A Framework to derive Segment Definitions for Web Analytics Suites based on Customer Cross Channel Usage Behavior

Guido Budziak

Department of Mathematics and Computer Science

Eindhoven University of Technology

A thesis submitted for the degree of

Master of Science

November 2008

Abstract

Companies and customers exchange information related to all kinds of issues over different *channels*. Besides traditional channels such as paper mail and the call center, the internet and e-mail have become extremely popular in recent years. Deploying a multi channel strategy offers opportunities to companies to increase the level of service offered to customers. *Internet self service* (services that enable customers to access, retrieve and modify personal information over the internet) is an important concept in this strategy. At the same time, the flow of internet self service web usage that results in a call is considered to be undesired, from the cost and quality of service perspectives. Currently, there is no integrated view on web usage behavior in relation to call center channel usage that allows for detailed analysis.

Web Analytics (WA) is the practice of analyzing visitor behavior within the scope of a single web site by means of tagging each HTML page with a dedicated script. The data collected with a WA implementation can be analyzed in an *web analytics suite* (an OLAP system). A key mechanism in WA is *segmentation* which selects a sub set of the data that meets a condition that is formulated in first order boolean logic. Besides, WA suites allow for tracking cross channel usage. These mechanisms have several limitations that have to be addressed in order to allow for a WA suite to serve as central platform in customer cross channel usage analysis.

In this thesis a framework is presented for a systematic derivation of WA suite segment definitions based on customer cross channel usage behavior. The focus is on customers that make a call after internet self service usage. The framework is elaborated as a *KDD* process. It proposes the set of steps to take in order to process raw data and apply of a rule classifier to obtain a segment definition. Once implemented in a WA suite, this definition will

test clickstream data for segment membership *real-time* in the WA report suite. The framework is an integral part in the process of optimization of internet self service aimed at reducing the volume of incoming internet self service related calls. An experiment conducted using NUON (Dutch energy supplier) data is presented in a case study.

Acknowledgements

Special thanks to my supervisor Dr. M. Pechenizkiy for his dedication and guidance throughout this master project. Many thanks to Bob Nieme, Matthijs Keij and Bartosz Kuzmicki from Adversitement B.V. (Uden, Netherlands). Without their shared vision on innovative technology and their support create a setting that allowed me to do the job this work would not have been produced. Thanks to Eyot Kramer from NUON (Amsterdam, Netherlands) and his colleagues who provided useful feedback and information related to NUON.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Companies and customers exchange information related to all kinds of issues over different channels. Besides traditional channels such as paper mail and the call center, in recent years the internet and e-mail have become extremely popular. Deploying an customer *internet self service environment* (that is part of a multiple channel strategy) offers a company many opportunities to increase customer satisfaction. The most prominent benefit is a 24h-a-day 7-days-a-week on-demand service availability to customers. However, from an analysis point of view, this setting poses many challenges. In this setting, the *information need* of a customer is (likely to be) expressed over multiple channels, in distinct moments of interaction. Each of these interactions contains (a) portion(s) of the information need a customer has.

Example 1 (Cross channel information exchange) *An example of cross channel interaction would be a customer that consults a web site on a Monday morning. Not satisfied with the information obtained, the customer decides to call the company in the afternoon. The call center agent provides a part of the answer and will send an e-mail in which the remaining information is provided. The customer receives the e-mail the next day but reads it on Wednesday.*

In the example above, in the sequence of customer-company interaction related to a particular issue up to three different channels are used: the internet, the telephone and e-mail. The customer's information need is consistent throughout this sequence of events. The interaction time window spans 3 days. Another key aspect is that the

e-mail is sent (with the corresponding registration in a dedicated database) on a particular day, whereas it is read at a different moment in time. The set of events and transactions that are an abstraction of the information need are registered. Clearly, all transactions over these different channels are related. Yet, there are no clear demarcation points with respect to the start and end of the process to resolve a particular issue. Moreover, elements that capture the information need are distributed over several data repositories.

The data from transactions from each channel is stored in a dedicated repository and analyzed in isolation. A *Web Analytics* solution allows for analyzing web site traffic.

Definition 1 (Web Analytics) *Web Analytics (WA) is the objective tracking, collection, measurement, reporting, and analysis of quantitative internet data to optimize websites and web marketing initiatives [12].*

This is within the scope of a single web site. Besides, the data from a call center will be reported on as well. But there is no integrated view on the relation between the web usage behavior in relation to the call center usage. This integrated view is desired, as it allows for evaluating the overall performance of the cross channel strategy. With the increasing popularity and familiarity of the internet, in many cases, this channel is the first source that is consulted by customers to fulfill their information need. Besides, *WA solutions* offer advanced tools to visualize and provide insight into web site traffic.

Definition 2 (WA solution) *A WA solution is the implementation of dedicated script code in HTML pages to facilitate the objective tracking, collection and measurement of clickstream data.*

These elements offer a solid basis to consider web usage behavior as central part in the analysis of customer cross channel usage behavior.

1.1 Business perspective

Internet Self Service Portals (SSP) and Customer Contact Centers (CCC) are part of to the Customer Care Program (CCP) of many companies. One of the objectives many

companies have is to increase the service level offered to customers while decreasing call center volumes[1] in order to reduce costs at the same time. By its very nature, maintaining a call center is expensive since the work is carried out by people.

Definition 3 (Self Service Portal (SSP)) *A Self Service Portal offers electronic services that enables customers to access and retrieve personal information and perform routine tasks (e.g. modifying personal related data) over the internet, without requiring any interaction with a representative of a company. This is also called customer self service (CSS). For customers, self service offers 24 hour-a-day service availability and immediate access to information without having to wait for an e-mail response or a returned telephone call.*

The success of internet self service depends on the quality and quantity of information available and the user friendliness of the environment. Deploying web self service applications benefits a company in a variety of ways. The most prominent motivation is the lower cost, as compared with telephone or email service by a company representative[2]. The loss of jobs as a (direct) result of decreasing the volumes of incoming calls is a debate in itself. In this thesis the focus is on contributing to companies that has the objective to reduce call center volumes to achieve this goal.

Among the set of incoming calls there is a large portion the CCP management considers to be *routine questions & tasks*. This type of question or task does not necessarily require the involvement of a human agent. They deal with information and services that are being offered in the SSP. This allows customers to fulfill their information need themselves. The rationale behind SSP deployment is that a well designed SSP deflects (large numbers of) incoming calls and as such reduces costs.

The basic question that has to be answered to realize an improvement of the quality of the SSP (in order to reduce the number of self service related incoming calls) is: *What makes it that customers that use the SSP subsequently make a self service related*

[1] incoming calls in general, but self service related calls specifically

[2] According to Forrester Research, the cost of the average web self service session is $1, compared to $10 for an e-mail response and $33 for a telephone call.

call?

Remark 1 *For a definition of the terms* ***page view****,* ***web visit*** *(or* ***session****) and* ***visitor*** *that are used throughout this thesis see Appendix A.*

Definition 4 (Self service visit) *A self service visit is a web visit in which services from the SSP are used.*

Despite investments made to improve the quality of SSP and to increase its use, it is recognized that there is a substantial flow from SSP web usage to the call center[3]. This flow is considered to be undesired. Since these calls originate from a (sequence of) web visit(s), this specific browsing behavior might reveal short comings in of the SSP quality.

Definition 5 (Unsuccessful web usage) *Unsuccessful web usage is a (set of) SSP visit(s) of a customer that results in a call.*

Definition 6 (Successful web usage) *Successful web usage is a (set of) SSP visit(s) of a customer that does not result in a call.*

Definition 7 (Self service related call) *A self service related call deals with an issue for which information and services are offered in the SSP.*

Definition 8 (Undesired flow) *Undesired flow is the (sets of) SSP visits that result in self service related incoming calls.*

Remark 2 *All services that are available through the SSP can be completed through the call center.*

Definition 9 (Call center topic) *A call center topic is the label assigned (out of a set of predefined labels) by a call center agent to a call based on its contents.*

Definition 10 (Self service concept) *A self service concept is a page, FAQ answer or form that provides information or a service that relates to a particular call center topic.*

[3]E.g. qualitative analysis of call center data shows that SSP registered users make calls that are considered self service related

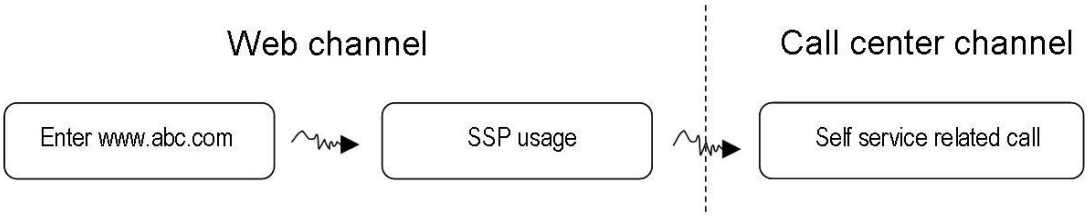


Figure 1.1 – The dotted line indicates the physical separation between both channels. Typically, customers have the objective to fulfill their information need regardless of the channel required. Moreover the *call threshold* (how quick is someone inclined to make a call) varies from individual to individual, from issue to issue. This makes the analysis of web channel performance in relation to call center channel use a hard task.

