Segmentation of the PAYE Anytime Users

Jessica Clancy, Giuseppe Manai and Duncan Cleary
Irish Revenue Commissioners, Ireland
Jessica.Clancy@revenue.ie

Abstract: PAYE anytime is a web application designed and implemented by the Office of the Revenue Commissioners of Ireland. The application allows the Pay As You Earn (PAYE) customers in Ireland to manage most of their tax affairs online. By using easily accessible technology, PAYE customers can update their information and process most of their tax credits and reliefs online in a clear and effective manner. This online system was designed and implemented in order to reduce the volume of direct contacts between Revenue and its PAYE customers, to decrease costs and improve overall efficiency and effectiveness within the organisation. Moreover, the usage of such an e-channel allows Revenue to record important information that can be analysed with the aim of improving overall customer service. Therefore, management of this strategic contact channel is paramount to Revenue's continued advancement and improvement of its online services. This paper describes a segmentation of PAYE anytime users. The segmentation was conducted to understand the profiles and behaviours of these customers. This unsupervised data mining method produces an unbiased, self directed portrait of PAYE anytime customers. The data analysed were extracted from the weblogs of the PAYE anytime online application, which contains information about the users' navigation. The data were linked to the users' information held in the Revenue data warehouse in order to access all recorded details about PAYE anytime users. This information consists of the tax credits claimed, the value of tax credits, time period and similar attributes. By linking the online behaviour with the users' information and mapping on the demographic details of the users, it was possible to identify the different segments and their profiles. The results of this segmentation improve Revenues understanding of the PAYE customer base. Knowledge gained with this project can be applied in a number of areas. Naturally, the profiles and behaviours associated with each segment can be strategically used for customer intelligence policies, allowing specific services to be tailored around customer profiles. Moreover, the analysis can point to improvements of the design and structure of future iterations of the PAYE anytime application.

Keywords: segmentation, weblog analysis, association analysis, data mining, customer behaviour

1. Introduction

PAYE anytime is a web application designed and implemented by the Irish Office of the Revenue Commissioners (Revenue), which allows Pay As You Earn (PAYE) customers to manage certain aspects of their tax affairs online. This paper describes a segmentation of the PAYE anytime users, which was conducted to inform Revenue's knowledge of the different online behaviours of PAYE customers. The results will improve Revenue's understanding of the PAYE customer base, especially in relation to the designing of customer intelligence policies, tailoring customer services and improving the efficiency of PAYE anytime.

This focus on understanding the PAYE anytime customers is vital at a time when Revenue has reduced direct contact to customers by approximately 2.5 million letters in 2010. This is due to an operational decision by Revenue to withdraw the direct issuing of the annual Tax Credit Certificate (a statement of Tax Credits for the coming year) to each PAYE customer. Instead PAYE customers can contact Revenue via PAYE anytime or other self-service channels to request a Tax Credit Certificate, if required. Due to this Revenue derived change to the direction of communication, it is imperative that customer-controlled contacts are examined now.

Examining the customer-led interaction on PAYE anytime may assist in a greater understanding of the needs, attitudes and behaviours of these PAYE customers. By generating a deeper insight into the PAYE online customer, it is envisaged that areas can be identified to improve Revenue cost efficiencies and customer service standards and to further improve the PAYE anytime customers' experience. Revenue believes that increased usage of online methods leads to increased in-house efficiency, i.e. 'do more with less.'

This paper will provide a background to Revenue and the PAYE anytime service, explain the steps taken in the segmentation process and discuss the results of the segmentation.

Reference this paper as: Clancy, J, Manai, G and Cleary, D. "Segmentation of the PAYE Anytime Users" *Electronic Journal of e-Government* Volume 8 Issue 2 2010, (pp105-120), available online at www.ejeg.com

2. Revenue in context

The core business of Revenue is the assessment and collection of taxes and duties. PAYE is the tax on income paid in the Republic of Ireland by individuals that are termed *Employees*. Employees are taxed *at source*, where the employer deducts the tax due each time a payment of earnings is made to an employee.

Based on personal circumstances every individual is entitled to tax credits, which reduce the amount of total tax payable. The penultimate responsibility for ensuring that PAYE customers receive the correct amount of tax credits rests with the PAYE Customer(s). However, Revenue, in its 2008-2010 Statement of Strategy, has included a goal, which seeks to '*Provide quality and innovative service that supports all of our customers'*. Under this goal a strategy to *'Help customers pay the right amount and to get their entitlements'* has been devised. One of the main performance indicators for Revenue, in relation to its PAYE customers under this goal, is to implement '*Easier to use PAYE self-service channels leading to greater take-up with a target by year end 2010 of 300,000 using the service at least once a year*' (Revenue 2008).

There are a number of contact channels, which PAYE customers can use to claim tax credits, such as letter or fax, sending a form, telephone calls through the 1890 Lo-Call phone lines, call in person to a Public Office, text messaging facility and PAYE anytime. The Research and Analytics Branch of Revenue has carried out PAYE customer surveys, and found that the satisfaction levels with customer contact methods are generally positive, however, the most common form of contact is via 1890 Lo-Call phone lines (Revenue.ie, 2008).

3. An introduction to PAYE anytime

PAYE anytime is the latest version of Revenue's online service for PAYE customers. It facilitates PAYE customers by providing services such as claiming tax credits, declaring additional income, viewing and updating their personal information held by Revenue. PAYE anytime has the following benefits for PAYE customers (Revenue.ie, 2010):

- Speed and flexibility;
- Quicker processing of refunds;
- Instant acknowledgement of receipt of updates;
- Paperless; and
- Reduced clerical errors.

As well as benefits to customers, PAYE anytime also provides benefits to Revenue (Revenue.ie, 2009):

- Increased efficiency and effectiveness, thereby allowing for staff redeployment to less routine work;
- Data are stored electronically in a single location thereby removing the need for paper files and duplication. This also makes the data easier to examine for audit and compliance purposes;
- Enabling Revenue to meet its commitment to the e-Government agenda (Department of Finance, 2009).

Electronic contact channels allow Revenue to efficiently distribute information directly to parties concerned in a cost effective way. It also permits Revenue to develop the on-line facilities to adhere to the World Wide Web Consortium's Web Content Accessibility guidelines (W3C, 2008).

