile based on: demographic data, reported expectation before use and subjective satisfaction after completion of interactive activities in e-government.

**K-Means**

There is a lot of research over grouping data with similar or related information, i.e. clustering [2], [8]. Among algorithms used with the aim of clustering, K-Means is currently one of the most used due to its performance and ease of use [12]. K-means algorithm is used to create clusters with different data types like images, texts, and other kinds of data [12].

K-Means algorithm performs the following steps to a particular group of non-clustered data [7],[11]:

1. Select n objects to be the centers of clusters;
2. Repeat steps 3 and 4 while there are changes in cluster;
3. Reorganize each object to the cluster that it’s closest to.
4. Recalculate the centers of clusters.

Equation 1 shows the formula used by K-Means to define the distance between objects and the center of a cluster. When we have a smaller result of this formula it means that the object is closer to the center. The minimization of Equation 1 means that K-Means ends its execution. [7].

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2$$

Equation 1 - Minimum Square K-Means

There are some inconveniences to this algorithm. To minimize those errors in an optimal manner, either to 2 or N clusters, it takes a long time, because this algorithm is NP-hard, because you can not verify it or perform it in polynomial time, and also this is a greedy algorithm that only converges to local minimums, and only this convergence to a global minimum if clusters are well separated [7].

Due to this problem K-Means is directly dependent on the startup of the centers of clusters [13]. This case implies the standard deviation is small and the initialization of the algorithm is sparse, finding a cluster in this data can become a very difficult task.

Figure 1 shows the formation of clusters with close data (project data), the left identifies the classes in a two-dimensional space and beside the result of the partitions created by K-Means, where it realizes that the algorithm does not perform the partitions isolating each class in its own cluster. In Figure 2 we can observe the behavior of the algorithm when these classes are sparsely in space (project data).

Even with these disadvantages K-Means is widely used in scientific research and commercial applications in data mining, even with the difficulty of the algorithm to find the best solution, it can find a solution that satisfies the problem in considerable time and it's easily applicable [11].

**PROCEDURE FOR CREATING PERSONAS**

To accomplish this paper, it was necessary a data analysis from user profiles obtained through questionnaires before and after services to e-Government. Figure 3 shows the steps necessary for the implementation of this process.

Next section presents the data preparation for the process, where data will be filtered and normalized. Following the clustering process will be detailed, from the implementation of K-Means to the formation of the personas, which represent their clusters. Finally, the analysis of the personas established is discussed showing the multidirectional part of this process.
Figure 3 – Procedure for creating personas

Data Collect

Pre Service Questionnaire

Post Service Questionnaire

Database

Remove Inconsistency

Denormalized Database

Ethnographic (Group 1)

Expectation (Group 2)

Satisfaction (Group 3)

Easy to Use (Group 2)

Data Normalization

Note:
The clustering process for group (2) (3) and (4) was difficult due to a low standard deviation of data

Median

Mean

Mode

How to identify the cluster’s persona?

From this step, only personas created by group (1) were used.

Data Analysis

- Calculate clusters representation
- Which personas are equals or similar?

Query personas in original database

Exists?

Yes

No

Try to find personas from group (2) (3) and (4) based on group (1)

Process End
Preparing Database
For this research, information was obtained from two questionnaires: a questionnaire to evaluate the expectation of citizens about the computer use on the web to reach their goal of electronic services (“pre-service questionnaire”) and a questionnaire applied after use to collect user satisfaction according to experience (“post-service questionnaire”).

A large database of exactly 6580 records was collected, being 3812 pre-service questionnaires and 2768 post-service questionnaires. Pre-service questionnaires enquire users about their expectations to perform the services, while post-service questionnaires enquire about the ease of use of these services and user’s impression regarding the past experience.

The first step is a quality of data process, in which user input data is verified against invalid information, for example, age equal to zero (very common when the user does not wish to reveal his/her age). In these cases the invalid or incomplete record was removed from the database.

In sequence, the information was separated into answer records to post-service questionnaire and pre-service questionnaire, because people who answered a questionnaire did not answer the other one and the number of questions between them is also different.

Both questionnaire models have identical demographic questions with the aim to allow the junction with the other information. Because of impact on the service completion time, only one questionnaire model was applied to each citizen. Another purpose of the division was to reduce the time of participation in research, increasing the receptivity of the citizen. For the pre-service questionnaire there were 14 questions, while for the post-service there were 28 questions, presented below.