Each internet self service concept maps to a call center topic. Pages hold a description or explanation where the customer can find particular information. The FAQ section provides answers to questions that are posed in the call center. Forms allow customers to change personal information.

Definition 11 (Topic related visit) *A web visit in which concepts are visited that correspond to a particular call center topic.*

Remark 3 *From this moment on, web usage is evaluated in the context of a specific call center topic. This means that web usage of a customer is considered to be unsuccessful with respect to a particular topic A if there is at least one topic related visit* prior *to an incoming call for this topic A. Web usage that is topic related but does not result in a call for topic A is considered to be successful web usage.*

Example 2 (SSP performance evaluation) *A company decides, from a cost perspective, to move a particular service to the SSP instead of being handled off by the call center. The success could be evaluated as the net result of the cost reduction in the call center (with respect to this particular service) minus the investments made in the web functionality. Yet such an evaluation fails to notice the* actual usage[4] *of the new SSP service. It might be that customers are not aware of the new feature in the web, that its availability is not indicated clear enough, or that the implementation of the service itself raises additional questions as it use is not intuitive.*

[4]Given a particular that is introduced in the SSP, as a consequence calls related to this service are classified as a self service related. If the service is not used then the impression might be that there is a high number of calls because of use of this service. But to what extent customer actually use this service is not addressed.

Table 1.1 shows the possible configurations for a customer with respect to SSP and call center usage for a particular issue given a particular *time window*. Calls that were not preceded by related web usage are not considered. Entries A and B deal with *successful web usage*. There is no call related to the web usage. Entries C through F deal with *unsuccessful web usage*. There is at least one call related to the web usage. All sessions from SSP registered customers where the customer did not log in are left out of consideration. This will not affect the research much. The use of the services in the SSP is subject of analysis. Without logging in, customers do not have access to these services.

		# Self service related calls		
		0	*1*	*>1*
# Self service web visits	*0*	-	-	-
	1	A	C	E
	> 1	B	D	F

Table 1.1 – Customer web & call center channel usage configuration options. The cases in which there was no self service related visit prior to the call are omitted as the flow from the web channel to the call center is subject of analysis.

1.2 Improving: reducing call center volumes

Principle: In order to reduce the volume of incoming self service related calls, the *familiarity* and *quality* of the SSP have to be improved.

Familiarity is a qualitative notion. It refers to the awareness of customers that there is a SSP that offers particular services. This aspect is not further addressed in this thesis. Quality refers to the degree in which the information offered meets the customers' information need. Besides it is concerned with the user friendliness of the environment (e.g. form usability). The services are offered in the web channel whereas their success (or *failure*) is measured in a different channel. This asks for an central platform that offers an integrated view on cross channel usage behavior. Currently, such an integrated view on internet self service in relation to incoming calls is not

available. A *WA suite* has the *potential* to serve as the central platform for reporting on cross channel usage behavior. However, certain limitations have to be addressed properly.

1.3 WA suite

A WA suite[5] is an OLAP system [28, p.131] that presents reports on the usage behavior of a web site. The input for this report suite is the clickstream that is collected from the web site. Currently, a WA suite does not have *machine learning techniques*[6] for *classification*, *clustering*, *frequent sequence* and/or *itemset mining* implemented that automatically extract *knowledge* from the data. Instead, a set of mechanisms is available that allows domain experts for organizing the data by means of labeling and sub setting. Discovering knowledge from the data, such as segment definitions that select all the data from a sub set of the web visitors that satisfy a particular condition with respect to the web usage behavior, is a human task. In the next two subsections a description is given of two key mechanisms that are available in a WA suite to gain analytical insight.

1.3.1 WA cross channel tracking mechanism

WA suites already offer the possibility to track cross channel usage behavior. Below is an example of a so-called *e-mail campaign* that is generally used in practice.

Example 3 *Consider a company that sends news letters by e-mail. E.g. to promote special offers or to refer to interesting items on the web site. The flow from the e-mail channel to the web channel can be brought to light in the following way. Each link in the e-mail to the web site contains a so-called tracking code in the URL query string (string of characters after the question mark in the URL). The link in the e-mail is considered to be the* referring URL. *The WA software is configured to track this code. When someone arrives at the web site after clicking a link in the e-mail, a response is registered.*

This mechanism provides an means to evaluate the effectiveness of the news letter. There are dedicated reports available that show statistics with respect to the flow from

[5] A general term is used. There are several vendors on the market. Each vendor offers a suite that has its own qualities and characteristics. In general, most of these suites have the capabilities that are outlined in this thesis.

[6] For a comprehensive overview of these techniques see [32].

the e-mail channel to the web channel. These reports provide an integrated view on the e-mail channel usage (for this particular news letter) in relation to the web channel usage.

A restriction of this mechanism to track cross channel is that it is only capable of tracking events in related channels that occur *prior* to use of the web channel. Besides, the WA suite can be configured as it is known *in advance* that the links in the news letter are relevant for tracking cross channel usage. On the contrary, the undesired flow distinguishes itself (by definition) by the fact that a (set of) web visit(s) that is followed by a call. Moreover, in many cases there is no *explicit* information disclosed by the customer on the web site that indicates whether she/he is bound to make a call or not. Therefore, direct application of this mechanism to the setting of this thesis is not possible.

However, *if it were possible*, then the WA suite would have all required functionality to report on and analyse cross channel usage behavior. It would allow for answering interesting questions such as *What are most frequently traversed paths in the web site for customers that (are bound to) make a call for topic A?* and *What are the top-5 failed keywords in the internal search for customers that (are bound to) make a call for topic A?*. These are questions that are asked by domain experts that are trying to find out why customers make a call after they visited the web site. However, they are not enabled to analyse this issue in order to find and answer.

1.3.2 WA visitor segmentation mechanism

Another key mechanism available in a WA suite is *visitor segmentation* which is defined as:

Definition 12 *Visitor segmentation: Selection of clickstream data that meets a given condition in order to obtain a homogeneous subset of the data to gain greater analytical insight.*

It allows for analyzing the data from a sub set of the visitors whose web usage is considered to be *homogeneous* with respect to a particular condition. Given this sub set of visitors, the complete report suite functionality is available. WA suites allow for

segment definitions in *first order boolean logic* based on parameters that are *present* in the clickstream. Given a particular segment definition, the clickstream of each visitor is validated against this condition. The result of this validation step is a verdict whether to include a particular clickstream in the segment or not.

In general[7], it is best to have as little segments as possible that divide the data as *pure* and *fine-grain* (based on the visitor usage behavior and interests) as possible. The mechanism of segmentation works properly, yet to define meaningful segments is in many cases a hard task. In general, segment definition complexity and its informative value go hand in hand.

Example 4 (Segmentation of registered users) *The SSP on a web site can be accessed by customers through a username/password combination. It is useful to consider data from these visitors in isolation or to exclude this data from visitors that just look at the so-called* open *content. When a customer logs in a value is assigned in a dedicated parameter that marks this visit as* "customer logged in"*. In the report suite a segment can be created in which the value of this variabele is used in the boolean condition. The segment that includes only data of customers that logged in is defined in a boolean condition "logged in during visit".*

The example above is a matter of *high level* or *coarse-grained* segmentation of registered visitors. It is much more interesting to create more specific segments among customers based on the actual information need. Figure 1.2 gives an overview of the flow from clickstream generation to segment analysis. In general, the content offered on a web site is *heterogeneous* in nature. Depending on his/her information need a customer (but this holds for web site visitors in general) will explore the site and view different content types. A part of this content will be relevant with respect to the information need, another part will be irrelevant. The ratio *relevant : irrelevant* is visitor specific.