4. Background to the research

4.1 PAYE customer segmentation (2008)

In September 2008, Revenue's Research & Analytics Branch carried out a segmentation of the 2.2 million PAYE customers in Ireland. As this was the first PAYE data mining project undertaken by the Research & Analytics Branch, it required extensive data identification, preparation and manipulation.

Ultimately, the segmentation used tax record and demographic data from a number of sources to identify four actionable segments. The segmentation was carried out on the tax profile of the PAYE customer and subsequently the demographic data were extrapolated onto the segments to identify their profile.
The segments were distinguished on income levels, age groups, and level of engagement with Revenue. The level of engagement is an important factor because PAYE customers are not required to make an annual tax return or have any self-directed contact with Revenue.

The four segments identified can be briefly described as follows:

- Middle Income, private sector employees, property owning and somewhat engaged
- Low earning working taxpayers – disengaged
- Non-national tax paying workers who rent, engaged
- Public sector / unionised / uniform wearing workers, engaged

This initial PAYE customer segmentation was carried out to enable Revenue to gain a greater understanding of its PAYE customer base. As this was a first attempt at PAYE customer segmentation many lessons were learned during the project, especially in relation to data identification and manipulation. These lessons have since been employed during other projects that have used PAYE data, and indeed in other segmentation exercises.

4.2 PAYE anytime customer segmentation (2010)

In December 2009 the PAYE anytime system was enhanced. As part of this upgrade a number of features were added to improve the service including the ability to log the activities carried out by the PAYE anytime users on the site.

The introduction of this weblog capturing system allows Revenue's Research & Analytics branch to gain insight into an engaged section of the PAYE population. This PAYE anytime customer segmentation is unique as it focuses on the profile of the users of this online service through their self-directed interaction with Revenue, i.e. weblogs. Demographic data are subsequently appended onto the segments to provide a demographic profile of these customers.

The PAYE anytime segmentation project was conducted to generate an overall view of the PAYE anytime population of online users. The main goal of the project was to answer the following questions:

- *Who* are the PAYE anytime customers?
- *What* do PAYE anytime customers use PAYE anytime for?
- *How* can the PAYE anytime service be improved?

The examination of PAYE customers that are in direct contact with Revenue is an extension of the general PAYE population segmentation that was carried out 2 years previously. As discussed in section 3, Revenue recognises the benefits provided by encouraging more customers to utilise self-directed contact channels. It is envisaged that by improving Revenue's understanding of its current PAYE anytime customer base more PAYE customers can be encouraged to use this service.

5. Analytical methodology

In this section the paper will explore the analytical methodology used to create the PAYE anytime segmentation process. The schema in Figure 1, details each of the steps taken in the analysis and the methods used. This schema will be explored in the remainder of this section.

5.1 Raw data

The PAYE anytime segmentation is based on weblog data, captured from the PAYE anytime online service. The weblog data are taken directly from the PAYE anytime weblogs recorded between 1st January and 7th of February, 2010. These weblogs consist of a record of each web page that the customer visited during a specific session; moreover multiple accesses from the same user are linked.

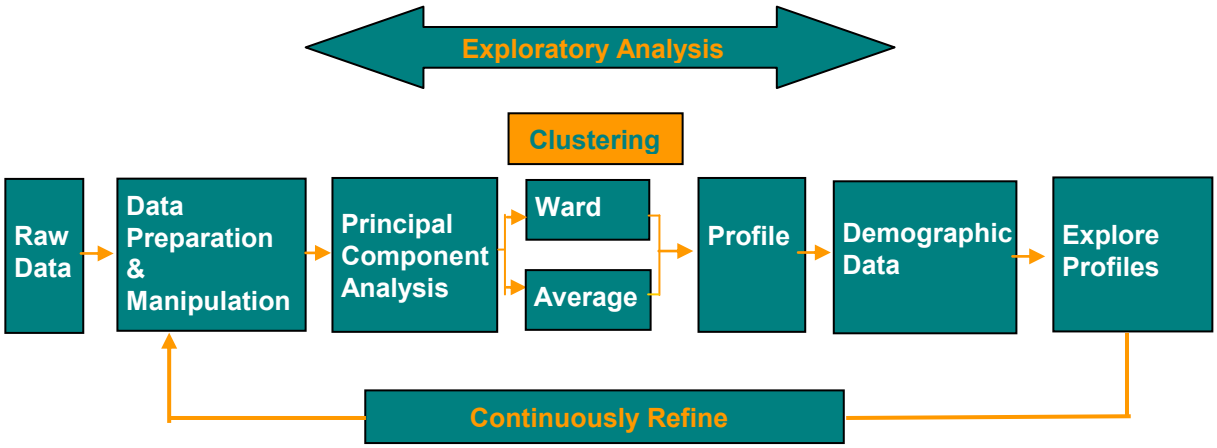

Figure 1: Schema of segmentation process

Once the segments have been identified the weblog data are then linked to the demographic data of the PAYE anytime users. This process allows the different types of PAYE anytime users and their demographic profiles to be identified.

Demographic data are extracted from the PAYE customer's records and include details such as gender, age, number of children and nationality, etc. As a matter of course, all data were anonymised prior to analysis.

5.2 Data preparation and manipulation

The weblog data were initially captured on a session basis with a line of data for each customer's impression on a web page. The data were manipulated in order to have a final summary dataset with one line of data per user. This dataset ensures that each member of the PAYE anytime population is represented once, i.e. users with multiple visits to the site are not over represented.

However, the data supplied by users with multiple visits not only to the site but to various pages on the site was not lost. These data were transformed from its original format into a frequency count of the number of times each user visited each web page. Once this step was carried out, the demographic data were linked using Revenue's unique customer number.

The demographic data were subject to a number of tasks that were performed iteratively as part of the continued refinement step of the segmentation process. These tasks included the creation of age groups from the original birth date and grouping nationalities by continent. Missing data were also reformatted, and in many cases the missing value was imputed as a zero value, this ensured that the record was included in the segmentation process.