Questions made in the pre-service questionnaire:
- Age
- Educational level
- Profession
- Monthly income
- Occupation for those who do not have an income
- Digital technology already experienced
- Experience using computers
- Frequency of Internet use
- Place where Internet access happens
- Time the user expects to spend on service
- Do you expect use to the computer? (5-point scale)
- Will it be easy to realize the service?
- Will it be tricky to use the Internet?
- Do you think it will be safe to use the Internet?

Questions made in the post-service questionnaire:
- Age
- Educational Level
- Profession / Occupation
- Monthly Income
- Justification for who does not have an income
- Which equipment did you use?
- Experience using computers
- How often do you use the Internet?
- Where do you usually access the Internet?
- How much help did the caregiver provide?
- Will you use this service again?
- Was it easy to perform the service?
- Did it take long to perform the service?
- Were you satisfied when you performed the service?
- Were you nervous using the service?
- Did you feel safe when using the Internet?
- Did you find the service page?
- Did you use the scroll bar?
- Did you read text on the screen?
- Did you understand the text?
- Did you click on buttons or icons?
- Did you print a document?
- Did you use the mouse?
- Did you use the keyboard?
- Did you fill in any boxes?
- Did you use navigation links?
- Did you use a search engine?
- Did you complete your purpose in the room?

In this step a de-normalization of the original database was performed, i.e., the answers of each questionnaire were grouped in a database table for easy browsing and the manipulation of information.

Based on the two tables created, variables were grouped to the formation of eight data groups. Table 1 gives a description of each group and the motivation that justified the need for this type of division.

<table>
<thead>
<tr>
<th>#Group</th>
<th>Description</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>All variables from pre-service questionnaire</td>
<td>To find personas composed by demographic and expectation data.</td>
</tr>
</tbody>
</table>
Before applying the clustering algorithm chosen for this research normalization data procedure was performed. The purpose of this procedure is to establish a standard between these values and stay inside the same range. This action supports the algorithm to obtain a good result when the data has many variables.

This normalization study was done between 0 and 1, as it is the most common way to obtain standardization [9].

Two steps were necessary to perform the normalization of data: the first one was code the answers to numbers, Table 2 shows an example of encoding performed; the second step was the application of Equation 2, where the new data was reached by dividing the current given by the maximum of the variable which it represents.

Table 2 - Data transformation for normalization

<table>
<thead>
<tr>
<th>Before coding</th>
<th>After coding</th>
<th>After normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDA</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Up_to_R$509.00$</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>From_R$510.00$ to_R$1529.00$</td>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>From_R$1530.00$ to_R$2549.00$</td>
<td>3</td>
<td>0.6</td>
</tr>
<tr>
<td>Above_R$2550.00$</td>
<td>4</td>
<td>0.8</td>
</tr>
<tr>
<td>No income</td>
<td>5</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 2 - Data normalization calculus

$$D_{new} = \frac{D(X)}{MAX(X)}$$

The entire process was implemented using MATLAB due to its ease of use, trustworthiness and code reuse. Thus those eight groups were uploaded to the MATLAB workspace in matrix shapes to better manipulate them.

Following the implementation of the data normalization to each group, the identification phase was performed to the clusters to which each sample belonged. To this end, we applied the algorithm K-Means native in the MATLAB library which was defined as parameters, groups A to H to perform an create clusters and the number of cluster that is expected to generate from each group, in this case 10. The number of 10 clusters follows the pattern of work [1] in the same context of the electronic government.

After identifying the clusters, the relevance of each cluster is calculated in respect to group used. This calculation is given based on Equation 3 below:

Groups A and B were created in an attempt to achieve personas from a relatively large group of variables which represent different kinds of behavior and / or characteristics. In groups C, D and E only variables related to the demographic data from users were separated, and as this data was common to both questionnaires, a combination of the two sources was performed in an attempt to achieve better detail for generated personas.

The others groups - F, G and H - were separated by specific data from each of the questionnaires, representing the service expectation information, satisfaction with the service and ease of use offered by the service, respectively.