By default, a *global segment* consists of the complete set of clickstream data. From an analyst point of view, reports on this global segment show numbers from a heterogenous set of information needs and usage behavior. Again, it is desirable to be able to

[7]Both from a cost perspective and an analysis point of view.

analyse data from visitors with similar information need or usage behavior in isolation. However, formulating a segment definition that captures a particular information need is a hard task. With the exception of particular events (e.g changing contact address) customers do not explicitly express their information need. Besides, given several visitors that have the same information need, they are likely to use the web site differently. Moreover a customer may have multiple topics of interests at the same time. Therefore the actual information need has to be *derived* and this is a challenging task.

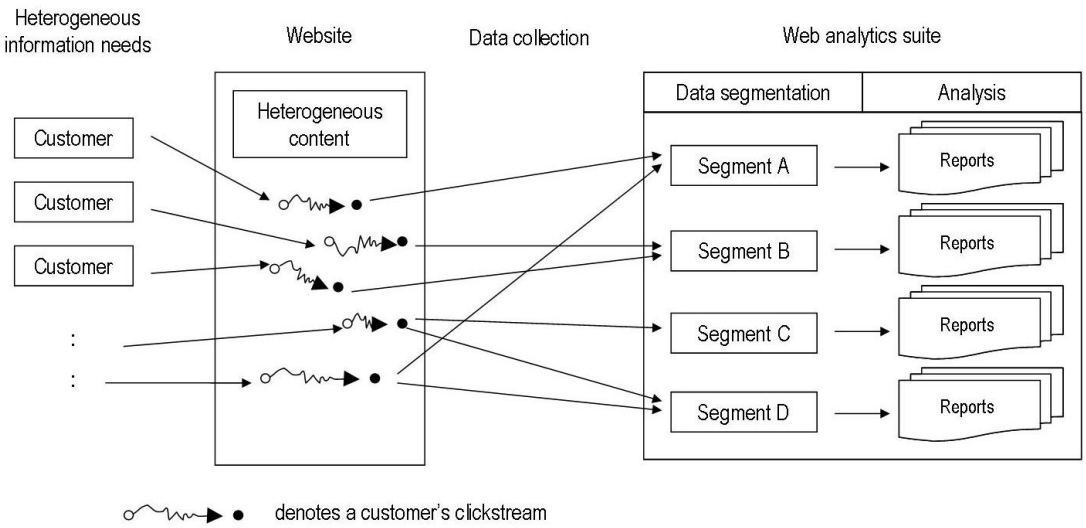


Figure 1.2 – The image shows the flow from the clickstream that is being generated by customers to the reports suite. Segmentation is done automatically and real-time by the WA suite. The benefit of analyzing data in a segment is that the complete range of functionality and report options are available for a sub set of the data.

Example 5 (Multiple segment membership) *Typically, a clickstream can be member of multiple segments simultaneously. Given segment definition* A: "Time spent in visit $\geq$ 1 minute" *and* B: "Number of pages viewed in visit $\geq$ 6"*. Both segments are defined based on a single metric (A: time spent, B: page views). Although the values of these metrics might be correlated, the segment definitions are independent with respect to the metrics used. So the clickstream of a visitor that visits the page for 3 minutes and views 11 pages will be assigned to both segment* A *and segment* B*.*

A more in-depth description of the discipline of WA will be presented in Chapter 2.

1.4 Motivation

A multi channel strategy that is part of the Customer Care Program is deployed by more and more companies. The web channel plays a central role in this strategy. Both from a marketing perspective (*how to keep existing customers satisfied*) and from a cost perspective (*how effective is internet self service in reducing incoming calls compared to the investments made?*) there is a desire to be able to evaluate the overall channel performance. It requires to have an integrated view on the actual channel usage. This allows for analysis based on which appropriate action can be taken. Apart from the perspectives mentioned above two trends related to SSPs are visible:

- Increase of the number of people that use the SSP
- Expansion of the number of services offered in the SSP

This means that in the future the amounts of data that will be collected will go up substantially. This requires a proper means to manage the data and keep the analysis feasible. All these aspects make it interesting to have a method to create an integrated view on channel usage. A WA suite has the potential to act as the central platform to report on and analyse web usage behavior in relation to other channels. However, additional steps are required to utilize this potential.

One of the aspects that has to be addressed to utilize this potential is the collection of additional parameters in the clickstream. These parameters are used to capture elements that address cross channel usage behavior. They can be used in the segment definition. A systematic approach for creating segment definitions is required for several reasons. This is pointed out in the next section.

1.4.1 Need for a systematic approach to derive segment definitions

Below an overview is given of the motivations arranged by subject.

1. Analysis point of view:
    - Currently, segmentation is based on web channel collected parameters in the clickstream data. Elements from other channels are not present in the clickstream and therefore cannot be used in a segment definition. Cross channel

data can be captured in the customer clickstream by means of scripting as a customer has to authenticate in order to access the SSP.

2. Practical point of view:
   - On the fly segment definition is not feasible due to the (in theory) countably infinite number of definitions that are possible. Common sense reasoning, intuition and expert knowledge will allow for excluding a large number of meaningless segment definitions, yet a large number still remains.
   - Visitor web site experience is known to be a very important qualitative aspect. Page load times have to be reduced to the minimum. This requires the number of backend database operations to retrieve data to be reduced to the minimum. This asks for capturing only the required set of values in the clickstream.
   - Additional parameter-value pairs have to be captured in the clickstream. This requires among other things scripting. Web implementations are known for their long time to complete and for failures. Therefore it is important to have clear requirements to be able to recognize difficulties beforehand.
3. Cost perspective:
   - Using segments comes at a price (will differ from vendor to vendor, but it requires database configurations and processing capacity).
   - Storing amounts of data comes at a price
   - Available variables to store data come at a price as well. Collecting and storing only the set of required values is desirable.

1.5 Thesis objective

The development of framework that allows for the derivation of segment definitions for unsuccessful web usage.

The value of the reports in a WA suite depend on the quality of the underlying data set as the *"garbage in garbage out"* principle holds. The framework proposed in this

thesis bridges the gap between the desire for informative segments that capture cross channel usage behavior and the availability of advanced reporting and visualization options available in a WA suite. Currently, segment definition is a task that is guided by domain and expert knowledge and carried out manually. In this thesis the contribution of *data mining* is shown to automatically generate segment definitions.

Definition 13 (Data mining) *Data mining is the process of automatically discovering useful information in large data repositories [28, p.2].*

As mentioned in section 1.3, WA suites do not have traditional machine learning techniques embedded that support the end user in gaining analytical insight. The framework proposed in this thesis uses data mining and machine learning techniques to be able to define more complex segments that will increase the value of the reports in WA suites. The aspects mentioned in the previous section constitute the basis for the framework that is proposed in this thesis. The framework is an integral part in the optimization process to reduce the flow of customer self service web usage that results in a call. In Chapter 2 the position of this framework in the business model will be outlined.

Sections 1.3.1 and 1.3.2 make clear that additional logic has to be implemented on top of the current WA suite functionality to realize an integrated view on web and call center channel usage. Membership of a clickstream to a segment (that is defined for a particular call center topic) that holds unsuccessful web usage has to be *predicted* based on the web usage behavior.

Research on clickstream data analysis and integration with other data sources is still in its infancy. As a first step, the potential of clickstream data that is collected by a WA solution as input for a data mining task has to be examined. Next, it has to be determined how to utilize this potential and which data representation is best for the mining task.

Internet self service is a concept with a clear goal and a clear target group: offer a *personal information environment* to *customers*. Companies hold valuable customer related data that can be used to enrich the clickstream data. All these aspects in combination with the availability of high quality clickstream data form a solid basis for

derivation of segment definitions for the flow from web usage to the call center. To the best of my knowledge there is no work in literature that directly addresses the issues that hold for the setting of this thesis.

1.6 Methodology

The derivation of segment definitions based on the web channel and call center channel usage behavior is *classification* task.