5.3 Exploratory analysis

The Exploratory Analysis phase of the Segmentation Process begins at the Data Preparation & Manipulation stage and runs in conjunction with the Principal Component Analysis and Clustering Stages. The results from the Exploratory Analysis contribute to each of the three applicable stages of the segmentation process and they will be described in this section.

5.3.1 Association analysis

In order to gain a better understanding of the weblog data, initially an Association Analysis was carried out. An Association Analysis is an unsupervised data mining technique used to identify relationships between variables in the dataset.

This analysis technique used the data in its original session based format. There are three main variables used in this analysis: the user, the web pages and the specific session. The analysis is

carried out in two steps; first, the relationship between the user and the web pages and second, the relationship between a specific session and the web pages, are explored.

The first analysis was conducted to understand the *links* between users and web pages, i.e. what pages are visited most. Results are presented in a link graph in Figure 2. A link graph shows nodes of items within the dataset that are connected to each other. The line linking the items represents a connection between the two items. The higher the confidence of the connection, the thicker the line between the two nodes (Battioui, 2006).

The link graphs in this paper show representative general categories of the areas of PAYE anytime utilised by the customers rather than the actual names of the web pages, to ease the link graph interpretation. For example, the "'06, '07 & '08 Entries" groups pages related to tax credits for tax years, 2006, 2007 and 2008. The link graph shows that customers visited a number of web pages related to different tax credit for these historical tax years.

The category "Pay & Tax" represents a specific section on the web site where customers are requested to enter values of pay received and PAYE paid if those details are not already on record. Not all PAYE anytime users are required to do this. If a PAYE customer is required to enter these details, then they must do so before they are permitted to make other changes to that historic tax year. On the link graph these pages are linked to each other but are not linked to other web pages. This may be due to the limitations imposed on the customer until the Pay & Tax details are entered.

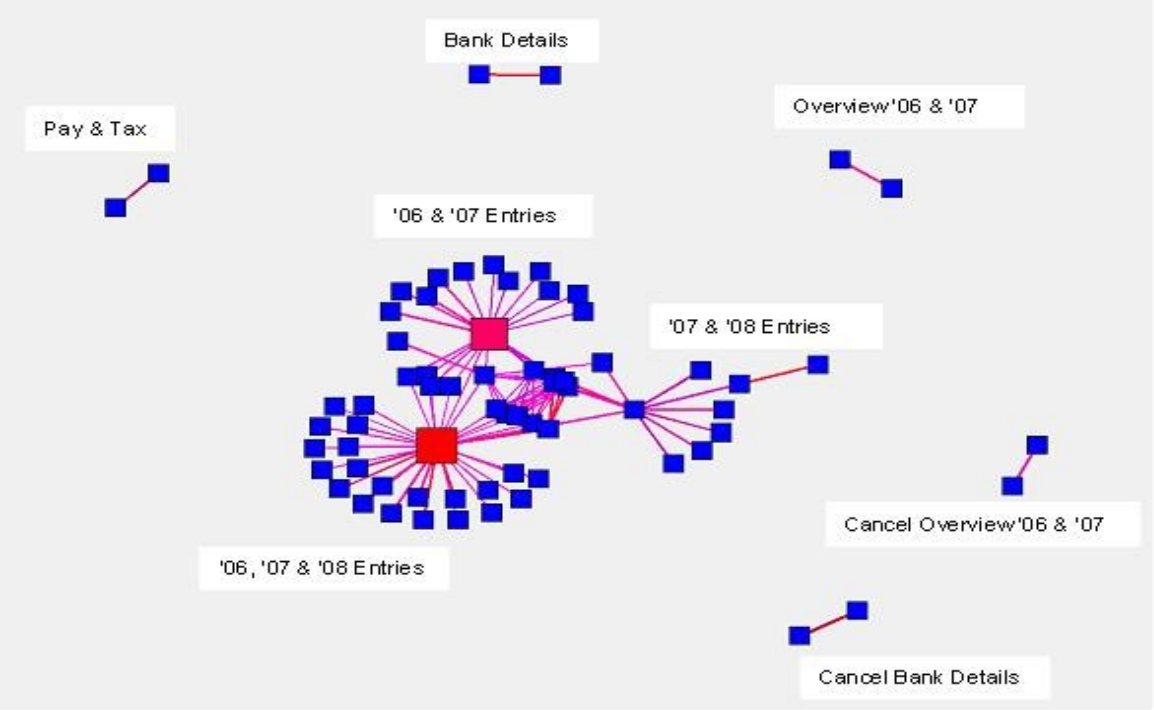

Figure 2: Link graph – PAYE anytime user and web pages visited from 01.01.10 to 07.02.10

A second Association Analysis was conducted to identify the web pages visited in a specific session. This analysis is presented in the link graph in Figure 3. This link graph displays differing node connections in comparison to the initial link graph.

The tax credit updates in this analysis have moved from 2006-2008 to the more recent 2008-2010 tax years. On a session basis the results indicate that users spend more time on the main information and overview pages, these pages did not feature in the previous analysis. The pages for tax years 2006 and 2007 are also visited on a session basis but are only linked to each other. This indicates that users visited these pages in isolation and not in connection with any other historic tax years.

A comparison of both link graphs shows that there is a vast difference between the type of interaction between a customer and PAYE anytime over a month long period as opposed to their interaction on a

session basis. The results of the Association Analysis appear to confirm a generally held concept within Revenue, that PAYE anytime users re-enter the site on a number of occasions to update their tax affairs. However, further research, including qualitative analysis, is required to ascertain why this happens.