Except for groups A and B which used all the variables, two issues were not considered for this study, because they were multivalued and during the execution of the algorithm could generate a range from little interest to the results of the research as discussed in the next topics. The questions excluded were:

- Which equipment did you use?
- How often do you use the Internet?

These issues were not used in this procedure to create clusters, but were useful in other models of data transformation, also used to characterize the persona after identifying the groups.

From this point the eight groups are ready to be implemented into the remaining procedure.

Clustering procedure

In the procedure to create clusters, we attempted to identify natural patterns among data, aiming to insert them into the same group to identify similar characteristics among the information belonging to the group.
\[ R(k) = \frac{E(k)}{E(total)} \]

Equation 3 - Cluster relevance calculus

Where:
- \( R \) is relevance
- \( E \) is number of elements
- \( k \) is number of the cluster

From this moment, the clusters are defined, and we also know the relevance of each to their origin group, but can not identify which persona that characterizes each cluster.

To identify these personas we used three arithmetic methods: mean given by Equation 4, median represented by Equation 5 and mode that presents data of higher occurrence within the variable.

\[ \text{mean} = \frac{\sum_{i=1}^{n} X_i}{n} \]

Equation 4 – Mean calculus

\[ \text{median} = \frac{(n + 1)}{2} \]

Equation 5 – Median calculus

Through these methods 30 personas were obtained, 10 for each method above, using the information of the 8 groups of data composed in the data preparation, totaling 240 personas generated by four different types (demographic, expectation of use, ease of use and satisfaction of use).

Analyzing Personas

After the identification of personas, personas were selected from data group C, which has demographic data from both questionnaires to analyze on a higher level of detail.

During the detailed analysis of the personas created by each method it was noted that some of the personas generated actually belonged to another cluster, due to their similarity or difference of tiny data. Thus, we eliminated duplication and increased the reference cluster values for every one which owned such similarity.

In the end of the analysis of each one of the three methods (mean, median and mode) the following results were obtained:

- For median, instead of a total of 10 personas ratified with 9 personas.
- For mode, instead of a total of 10 personas ratified with 9 personas.

At this point a search in the database was carried out for each of the personas formed by checking the existence of those, and at the same time finding some of the personas of user satisfaction, expectation and ease of use, which corresponded.

Table 3 presents how many instances of each persona were found in the databases.

<table>
<thead>
<tr>
<th>Person</th>
<th>Pre</th>
<th>Post</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>39</td>
<td>89</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>25</td>
<td>46</td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>58</td>
<td>108</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Person</th>
<th>Pre</th>
<th>Post</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>21</td>
<td>39</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Person</th>
<th>Pre</th>
<th>Post</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>10</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>40</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>34</td>
<td>40</td>
<td>74</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>60</td>
<td>30</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 3 - Count the number of occurrences per persona
The information in Table 3 shows the existence of personas with great representation in the database for both the questionnaires. Furthermore, for the cases of personas generated by the methods, mean and median, are personas composed of information from the database, however there is no similar combination of those stored in the database.

To identify the cases of complex personas just identify those personas which have no occurrence, for cases that have obtained a result in the selection greater than zero, it was possible to find a corresponding persona, created from the data of the groups F, G and H, because these presented the same combination in some records from the database.

RESULTS
When clustering was performed with data of the groups A and B those results were not satisfactory. The K-means algorithm did not achieve a good behavior while working with too many variables - 41 from the database of post-questionnaires use and 27 from the database of before questionnaires use - and personas generated in 80% of the cases are equal to each other. This occurred because the greater variation of data was between the answers of multivalued issues, which affected growth of the other values.

As for those groups F, G and H, creating clusters achieved a satisfactory value because there was good diversity among those generated personas. However, these clusters when generating the algorithm were very intermittent, due to the standard deviation between this data lies very low making those clusters often stay empty providing an error in the algorithm K-Means.

Between groups C, D and E, clustering was successful as expected and very similar among the 3 groups, but it is clear that when using a larger amount of data the result presented is better.

With this fact just personas generated from the group C have been chosen to perform analysis in more detail, as it contained data from both questionnaires.

Verifying those numbers of occurrences of each persona in the original database and identifying that the mode was the only method that showed occurrences from each of the personas directly in the database.

The mean or median may slightly change the values and end up in values which originally there is no occurrence in the database, unlike the mode which only returns the highest frequency value in the data. Therefore execution of the method for mode can achieve the most return queries in the database.