Definition 14 (Classification) . *Classification is the task of learning a* ***target function f*** *that maps each attribute set* $\boldsymbol{x}$ *to one of the predefined class labels* $\boldsymbol{y}$ *[28, p.146].*

In the context of this thesis, the clickstream is the attribute set. The segment definition is the target function (or *classification model*) that acts as a *binary* classifier. Given segment S_A for a particular topic A that has segment definition SD_A which is formulated in first order boolean logic. Each clickstream is validated against condition SD_A. There is no class label assigned, but each clickstream is tested for its *membership* of the segment. Either a clickstream is to be included in the S_A or the clickstream is not to be included in S_A.

The largest part of the work consists of *data preprocessing* [28, p.3] in order to obtain an attribute set that is fit for the classification task. Clickstream, call center and customer related data have to be successively collected, selected, linked and transformed before it is suitable for the classification task. This work is elaborated in detail in Chapters 3 and 4. In the remainder of this section, the classification task itself is described.

There are two main purposes for applying classification. It can be used for *descriptive* modeling. The classification model is used to serve as an explanatory tool to distinguish between objects from different classes. On the other hand it can be used for *predictive modeling*. The classification model predicts a class label of unknown records [28, p.146].

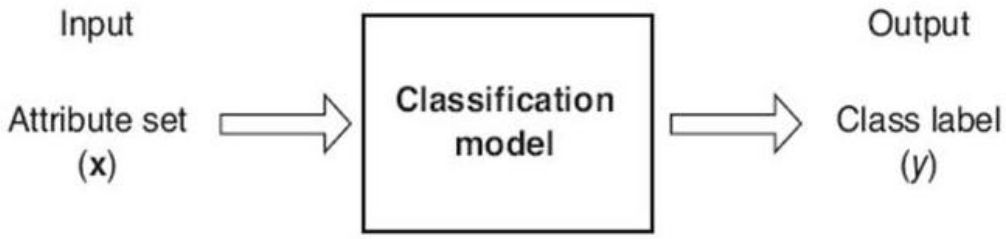


Figure 1.3 – The image shows the general process of classification. In the context of this thesis, the attribute set consists of the set of clickstreams. The segment definition is the classification model that determines for each clickstream whether to include this clickstream in the segment or not. Rather than assigning a class label, a clickstream is tested for its membership of the segment. Figure adopted from [28, p.146].

The classification process in the context of this thesis serves is a form of predictive modeling. The segment definition is implemented in a WA suite. The segment definition for a particular topic predicts (real-time) the membership of each clickstream of the segment. Table 1.2 shows an example of input and output.

Clickstream ID	Features	Member of segment A?
C_1	$\{ f_1, ..., f_N \}$	yes
C_2	$\{ f_1, ..., f_N \}$	no
C_3	$\{ f_1, ..., f_N \}$	yes

Table 1.2 – Given the segment definition for segment A, each clickstream is tested for membership of this segment based on its features.

Figure 1.4 provides a high level overview of the derivation and application of a segment definition for a particular topic. Historic clickstream and call center usage data that form the *training set*. A *rule classifier* serves as the *learning algorithm*. The *model* that is created is transformed into first order boolean logic. Next this segment definition is implemented in the WA suite. Here it will test each generated clickstream for segment membership *real-time*.

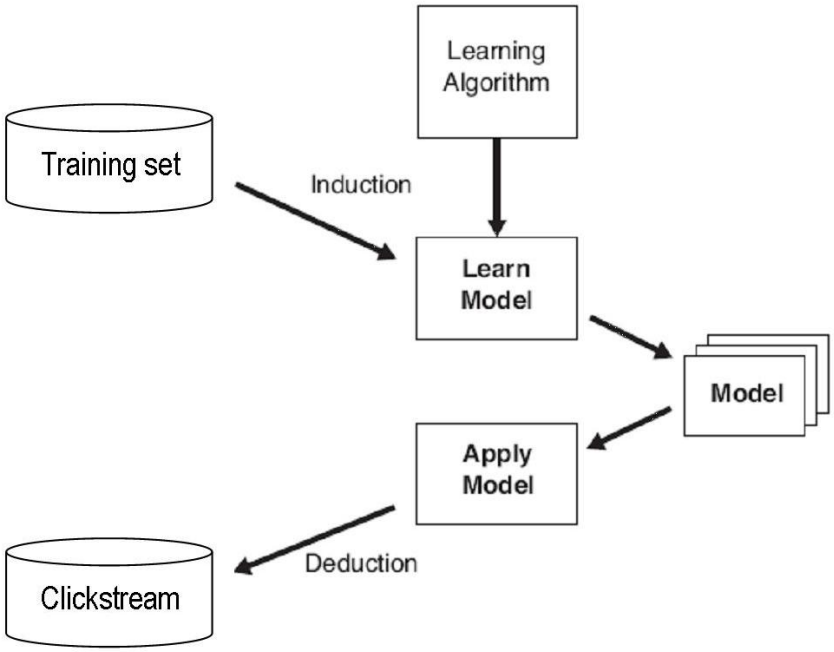


Figure 1.4 – High level overview of the classification task that is carried out in the framework that is proposed.

1.6.1 Framework overview

The elements described above are placed in a framework. Figure 1.5 shows the framework to derive segment definitions. The items above the dotted line are assumed to be present. All the elements below the dotted line up to the *Customer Topic Feature Set (CTFS)* creation deal with data preprocessing. From the WA suite the set of sessionised clickstreams is used. Onwards a *session info* record is created for each customer visit that summarizes the events (that relate to call center topics) that occurred during this visit. Next, this session info table is integrated with incoming call center data. From this a *Customer Topic Event Set (CTES)* is created. A CTES aggregates information from all sessions from a single customer with respect to a particular call center topic. Besides, a division will be made to the set of clickstreams into successful and unsuccessful web usage.

Onwards, this data is enriched with customer data (e.g. demographic and payment related data) in order to obtain a CTFS. A CTFS is a record (defined for each customer

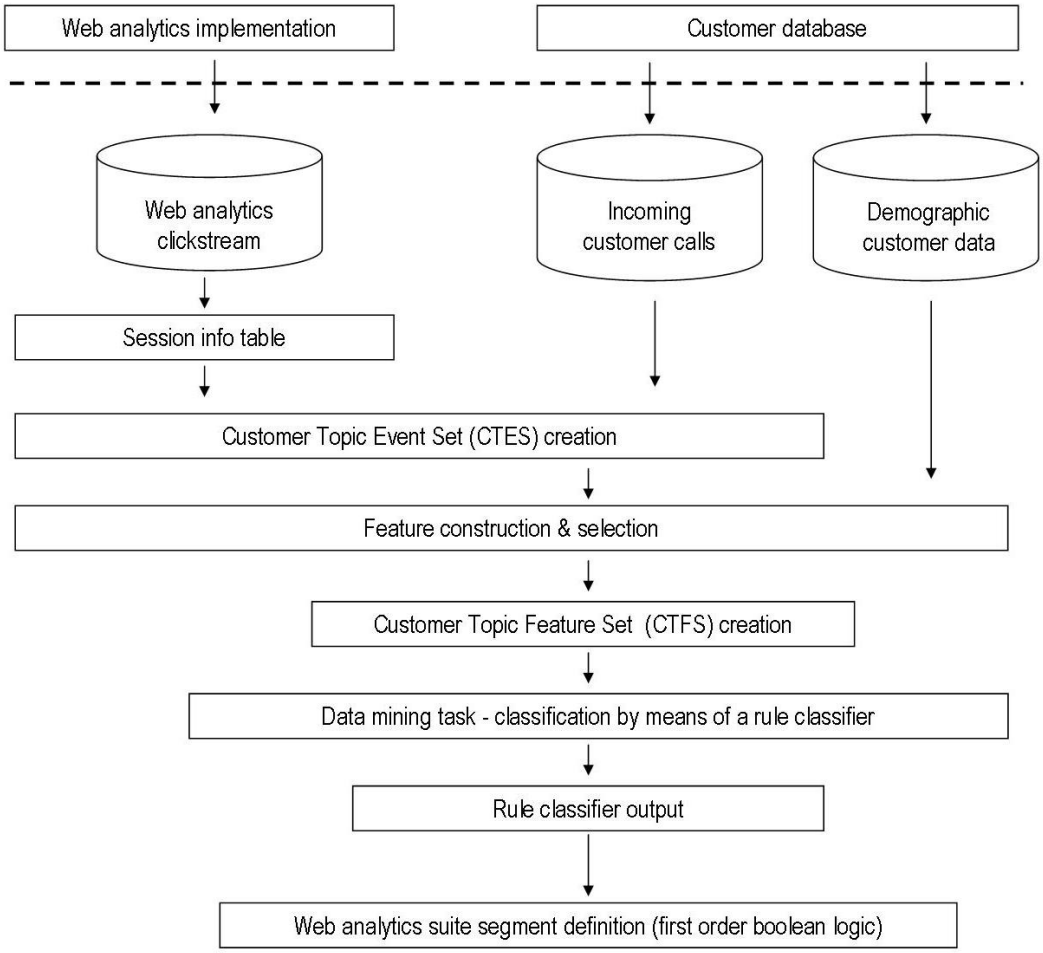


Figure 1.5 – The figure shows the contribution of the framework. Web usage and call center usage are integrated in order to extract the relations. The framework is an integral part of the online business optimization. This model is discussed in Chapter 2.

in the clickstream) that holds all features related to the call center topic that is subject of analysis. The set of CTFSs is input for the data mining task. Here a rule classifier is used to generate rules that discriminate between successful and unsuccessful web usage. The subset of output rules that cover unsuccessful web usage are transformed into first order boolean logic in order to obtain a segment definition that is directly implemented ([8]) in the WA suite.