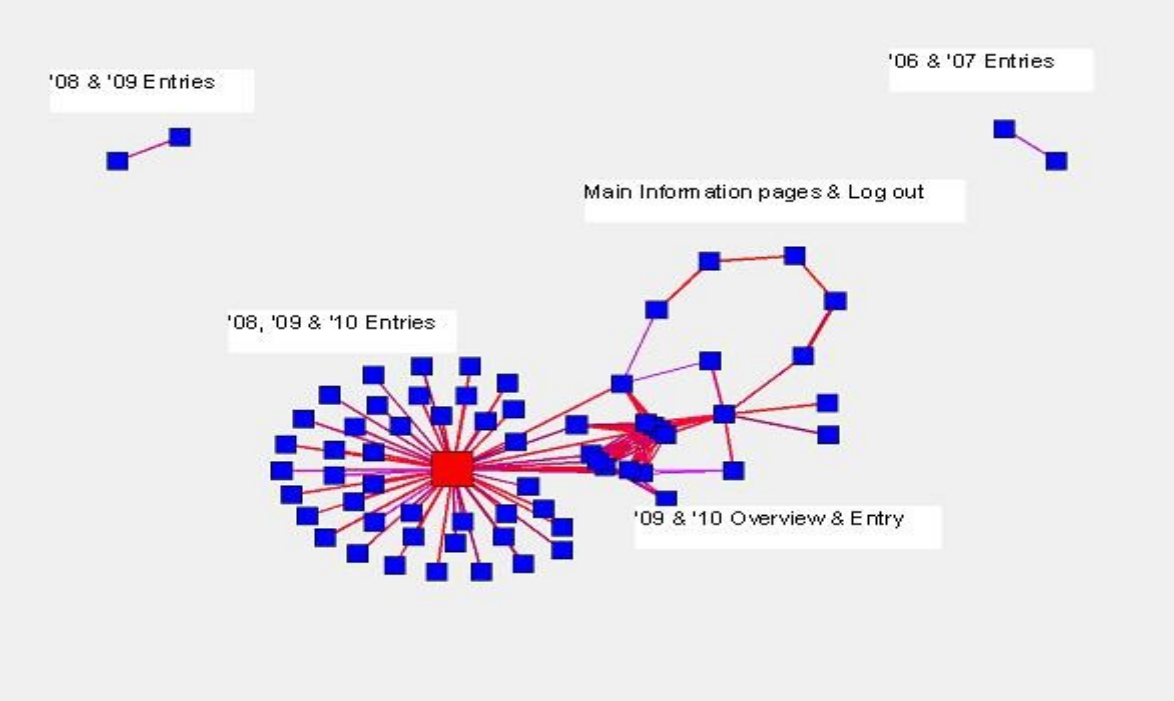

Figure 3: Link graph– PAYE anytime user and web pages visited per session

5.3.2 Distribution analysis

In order to understand the PAYE anytime population a Distribution Analysis of the demographic data was carried out where appropriate. This analysis provided a different view of the base population. It also enabled outliers and anomalies to be identified and explored. As an example Figure 4 shows the Distribution Graph of the Age Groups. This graph clearly shows that the population of interest is typically aged between 25 and 34 years.

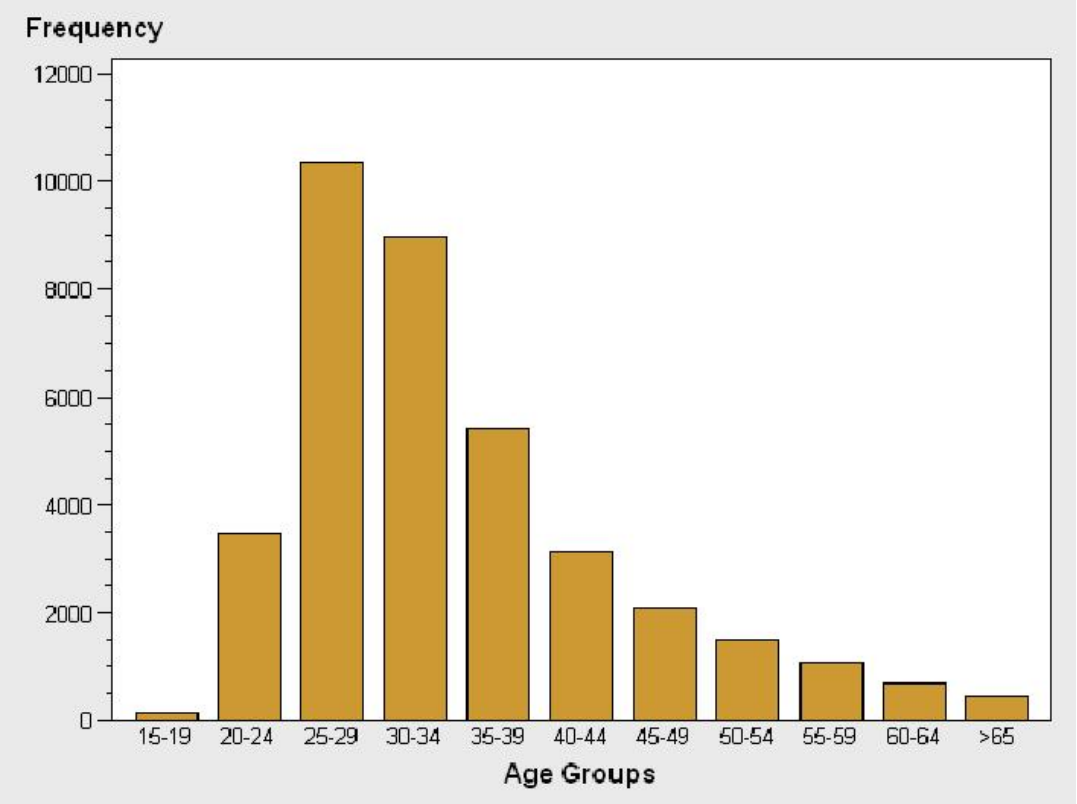

Figure 4: Distribution graph of age groups

Table 1 below, details the range and the percentage of customers in each of the Age Groups. As Figure 4 and Table 1 indicate the PAYE anytime population tend to be younger individuals, in fact over 90% of the users are under 49 years of age.

Table 1: Age groups

Age Groups	**Percentage (%)**	**Cumulative (%)**
15-19	0.36	0.36
20-24	9.31	9.67
25-29	27.81	37.48
30-34	24.13	61.60
35-39	14.58	76.19
40-44	8.39	84.58
45-49	5.57	90.14
50-54	4.04	94.18
55-59	2.84	97.02
60-64	1.80	98.82
>65	1.18	100.00

5.3.3 Frequency analysis

In terms of categorical demographic attributes such as nationality and martial status, Distribution Analysis is not appropriate. However a Frequency Analysis provides data on the content of categorical attributes, for example over 50% of the PAYE anytime user population during this period was Irish. This could be considered as a relatively low value, as the population of the Republic of Ireland is made up of approximately 80% of individuals with a nationality of Irish.