Though we can observe which of those personas was generated from mean and median, which have high relevance, we also found through mode, thereby serving as a way of affirmation for these personas. In this case those other methods support the ratification aiming procedure of a more representative persona.

Another interesting result is that in most cases for each persona of group C reached in the database, it was also found such as most frequent values, personas of user satisfaction, expectation and ease of use - belong to group G, F and H respectively - match, see Table 4.

As a last step of procedure, it was noted that there was a correspondence between personas through the opposite path, starting from the personas of user satisfaction, expectation and ease of use to demographic and the result was positive for personas to occurrences in the database presented Table 3.

In general, the highest combination of occurrences found, existed between personas considered appropriate after clustering and analysis in any of the paths opted to combine them. See example below in Table 4:

<table>
<thead>
<tr>
<th>Demographic Persona</th>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>Educational Level</td>
<td></td>
<td>high school</td>
</tr>
<tr>
<td>Monthly Income</td>
<td></td>
<td>from R$ 510.00 to R$ 1.529.00</td>
</tr>
<tr>
<td>Without Income</td>
<td></td>
<td>not answered</td>
</tr>
<tr>
<td>Occupation</td>
<td></td>
<td>Adjutant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assistant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Attendant</td>
</tr>
<tr>
<td>Computing Experience</td>
<td></td>
<td>I use the computer and have made courses</td>
</tr>
<tr>
<td>Frequency of Internet use</td>
<td></td>
<td>Frequently</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expectation Persona</th>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where do you use the computer?</td>
<td></td>
<td>Alone</td>
</tr>
<tr>
<td>Complicated to use the internet?</td>
<td></td>
<td>Disagree</td>
</tr>
<tr>
<td>Do you feel safe on the Internet?</td>
<td></td>
<td>Agree</td>
</tr>
<tr>
<td>How long did it take to use the service?</td>
<td></td>
<td>Up to 5 minutes</td>
</tr>
<tr>
<td>Was it easy to use the service?</td>
<td></td>
<td>Agree</td>
</tr>
</tbody>
</table>

| Number of Demographic Occurrences from Expectation - 7 of 50 |
| Number of Expectation Occurrences from Demographic - 13 of 212 |

Table 4 – Comparing personas
The result presented in Table 4 shows the possibility to create personas typified from a database and make the combination of personas to attend the needs of the designer. In the last two rows of Table 4 the number of times the expectation persona appears is located when the demographic persona was searched in the database, for example, searching exactly demographic persona data displayed in the table identifies 50 occurrences of the expectation personas, bringing together these results the highest number of occurrence identified - 7 - is exactly the expectation persona displayed in the table.

As the process performed is multidirectional searching the expectation persona, the greatest number of data occurrence found in the demographic personas composes exactly the same as Table 4.

CONCLUSION

Upon execution of this study, it is concluded that despite some limitations the k-means algorithm showed was a tool to a positive result in the application of this procedure, even with the difficulty of the data clustering groups A, B, F, G and H.

We have presented a procedure to build personas from the analysis of clusters identified by K-Means. Furthermore it is verified that personas can be accurately identified by the process of derivation of the profile to data from database or derivation to identity between the typification of personas.

The creation of personas can occur by simplified methods. The method is simple when using limited data set or the characters are created by the imaginary scenario process. The method presented in this paper is effective for large databases. Especially when the data are group data: behavioral data in this work. It is a method of creating personas is not simplified; it applies data mining process and proof of training data. The personas created with this method have been recognized as characters that identify individuals representative of the environment that the data were originally collected.

With the objective of this study accomplished to the definition of the process established, efficiency measurement research of the application process has continued. It is an important step to establish parameters to complete the automation of the process with the aspects of artificial intelligence.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of the team from e-Poupatempo represented by Iara Barbarian and Carlos Torres of the Operations Superintendent (SCO) of the São Paulo State. Special acknowledge to "Fundaçao de Amparo à Pesquisa do Estado de São Paulo" (FAPESP) for financial support for this publication.

REFERENCES

7. Jain, A. K. Data clustering: 50 years beyond k-means. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR), 19th International Conference in Pattern Recognition (ICPR), 31, 8 (June 1, 2010), 651-666. doi: 10.1016/j.patrec.2009.09.011.