Additional scripting might be required to capture particular parameter-value pairs that are currently not captured in the clickstream. This task is outside the scope of this framework as well as the actual implementation and evaluation of the segment definition in the clickstream. This will be motivated in the next section.

This work combines traditional data mining techniques (classification by means of a rule classifier, bagging, boosting, dealing with class imbalance by means of sampling and cost sensitive classification) with insights from *WA* (web concept taxonomy definition, clickstream data collection, segmentation). Domain and expert knowledge are a key part in the analysis which will be pointed out in the case study in Chapter 5.

1.7 Results

The output of this thesis is a framework that allows for deriving segment definitions for WA suites based on customer cross channel usage behavior in the context of internet self service. Application of this framework will output a rule set that discriminates between successful and unsuccessful web usage. This rule set can be transformed into a segment definition. An example of the application of the framework is presented in a case study with NUON data.

Besides, a complete data base configuration template is delivered that automates most of the work of the data *selection*, *preprocessing* and *transformation* steps that are described in Chapter 4. It consists of the required table definitions, additional operations (insertions, updates etc), functions and procedures that are optimized for the

[8] Mostly by making use of a wizard.

transformations carried out in these steps. Depending on the specific domain, adjustments have to be made to the concept values that apply.

The contribution of the framework proposed in this thesis are recognized by both Adversitement as well as NUON (see Chapter 2 for description of companies). Both parties cooperate in a strategy to optimize online business and the work proposed in this thesis is useful input for this strategy. Table 1.3 gives an overview of the issues and aspects that came to the fore during the development of this framework and that are addressed in this thesis. The issues are listed in the order in which they are addressed in this thesis.

Aspect or issue	Related section
• Set of definitions related to the context of web and call center usage integration and mining in the context of *Internet self service*	Chapter 1
• Web visit dependency: implications for classification task	Section 3.2
• Potential of WA clickstream data for data mining	Section 3.4
• Limitations of web server log clickstream data for data mining	Section 3.4.1
• Information need extraction from WA data	Section 3.4.1
• Web and call center topic mapping	Section 4.2.3
• Clickstream and call center data relation and integration	Chapter 4
• Customer Topic Event Set definition and creation	Section 4.4.5
• Customer Topic Feature Set definition and creation	Section 4.5
• Web visit data exploration	Section 5.3.3
• (Clickstream) feature construction	Section 5.5
• Derivation of segmentation parameters	Chapter 6

Table 1.3 – Issues and aspects addressed by this framework

The development of the framework is worked out and illustrated in a case study on

the basis of data from a single company. This does not affect the generic aspect of the work in this thesis. The framework addresses all aspects that have to be taken into account and tasks that have to be carried out for settings similar to the one in this thesis. SSPs deal with specific issues that are independent of the domain or context of the web site. Data selection, preprocessing and transformation are tasks that have to be carried out in all cases. Feature construction and selection is a general task as well although some of the features will be domain specific.

1.8 Thesis structure

In Chapter 2 a background is given on Adversitement B.V., the discipline of WA and NUON. In Chapter 3 the principles of this framework are presented and related work is pointed out. Besides the potential of WA clickstream data is examined. The WA clickstream collection architecture is discussed in detail and compared to the traditional approaches in literature. Lastly, the data representation that is input for the mining task is presented. Chapter 4 outlines the steps taken according to the classical *KDD* process to obtain a rule set that can be transformed into a segment definition in the first order boolean logic. Chapter 5 presents a case study of the NUON data and context. The *payment amount change* service which is both a web concept and a call center topic taken as example. Chapter 6 presents an experiment conducted with the NUON data. Chapter 7 presents conclusions and directions for future work.

Chapter 2

Background

This chapter addresses the activities and interests of Adversitement B.V (WA consultancy). Based on their objective to support customers in optimizing online business, the task of the development of this framework was formulated. Besides, the discipline of WA is described in further detail. The place of the framework in the business model to optimize online business is pointed out. Next, a description is given of NUON, a customer of Adversitement. The NUON data is used to work out the framework that is proposed in this thesis. NUON currently has a WA solution that is implemented by Adversitement. One of the objectives Customer Care Center of NUON has is to reduce the flow of customers that visit the SSP and onwards make a self service related call.

2.1 Adversitement B.V.

Adversitement (Uden, Netherlands) is a consultancy that supports companies in online business optimization by implementing dedicated software in websites that allows for analysis and reporting on web visitor behavior. Adversitement has of over 150 companies (from a wide range of market segments) that have their web site implemented with a WA solution[1] and 200+ implementations (see Appendix B). As the web channel is becoming more and more an integral part of business operation there is a demand for appropriate reporting mechanisms. As such Adversitement is working on the extension

[1]There are many different vendors on the market among which Omniture (SiteCatalyst, HBX Analytics), Google (Google Analytics), DoubleClick, NedStat, Webtrends. Although each tool has its own features and characteristics, the data collection architecture is similar. An 1x1 pixel image request with specified parameters in the URL query string is sent to dedicated data collection servers.

of its provision of services by developing and working on standards to integrate the analysis of the web channel with other channels. Adversitement considers a WA suite to be the central platform in cross channel usage behavior in which the web channel is involved. From this need the problem definition of this thesis was formulated.

2.2 Web Analytics (WA)

WA is a young rapidly evolving discipline that is engaged in analyzing and reporting on web visitor behavior. The strive for standards and formats is an ongoing process. The first result is a definition framework drawn up by the *WA Association*[2]. Currently, there is little explicit scientific work on WA. [13] is a scientific publication that addresses the issues involved in implementing and using web metrics to track web site performance. In WA, the web site *objectives* are subdivided into four *categories* as shown in Figure 2.2. In short, *E-commerce* deals with online purchases, *lead generation* with the activity of attracting new customers, *media* deals with advertising and *self service* is part of the customer care program. Each objective has its own characteristics as far as the content and audience types are concerned.

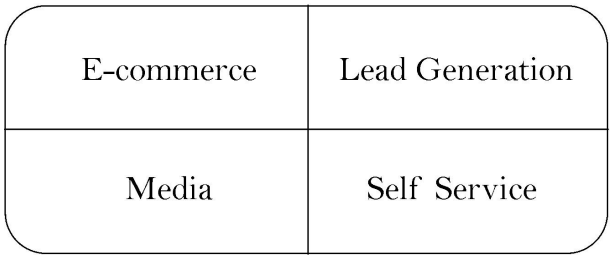


Figure 2.1 – Web site objectives. These objectives are not mutual exclusive. Common practice is web site that provides a combination of e-commerce and self service content. The web shop aims at selling products to (new) customers while the self service portal is available as part of the customer care program.

A key activity in WA is the clickstream data collection. In general the web site is considered to be a set of HTML pages that can be grouped by the type of content.

[2]http://www.webanalyticsassociation.org/

Together the set of pages make up a concept hierarchy. Given this set of pages, a basic WA implementation consists of two steps:

1. Definition of a taxonomy of concepts based on the set of pages
2. Tagging each HTML page with the required script and the concept that is name defined

The taxonomy of concepts is created by assigning a unique name to each page that addresses both its content and its place in this taxonomy. Best practice in WA concept naming is to have the website's sitemap reflected in the model. A sitemap denotes the web site structure and provides web visitors an overview of the web site structure at a single glance. The web site structure is the result of creating a web site that is intuitive to the users. Besides, a 1:1 match between the web site structure and the concept structure in the WA suite will be intuitive to the company end users (data analysts, usability engineers, management) that have to interpret the reports in the WA report suite. This creates a taxonomy of concepts that is *inherently imbalanced*.