Frequency Analysis can also be used to answer other questions about the dataset, such as how many users have entered the site more than once in the time period. Table 2 shows that over 40% of PAYE anytime users in January 2010 have logged into the website more than once.

Table 2: Frequency count of number of PAYE anytime sessions per customer

No Of Sessions Per Customer	Number Of Customers	% Of Customers
1	23,056	58.14
2	8,578	21.63
3	3,678	9.27
4	1,713	4.32
5	912	2.30
More than 5	1,719	4.33

This frequency table also draws attention to the issue of customers entering the site more than once, to complete their original objective. This applies for example when a PAYE customer needs to confirm the accuracy of historical financial information prior to updating their records or when an additional claim for a tax credit for a previous tax year is being made.

5.3.4 Sequential association analysis and path analysis

Web analytics is the process of analysing the behaviour of a visitor to a website. This paper describes two different methods of web analytics, Sequential Association Analysis and Path Analysis, to uncover patterns of interaction between the PAYE anytime users and the website. Both of these methods of analysis employ transactional level data.
Sequential Association Analysis is an extension of the association analysis described in the previous section to include a time element or sequence. The time element in Sequential Association Analysis is used to identify the order in which the PAYE anytime users visit pages on the website.

Path Analysis is a web analytics function that distills visitor session click data between certain pages and calculates the frequency of unique paths that visitors actually take between those two pages. (Bridgeline Digital, 2010) In effect, Path Analysis reviews the combinations of paths that a visitor can take on a website.

Findings from the path analysis indicated an administrative "Dead End" on the site. Dead ends are pages that a user encounters before logging out. In order to accurately calculate PAYE customers end of year position Revenue must have details of the money earned and the values paid in PAYE during that tax year. PAYE anytime users do not always have these Pay and Tax values to hand when they are requested to enter them on PAYE anytime. PAYE anytime users therefore cannot complete their transaction and the Pay and Tax screen becomes a temporary administrative dead end, until these values are subsequently provided.

Both the Path and Sequential Association Analysis have indicated a strong relationship between two of the summary information pages on PAYE anytime. This could be an indicator to investigate the design and interaction of these pages.

Further findings from these analyses, including the most significant paths identified, were mapped onto the final segments and are discussed in the segment descriptions in the results section that follows.

5.4 Principal Component Analysis (PCA)

PCA is a widely used data dimensionality reduction technique. PCA was applied to the PAYE anytime weblogs as it *'transforms a set of correlated variables into a set of uncorrelated components'* (Woods & Hasted, 2008). *'The goal of PCA is to find a new set of dimensions (attributes) that better captures the variability of the data'* (Ten, Steinbach, Kumar, 2006).

The first attribute of the PCA will contain as much variability as possible and each subsequent variable will contain a lesser amount of the variability in the dataset. For example, in the PAYE anytime dataset 98% of the variability was contained in the first 41 principal components. The table below details the percentage variability captured by the first six components of the Principal Component Analysis.

Table 3: Variability captured by first six principal components

PCA	Variability Captured by PCA (%)	Cumulative Variability Captured by PCA (%)
1	43.01	43.01
2	12.89	55.9
3	10.20	66.1
4	7.08	73.18
5	4.87	78.05
6	3.11	81.16

Principal components can be extracted from the dataset by a correlation or covariance matrix. The covariance matrix was chosen for the PAYE anytime dataset as all of the variables are measured on the same scale, i.e. the number of times each user entered each web page. This method reduced computational time and space required as it avoided the standardisation of variables, which takes place under a correlation matrix.

5.5 Clustering

Generally, segmentation analysis identifies notional groups, in this case, customers, that are similar to each other and yet different to those customers in other groups, i.e. homogeneity within the groups and heterogeneity between the groups. The data mining processes used to identify these groups are unsupervised methods and are broadly grouped under the term Cluster Analysis.

Cluster Analysis methods use a proximity measure of the distance between objects in the dataset. This proximity measure identifies which objects are closer or more distant from one another, thus forming clusters of the objects within the original dataset. It should be noted here that, as part of the iterative cycle of clustering formation, outliers in the dataset were removed and analysed separately.

The cluster analysis was conducted using the attributes identified during the Principal Component Analysis. The demographic data extracted from the warehouse were mapped onto the segments using inferential analysis in order to gain a deeper insight of the customer profile in each of the segments. There are numerous clustering methods available and they are largely distinguished by the proximity measure used to create the clusters. Three methods that employ different proximity measures were used during this segmentation process, Centroid, Average and Ward. The Centroid method was found not to give good separation of the PAYE anytime dataset. The resultant segments were not actionable and so Centroid was discounted as a viable method early on in the process. The Average method was applied to the dataset and 8 clusters were identified. When the demographic data were appended, the segments were found to be clear, well separated and actionable.

When the Ward method, was applied to the dataset, 3 clusters were identified. These clusters were deemed as actionable and well separated. However, when compared with the Average method, various profiles in the dataset were lost in the larger groups of the Ward clusters. These profiles were deemed too important to be ignored and the Average method was chosen as the most appropriate. However, this does not imply that the Ward method will be discounted for future iterations of the PAYE anytime segmentation, larger segments may be useful depending on the requirements of the business users.

6. Results

While the segmentation process is usually termed as unsupervised, ultimately the segments must be actionable for the business purposes identified. The Research & Analytics Branch asked:

- *Who* are the PAYE anytime customers?
- *What* do PAYE anytime customers use PAYE anytime for?

Therefore this segmentation process is semi-supervised as the segments that best answer these questions were chosen from three possible options. The remainder of this section discusses the 8 segments identified using the Average clustering method; these segments are presented in Figure 5.

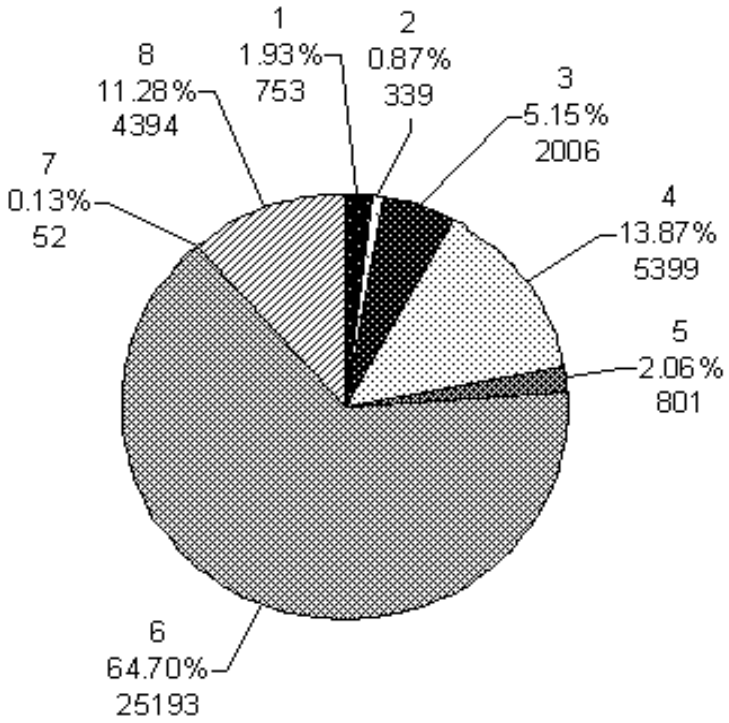

Figure 5: Pie chart of PAYE anytime clusters

Based on the demographic variables, which were mapped onto the weblog based clusters, the following is a narrative of the type of PAYE anytime customers that are generally associated with each of the segments outlined in the table above. The description below shows the general tendencies of the segments and does not describe each member distinctly.

6.1 Segment 1 (1.93%) – newly wed – middle income – repeat visitors

This segment tends to include males aged from 25-34; residing mostly in Dublin and the East of the country and have a tendency to be non-Irish. Recently married, they tend to have children and have at least 1 current employment. Earning between €20,000 and €60,000 per year, they are mostly living in rented accommodation.

This group has used PAYE anytime previously and have received a tax rebate. They have elected to provide Revenue with bank details and to go paper free in relation to correspondence. The path analysis indicates that this segment is likely to navigate between the main application page and the overview screen of the current and most recent tax year. As website users this group has a high hit rate, i.e. they visit a large number of pages, especially those relating to:

- Flat Rate Expenses
- Rent Tax Credit
- Service Charges Relief
- Trade Union Subscriptions
- Annual Balancing Statement Request

6.2 Segment 2 (0.87%) – Irish single fathers

These mid to low-income earners pay between 0 and 10% of their income in PAYE. They are mostly 25 to 34 year old, single Irish males who have just one current employer. They have children and claim One Parent Family Tax Credit. During 2009 these individuals have received a refund of PAYE and are also likely to have ceased employment.

These repeat visitors to PAYE anytime have chosen to correspond electronically with Revenue and have any rebates paid via electronic fund transfer. The path analysis also confirms that this segment has a high propensity to visit the pending requests page. When logged onto PAYE anytime this segment visit a large number of pages. They also tend to visit the following pages for each of the four historic tax years available:

- Health Expenses
- Service Charges Relief
- Annual Balancing Statement Request

6.3 Segment 3 (5.15%) – repeat visitor – non – Irish – single females

While these middle income earners pay up to 20% of their pay in tax, they also claim tax relief's for Health Expenses incurred, Rental Expenses and being in a One Parent Family. These PAYE anytime customers are 25 to 34 year old single and separated European (non-Irish) females. They have received a refund of tax in 2009 and are engaged with the PAYE tax system.

This group have used PAYE anytime prior to January of 2010 and do not visit a large number of pages. However, they apply for tax credits such as Service Charges Relief and request a Balancing Statement for historical tax years. This group has not elected to go paper free in relation to Revenue correspondence but they have chosen to have any rebates paid directly into their bank accounts. An interesting facet of the online behaviour of this segment is that on a per session basis they are likely to evaluate pending transactions and requests. They also tend to actively review profile changes, check their correspondence history and even archive historical items.

6.4 Segment 4 (13.87%) – non – Irish – single females

These middle to high-income earners are 20-34 year old, non-Irish females who are single and separated. They have one employer and pay up to 30% of their pay in PAYE. They are engaged with Revenue as they claim several tax credits including Health Expenses, Rent Tax Credit and Service Charges; this may be due to the volume of state employees in the segment. They have also received a refund of tax in 2009.

These repeat visitors have chosen to submit bank details to Revenue and to go paper free. This segment differs from the other segments in that they visit almost all web pages that are related to tax credit claims; they have visited the following for each historical tax year available:

- Flat Rate Expenses
- Health Expenses
- Rent Tax Credit
- Service Charges Relief
- Trade Union Subscriptions

However, in relation to actual request submission, this segment tends to focus on Health Expenses claims for 2009. They also view Revenue sourced documentation including the latest Tax Credit Certificate and their Annual Balancing Statement.

6.5 Segment 5 (2.06%) – first time – high impression visitors

These tend to be young workers in low paying jobs that pay little or no tax. They are likely to be non-Irish single fathers. However, they claim Health Expenses and Rent Tax Credit and have received a rebate of PAYE in 2009. They are aged between 20 and 34 years of age. They have one current employer and pay 10% or less tax in 2008.

This segment has visited the first time users PAYE anytime web page, and have chosen to take a tour of the site. They are progressive in that they have entered bank details and chosen to go paper free. These first time visitors have a very high impression rate that is; they have visited a large number of web pages. This exploratory style appears to be user behaviour specific to this segment. Members of this segment have visited the following pages for at least 3 out of 4 available historic tax years:

- Blind Persons Tax Credit
- Dependent Relative Tax Credit
- Flat Rate Expenses
- Guide Dog Allowance
- Health Expenses
- Home Carers Tax Credit
- Income Continuance
- One Parent Family Tax Credit
- Rent Tax Credit
- Service Charges Relief
- Trade Union Subscriptions
- Annual Balancing Statement Request

6.6 Segment 6 (64.7%) – older married Irish – low volume users

The largest segment tends to contain the over 35's that are or were married, i.e. married, separated and widowed. There are a large number that do not have a current employer. The annual earnings in this segment is split between those that earn very low wages, less than €10,000 and those that earn high wages i.e. greater than €70,000. They either pay no tax or less than 20%. They have children, claim Mortgage Interest Relief and Retirement Annuity Relief.