Example 6 (WA concept naming strategy) *The URL*
http://www.nuon.nl/veelgesteldevragen/?action=show&objid=1299&from=FAQ
links to an FAQ answer related to relocation in the domain http:www.nuon.nl. This URL contains refers to dynamically generated content in the URL query string - part after the ? symbol. The definition of the page name is /faq/relocation/how do I pass on my relocation?. This string addresses both the content at page level and the position in the concept taxonomy. It consists of three levels - the "/" is used as a level delimiter. The main concept level is FAQ. The second level addresses the topic and the third level expresses (for this concept) the specific question.

The collection of WA clickstream data is achieved by means of the implementation of dedicated script code in all HTML pages (see Appendix C). This script is executed when the page loads. It generates an image request to the data collection servers. These servers collect, parse and provide the data input to the WA suite. The definition of *clickstream* that is used from this moment reads:

Definition 15 (Clickstream) *A clickstream is defined as the sequence of WA image requests that are sent to WA data collection servers.*

Below is the beginning of the URL definition that corresponds to the HTML image request that is sent when the page from the previous example is loaded in a browser. Apart from the concept information (displayed in **bold**), several other parameter-value pairs are passed. Many of these parameter are empty in this example. Each parameter serves a particular purpose.

Example 7 (Image request) *The image request that corresponds to the page load from the previous example is configured as:*
http://GATEWAYNAME/HG?hc=&hb=ACCOUNTNUMBER&cd=1
*&hv=6&**n=/how do I pass on my relocation?&vcon=/faq/relocation***
&tt=auto&ja=y&dt=12&zo=-120&lm=1220350017000&bn=Netscape&ce=y
*&ss=1400*1050&sc=32&sv=16&cy=u&hp=u&ln=nl&vpc=HBX0200u*
&vjs=HBX0201.03u&hec=0&pec=&cmp=&gp=&dcmp=&dcmpe=&dcmpre=&cp=null
&fnl=&seg=&epg=&cv=&gn=&ld=&la=&c1=&c2=&c3=&c4=&customerid=
&ttt=lid,lpos&ra=&pu=&rf=bookmark&pl=Mozilla .....

A characteristic of a WA solution is that it is based on *client-side* data collection. The clickstream to be collected is directly sent from a browser (= *client*) by means of an image request. In Chapter 3 the potential of WA clickstream data as input for a data mining task it is pointed out. Besides, the benefit of WA clickstream data over *server-side* collected clickstream data will be addressed[3] as input for data mining tasks will be addressed.

2.3 WA report suite

Through the WA report suite (often a web interface) the statistics from the collected clickstream are available to the end users. This report suite is an OLAP system [28, p.131]. Reports show numbers and counts for *aggregated* data. This data can be explored in its dimensions through *drilling down* and *rolling up*. There is a large set of predefined reports on a variety of metrics available. Table 2.1 shows a selection of the most important reports that are available in a WA report suite. Onwards, there is a set of custom reports. These reports can be used to report on metrics as desired. Dedicated variables have to be configured in the script tags to collect the appropriate

[3]This all under the assumption that the clickstream data is obtained from a WA solution that is properly implemented.

Category	Metrics
Traffic	Page views, visits, (unique/returning) visitors, clickpaths
Navigation	Link clicks, traversed paths, keywords
Origin	Referrer, search engine (keywords)
System	Browser, resolution, plugins
Commerce	Order funnel, purchase related information

Table 2.1 – WA suite: overview of main report categories.

data to fill these reports. Apart from the definition and implementation of WA solution and visitor segmentation, *testing & validation* is also part of a WA solution.

2.4 NUON

NUON is one of the Dutch energy supplying companies. NUON concentrates upon both the consumer as well as the business market. Quality of service is an important aspect of the company strategy in the present juncture as the market for energy is released for free competition. Unsatisfied customers are free to switch to a different supplier. *MijnNUON* is the Self Service Portal offered to customers (from the consumer market) that allows for payment management and the arrangement of several routine tasks over the internet.

MijnNUON deals with *costs*. This in contrast to e.g. web shops where people consider the purchase of luxury goods (= *expenses*). Moreover, energy is one of human's bare necessities of life. So, regardless of the *quality* (e.g. usability of the user interface), invoices have to be paid no matter what. However, the quality of the information services contribute to a certain extent to the customer contentment.

NUON has an active strategy to have customers use MijnNUON. Currently about 20% of all customers are registered as MijnNUON sign ups. The goal is to increase both the number of registered customers as well as the yearly contact frequency substantially. Currently, MijnNUON can be considered to be a customized information

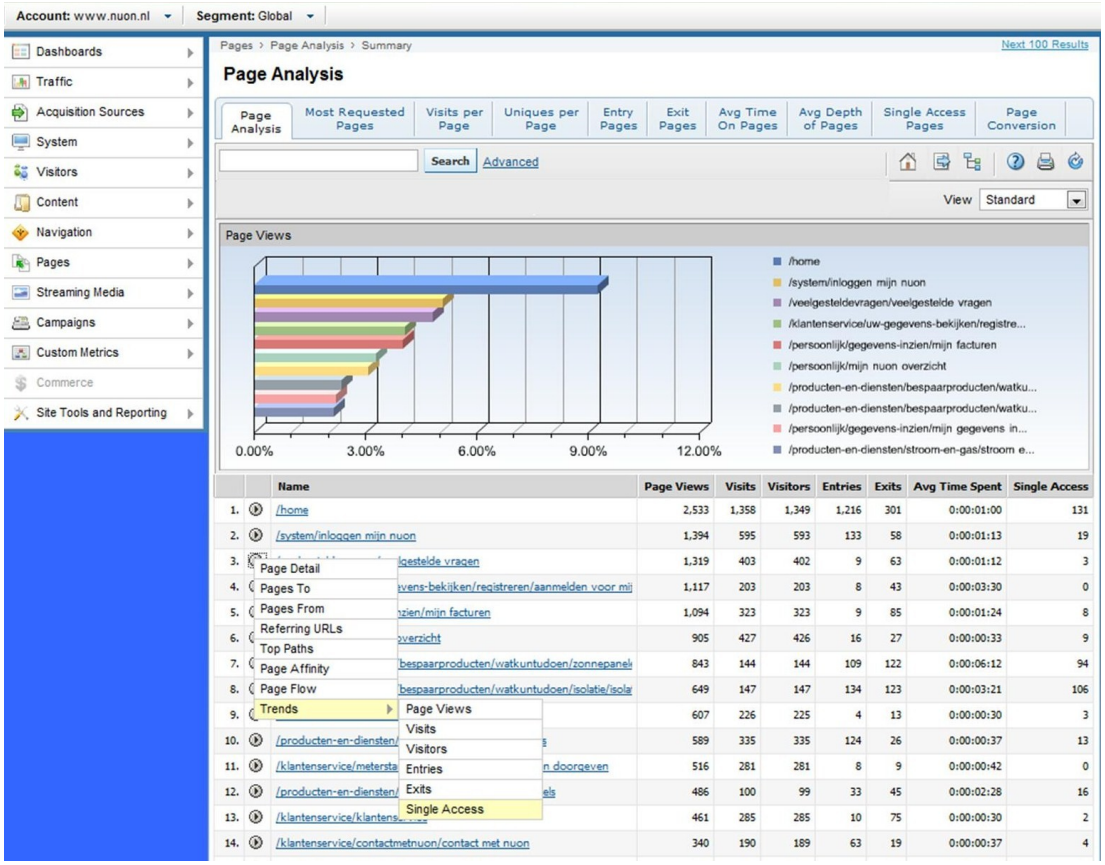


Figure 2.2 – Screenshot from the Omniture HBX Analytics report suite. The screenshot is taken from the NUON report suite.

system on its own. In the future it is the objective to turn this into an adaptive environment in which energy consumption and product offerings are harmonized with each other to offer customers the best deal possible.