This segment reviews the tax years 2010, 2009 and 2008. They tend to change their paper free status and provide bank account details to Revenue. The members of this segment are inclined to log out having viewed their most recent Tax Credit Certificate.

6.7 Segment 7 (0.13%) – low impression – non – Irish males

This segment tends to include low income earning non-Irish males that live in rented accommodation and pay low levels of tax. They are aged between 25-34 and are married or are newly weds. They also tend to have children and have previously received a refund of PAYE tax.

They have used PAYE anytime before and have provided Revenue with bank details. This segment has also chosen to magnify the pages on the site, which might indicate that they suffer from visual disabilities. This is a specific usage segment; they only visit the required web pages and therefore have a low impression rate. Their usage of PAYE anytime indicates that more pages are visited for recent tax years, i.e. they could be regular users that update their records annually and therefore the most recent year has the most visited pages. This segment also has a propensity to view the Rent Tax Credit screen of the current year.

6.8 Segment 8 (11.28%) – low income – first time users

This low-income segment earns less than €20,000 per year and therefore pays little or no tax. They tend to be male, non-Irish including a large number of Asians and are generally married. The age group in this segment is split between the early twenties and the late thirties. They also tend not to have a current employer. They have children and have claimed Rent Tax Credit and Service Charges.

This segment has indicated that they are first time users of the site, they have not chosen to submit bank details, but have chosen to go paper free. They have also elected to take a tour of the site. Segment 8 are even more conservative in terms of web page browsing than segment 7 but less so than 5. This segment tends to view the Tax Credits allocation and income and the band rates pages. They also view Historical Tax Documents, Profile change summary and Tax Credit Certificates. In fact, this segment tends to head directly towards their latest Tax Credit Certificate or to the Rent Tax Credit Claiming page upon log in.

Broadly speaking there appears to be a national and non-national split between the users of PAYE anytime. Within these two broad categories there are 8 distinct segments. Table 2, outlines the segments main features and gives a brief description of the PAYE anytime customers that tend to populate each segment.

Table 4: Segment description

Category	Segment No	Segment Name	Segment Size	% Of Total Population
Non-national	1	Newly Wed – Middle Income – Repeat Visitors	753	1.93%
National	2	Irish Single Fathers	339	0.87%
Non-national	3	Repeat Visitor Non – Irish Single Females	2,006	5.15%
Non-national	4	Non –Irish – Single Female	5,399	13.87%
Non-national	5	First Time – High Impression Visitors	801	2.06%
National	6	Older Married Irish Low Volume Users	25,193	64.70%
Non-national	7	Low Impression Non – Irish Males	52	0.13%
Non-national	8	Low Income First Time Users	4,394	11.28%

As previously stated, the PAYE anytime segmentation is based solely on weblog data, and naturally the online interactions of each segment are different. This is especially evident in relation to Segment 5, exploratory first time users and segment 3, archiving regular users.

However, the path and sequential association analyses have identified some similarities between the segments based solely on the pages that they encounter on the website. The large older married Irish segment, number 6, shares similar features to the smaller low income first time users in segment 8.

Currently these segments show similar characteristics and features in terms of web path usage, with additional data, future segmentation analysis could identify if members of the first time user segment transition into the older married segment. Perhaps the age split between the early twenties and late thirties in segment 8 signifies this.

Figure 6 is a projection of the segments in a two-dimensional space, for illustration purposes only. Each of the segments must be interpreted in a multidimensional space in order to appreciate the actual proximity of the clusters.