As indicated above, NUON uses a WA suite to report on the web visitor behavior. A high level segment definitions is used to select the sub set of visitors (= customers) that access the SSP during a visit. Yet there is a need to be able to define more fine grain segments based on the actual usage behavior to be able to interpret the statistics obtained. Moreover, integration with data from other sources (call center, payment history) is of use as well given the aspect of the *undesired flow* as defined in Chapter 1.

2.5 Business model for online optimization

The business model that is being deployed by Adversitement to optimize online business is shown in figure 2.3. It is a cyclic process that consists of 5 steps. The framework proposed in this thesis affects the *Collection* and *Report* step and it supports the *Analyse* step. Below is an overview of the contribution of the framework to the business model:

- Collect step:
  - The framework outputs a model that discriminates between successful and unsuccessful web usage. The part of the model that covers unsuccessful web usage is transformed into first order boolean logic. Parameters that are present in the formula but currently not captured in the clickstream have to be implemented.
- Report step:
  - The segment definition that is being implemented is allows for reporting on the sub set of the data that is a member of this segment. The full range of reporting functionality is available for this segment.
- Analyse step:
  - The segment definition allows for selecting a sub set of the data that needs to be analyzed in detail as it deals with customers that cause the undesired flow from the web to the call center channel. The result of the application of the framework is that domain and WA experts are enabled to find out why customers have this behavior by using the WA suite.

2.6 Contribution of the framework to daily business operations

The framework contributes to the business operations of both Adversitement and NUON. Adversitement aims at the development of techniques and methods on top of the *off the shelf* WA suites. That is where their added value is in the process of

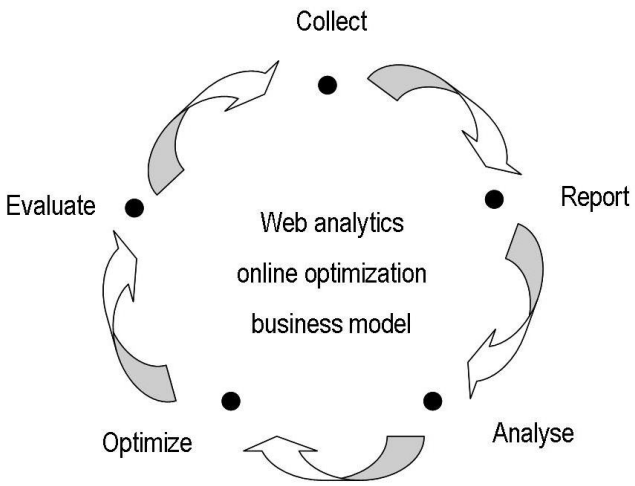


Figure 2.3 – The framework proposed in this thesis fits as a module into business model for online optimization. It affects the *Collect* and *Report* step and it supports the work in the *Analyse* step.

online business optimization. The framework is applicable to the domain of several corporate customers in the customer portfolio of Adversitement. This framework can be deployed in an environment where the web site that has an SSP and that is implemented with a WA solution. For NUON, deploying the output of this framework is a step in the process of optimizing the SSP based on the analysis of an integrated view on customer cross channel usage. It allows for a specific analysis of the relation between customer web usage behavior and related incoming calls.

Chapter 3

Framework principles

In this chapter the principles of the framework are presented. Issues and aspects that were encountered during the development of the framework are described. General lessons learned from literature with respect to *Web Usage Mining* (WUM) as well as related work that was studied are presented. A characteristic all works in WUM share is the strive for a means to derive the information need that underlies the generated clickstream. This holds for this thesis as well. The potential of WA clickstream as input for the mining task that is proposed in this thesis is determined. This is based on a comparison between this type of data and the *server side collected* data that is traditionally used in WUM. Finally, the data representation that is used in this framework is outlined.

3.1 Lessons learned from literature

There are several critical aspects taken into account in order to be successful in the development of the framework. These aspects are recognized based on the observations and lessons learned in literature with respect to the achievements in WUM so far. The next two observation served as a guideline for this thesis:

"One reason for the limited success (of Web Usage Mining) has been a component of Web Usage Mining that is often overlooked the need to ***understand a websites content and structure****."*[7]

WA plays a central role in this thesis. By definition WA is concerned with website content and structure. In this guideline is taken into account directly. But also in the data mining task that abstracts over the context this aspect is taken into account.

"There is no substitute for getting to know your data."[32]

This message is propagated in two ways in this thesis. To understand the data, data *collection* and the *exploration* are considered to be critical aspects. The first aspect is addressed in this Chapter, the second one is pointed out in the case study (Chapter 5).

Another important observation related to clickstream feature construction is taken from [14].

"There is little explicit work on clickstream feature extraction."

Clickstream data mining is clearly in its infancy with respect to this aspect. In this thesis the feature construction and selection task is the result of knowledge from WA best practices in segment definitions and common sense reasoning.

3.2 Addressing information need & session dependency

In general, WUM tasks take a web session as the basic element for analysis. Multiple web sessions from a single person are (implicitly) assumed to be independent. This is mainly due to technical constraints as in many settings it is hard to link multiple sessions from a single user. Besides several privacy issues are of concern. However, as shown by the statistics in section 3.2, this aspect has to be addressed more carefully. This thesis proposes an approach in which the information need is the cornerstone in the analysis, rather than a web visit. *Information need* is considered to be the input for the data mining task. A web visit is considered to be a technical notion that is part of a larger whole. This in contrast to most WUM projects that use web visit as the basic element in the analysis. Qualitative analysis of the data used to develop the framework shows that information need (in the context of this thesis) is:

- expressed over multiple channels
- persistent during a particular time interval
- expressed during different moments in time
- revealed over multiple channels - in each moment of interaction pieces of the puzzle fall into place

These characteristics apply to the context of the thesis, yet they are likely hold to in general. Data from e-commerce sites collected from thousands of purchases made by thousands of distinct people over a year in different market areas show that the number of customers that buy within a single visit has an upper bound of 50%[1]. Correction for this number due to *cookie deletion*[2] would only decrease this number. Similar numbers for self service environment data are not available, yet the main point to be made is that this (implicit) assumption is not substantiated by the facts. Qualititative analysis (as part of the *data exploration*) of the data used in this thesis shows that a substantial part of the customers have multiple visits related to a call.

In this thesis, a variety of data mining techniques is used or considered among which *clustering*, *labeling/classification* of web visits based on information need, *feature construction & selection* and *data sub setting* to have proper input sets for the data mining task. As there is no similar work in the context of this thesis, techniques used are taken from related work and applied to the setting in this thesis. In the next sections an overview is given of the literature studied including an evaluation of the utility in this thesis.

[1]Statistics obtained from the data that is collected by the WA framework that collected the data used in this thesis. Market areas: flight tickets, online phone shops, travel agencies. Period: 01-08-2007 through 31-07-2008.

[2]A cookie is a small file that is stored in the browser by the web host. It allows for storing small amounts of information in order to recognize a browser over multiple visits and to remember visitor preferences.

Session ID	F_1	...	F_k	Conversion
S_1	$value_{F_1,S_1}$	...	$value_{F_k,S_1}$	**no**
S_2	$value_{F_1,S_2}$	...	$value_{F_k,S_2}$	**no**
...	...	...	...	...
S_i	$value_{F_1,S_i}$	...	$value_{F_k,S_i}$	**no**
...	...	...	...	...
S_N	$value_{F_1,S_N}$	...	$value_{F_k,S_N}$	**yes**

Example 8 *Given a clickstream from an e-commerce site that holds N sessions. In the extreme case all N sessions belong to a single person. Only a purchase is made in the Nth visit. The all preceding $N-1$ visits are labeled "no". If this were the input for a classification task that takes web sessions as input to derive discriminating features between* buyers *and* non-buyers*, it would result in incorrect output.*

Figure 3.1 – Example to illustrate that session (in)dependency should be addressed carefully.

3.3 Related work

The approach in [4] corresponds in the basis with the approach in this thesis. *Information need* is an input for the model. Two objectives are formulated in this paper. Besides the *prediction*[3] the objective of *inferring* is formulated: for a particular pattern of surfing an information need (or *goal*) is inferred. Information need is expressed as a $TF \times IDF$ weighted vector of keywords extracted from the documents (= web pages viewed). [1] is a work that uses weighted keyword vectors in a similar way to compute the information need. In this thesis the taxonomy of concepts is used instead of the actual web page content. The motivation will be addressed in section 3.4.

Clickstream mining is considered to be an instance of *Web Usage Mining (WUM)*. WUM itself is an instance of *Web Mining*.