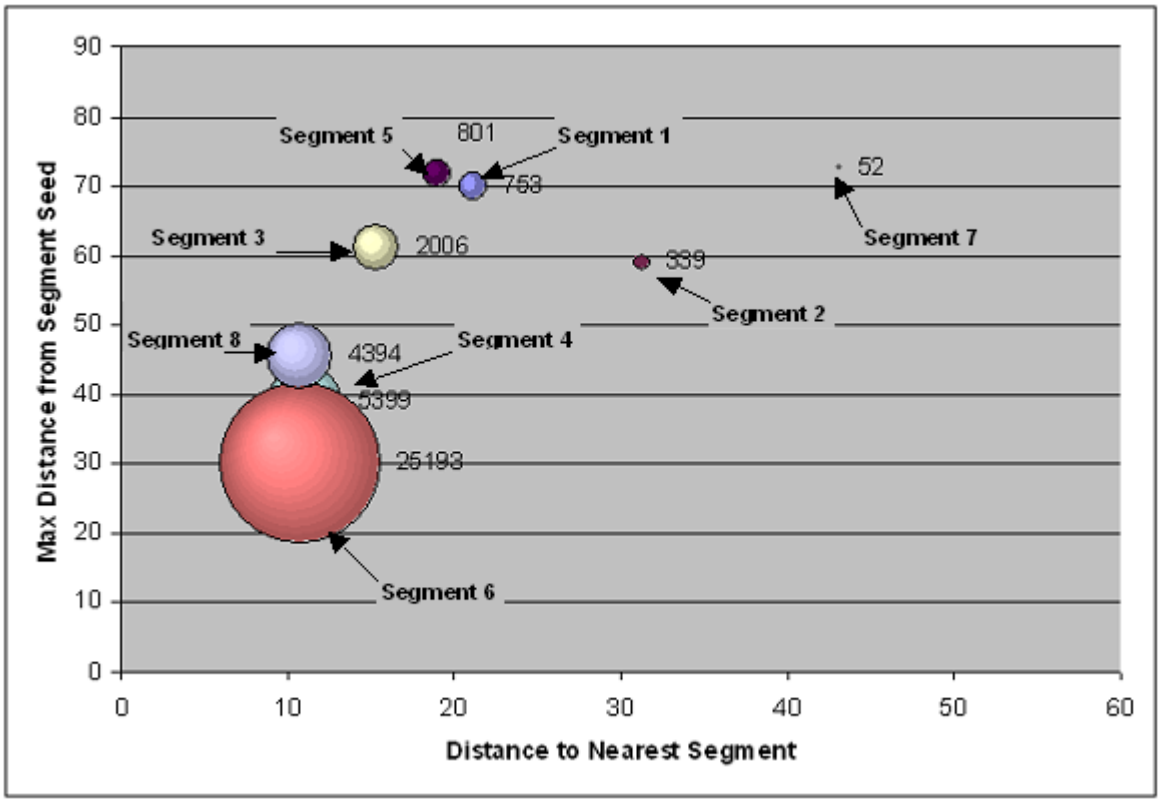

Figure 6: Two-dimensional projection of the PAYE anytime user segments

7. Conclusion

This paper describes a segmentation of PAYE anytime users, which was conducted to inform Revenue's knowledge of the different online behaviours of PAYE customers. The main goal of the project was to answer the following questions:

- *Who* are the PAYE anytime customers?
- *What* do PAYE anytime customers use PAYE anytime for?
- *How* can the PAYE anytime service be improved?

Ultimately the iterative segmentation process identified 8 actionable segments. The segment profiles indicate that there appears to be a national and non-national split between the users of PAYE anytime. Two first time user segments were identified as well as a large older married segment.

7.1 Segment applications

In this section possible Revenue applications of the knowledge gained in the segmentation process are described. A number of these applications are designed to assist the Research & Analytics Branch to identify possibilities for improving the PAYE anytime service.

7.1.1 Targeted marketing

Revenue would like to shift PAYE customer contacts to self-directed e-channels such as PAYE anytime. Mapping the characteristics of the first time user segments onto the general PAYE population could identify potential PAYE anytime users and could result in better-targeted marketing campaigns by the Revenue Online Services marketing team. This could lead to more effective advertising and 'doing more with less'. Also by encouraging more PAYE customers to become PAYE anytime users Revenue can reduce the number of customer contact staff and make direct cost savings.

7.1.2 Streamlining PAYE anytime customer experiences

Segments 5 & 8 reflect two distinct types of first time users of PAYE anytime. Further analysis of the behaviour of the members of these segments could be used to determine why these two segments behave differently, thus perhaps leading to a better first time user experience.

The Association Analysis results showed that the interaction of a PAYE anytime user varies greatly between a month long period and on a session basis. Frequency counts indicate that customers re-enter the site. The customer experience would be improved by allowing customers to complete their PAYE requirements in one visit. A single interaction with the website would require less commitment on the customer's side and perhaps encourage greater use of the PAYE anytime facility.

Further analysis on this issue could be conducted to identify the reasons why customers re-enter the site and perhaps allow the design and implementation of specific improvements to allow customers to complete their task in one visit. For example, by advising customers of the possible historical financial information required to claim tax credits, the customer can have the information prepared prior to entering the website. Information uncovered using qualitative research methods would greatly assist in improving the PAYE anytime customer experience.

7.1.3 Improving Revenue efficiency

A reduction of paper based correspondence and greater utilisation of electronic fund transfers will result in greater cost efficiencies within Revenue. The analysis of the weblog data has identified which type of PAYE anytime customer chooses to correspond with Revenue electronically and those that choose to have any PAYE rebates made directly into their bank accounts. Further analysis on these individuals could be implemented to identify like-minded individuals.

7.1.4 PAYE anytime developments

The segment profiles can be utilised to influence strategic decisions in the further development of PAYE anytime. However the results described in this paper refer to data from a relatively short time window, i.e. January and February 2010. As more data are gathered, further iteration of the PAYE anytime segmentation should reveal new aspects related to seasonal variations of PAYE anytime usage and broaden the profiles of the segments identified. Iterative segmentation can also identify transitional segments.

Path and Sequential Association Analysis help to provide an understanding into PAYE anytime customers' interaction with the website. The provision of additional weblog data will improve the insights gained with these web analytics tools.

Revenue facilitates PAYE customers with visual impairments to test PAYE anytime developments to ensure that the needs of these individuals are met. This testing method could be broadened to include PAYE customers from the various segments identified to ensure that their needs are also catered for.

7.2 Summary

While the weblog information used to conduct this segmentation was taken from a specific time period (January 2010), it has assisted Revenue to understand what PAYE anytime users actually use PAYE anytime for. This clustering has identified 8 segments, which group different types of PAYE anytime customers. This analysis highlighted various actions that can be put in place to enhance the PAYE anytime service, the PAYE anytime customers experience and the e-contact channels in general. The Research and Analytics Branch envisage carrying out further segmentation analyses on an ongoing basis.

References

Battioui, (2006), Data Mining Techniques to Analyze a Library Database, www2.sas.com/proceedings/sugi31/076-31.pdf accessed between January and March 2010

Bridgeline Digital, (2010) Getting the most from Path Analysis http://blog.bridgelinedigital.com/2008/12/getting-the-most-from-path-analysis/ accessed between September and October 2010

Department of Finance (2009) Definition of eGovernment http://ict.gov.ie/docs/eGovernmentDefinition.pdf accessed between January and March 2010

Office of the Revenue Commissioners, (2008) Revenue Statement of Strategy 2008 – 2010, Dublin: Press Office

Revenue.ie (2008), PAYE Customer Survey 2007, Results and Analysis, http://www.revenue.ie/en/about/publications/paye-survey-report-2007.pdf accessed between January and March 2010

Revenue.ie, (2009), Revenue Online Service (ROS) & PAYE anytime, http://www.revenue.ie/en/about/foi/s16/income-tax-capital-gains-tax-and-corporation-tax/part-38/38-06-01.pdf accessed between January and March 2010

Revenue.ie, (2010) Frequently Asked Questions about PAYE anytime, http://www.revenue.ie/en/on-line/paye-anytime.html accessed between January and March 2010

W3C, (2008), Web Content Accessibility Guidelines, http://www.w3.org/ accessed between January and March 2010

Woods T. & Hasted A., (2008) Segmentation of a Customer Database, StatApp Ltd