Definition 16 (Web Mining) *Web mining is the application of data mining techniques to extract knowledge from web data.*

[3]Given an information need and some set of pages as starting points, information *scent* is used to predict the expected surfing patterns.

Discipline	Definition
Web Content Mining	Web Content Mining is the process of extracting useful information from the contents of web documents
Web Structure Mining	The process of discovering structure information from the web. This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level
Web Usage Mining	The process of applying data mining techniques to the discovery of usage patterns from web data

Table 3.1 – Web Mining taxonomy overview. Definitions taken from [27]

3.3.1 WUM setting

Several works on WUM are described in literature which are mainly set in an e-commerce context [15] [18]. An e-commerce environment deals with *expenses*. An SSP (mainly) deals with *costs*. This is a fundamental difference with respect to the development purpose and usage between these concepts. There are two other important difference between mining e-commerce data and self service data. In an e-commerce setting, if a conversion occurs all related activities that precede the conversion all occur in the same channel.[4]. In a web shop there is a linear order process that has to be completed[5]. In the setting of this thesis, a call that results from SSP usage might be considered to be the conversion. Unlike the unambiguous conversion point in the web shop (the order confirmation page), the customer decides (last page viewed in the website after which the customer decides to make the call) *him/herself*. This aspect makes it hard to relate the web usage to a call. The strategy to tackle this issue is described in Chapter 4. A further discussion of the difference between a self service and an e-commerce context is described in Appendix F.

[4]Apart from the fact the people may orientate on products outside de website or that people use the website to orientate and to make the purchase in a shop.

[5]Order composition followed by filling in personal and bank account information, followed by shipping info and finally the order confirmation step.

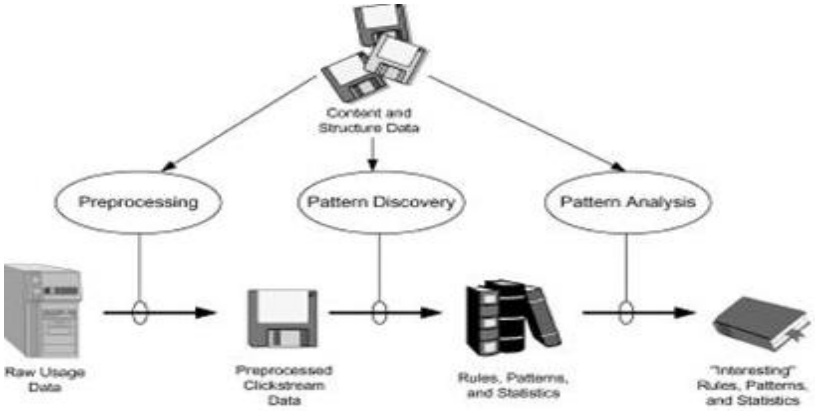


Figure 3.2 – WUM process overview
The figure shows the typical process of Web Usage Mining. Figure adopted from [26].

3.3.2 Website structure & content

In [7] it is argued that the need to understand a websites content and structure is a key factor in WUM and one that is often not taken into account. As pointed out in the previous section, the data used in this thesis is collected as part of a WA solution. Taking the site content and structure into account is an integral part of this implementation. Each page has a unique name that addresses both the content at page view level and the position in the concept taxonomy. The structure allows for selecting different levels of granularity in the concept taxonomy. This is an important aspect as the dimensionality of the data is inversely proportional to the level of detail. Besides, high dimensionality affects the performance of common data mining techniques. This is known in literature as *the curse of dimensionality* [28, p.51].

The set of pages make up a taxonomy or concept taxonomy. Each page name has levels. The full page name denotes the lowest level of granularity. Going up one level of abstraction. The traditional trade off between level of detail and dimensionality is present here. Taking the set of concepts a the lowest level - full page name - gives the most detail. Yet, it results in high cardinality (that implies also high dimensionality) of the data. The optimal balance between level of detail and dimensionality has to be determined in order to have the machine learning techniques used perform best. In [18] steps are proposed to create an *abstract* from a web page name to make it fit as input

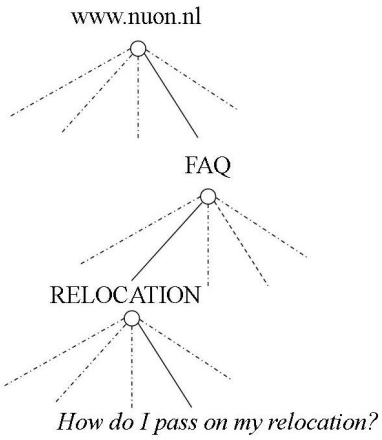


Figure 3.3 – The concept taxonomy is *unbalanced*, that is not all leaf concepts are at the same level of the hierarchy

for session alignment with the Needleman-Wunsch algorithm. As observed in section 2.2, the taxonomy of page names is imbalanced. Computation of session similarity based on (a transformation of) the page name is not used in this thesis. Instead the full page name is used.

3.3.3 Focus points

In [10] a methodology is described that takes *time spent* on web pages as the basic element in the analysis. This measure is considered to be of great importance in other works [14] as well. It is used in the context of this thesis as well. Time spent on page is influenced by many different factors that cannot be brought to light. This degrades the value of this measure. However, it gives an impression of the attention and interest a visitor has for a particular concept. The method described to remove outliers is used. Moreover the concept of *focus point* that is used as a means to address *information need*. A focus point reflects the interest of a web visitor. In this thesis an *event driven* approach is taken in the marking of focus points. A self service concept is considered to be a focus point if it relates to the call topic that is subject of analysis.

3.3.4 Clickstream clustering

Segmentation is to a certain extent related to *cluster analysis* as it deals with the division of data into groups [28, p490]. Clustering is an automated technique.

Definition 17 (Cluster analysis) *Cluster analysis aims at grouping objects such that the objects within a group are similar to one another and different to the objects in other groups.*

Several works proposing different techniques on web session (sequence) clustering have appeared in literature [2] [30] [29] [21] [17][18]. In all works, a session is considered to be the basic element for analysis. For this thesis, experiments with session sequence alignment by means of the application of the Needleman-Wunsch algorithm as proposed in [30][18] were carried out in attempt to group customer sessions based on the information web usage behavior that they share. The algorithm was used to compute a session similarity matrix. This matrix was used as input for the scientific clustering toolkit *CLUTO*[6]. Several clustering algorithms and configurations were used. The experiments did not yield meaningful results. Clusters were heterogeneous with respect to the web usage behavior. Besides, in the publications on sequence alignment mentioned, it is recognized that several thresholds and values in the page similarity computation are dependent on the domain. There is no means yet to evaluate the quality of the values chosen.

A possibility to overcome the issues related session dependence would be to represent all separate but related visits from the same visitor as a sequence of page views and then use it for the algorithm. But the implications of this input data (e.g. repetition of subsequences, number of visits and weighing) are not addressed in literature. As there are many issues to be addressed related to sequence clustering, the automated grouping of sessions was not considered further. A manual method is proposed to label sessions according to its contents.

3.3.5 Features

As mentioned, in [14] it is observed that there is little explicit work on clickstream feature extraction. As far as the format is concerned, there are two basic approaches in clickstream feature representation. A *native varying length* format in which the pages are features (allows for frequent sequence mining) and a *fixed length* format (all instances have the same dimension). In this thesis a fixed length format is used as it

[6] http://www.cs.umn.edu/ karypis/cluto

allows for the integration of web data with data from other sources relatively straightforward.

As mentioned before, time spent at page level is recognized as an important measure of user intention and page relevance [10]. In [10] a *distraction factor* is defined that takes into account the fact that several activities that run in parallel (e.g. chatting, answering telephone) influence the value for time spent at a page. Constructing features at page level that take into account the time spent in the setting of this thesis complicates the feature construction task - the technique is proposed for the level of granularity of individual sessions. Moreover it would result in a data representation of high dimensionality which will negatively affect the data mining task. In [14] an attribute *time* is constructed that expresses the percentile of the visitor's time spent on the website relative to the values of all other visitors.

[5] proposes two features for mining web visits. The *First-N-Last-M* feature defines frequent session subsequences as the first N pages and the last M pages viewed in the session. The *K-Most-Frequent-M* feature uses the top K ranking of measure M, where M is e.g. time spent or number of page views. This approach is used to construct features related to entry and exit content. It can be considered to be a form of frequent subsequence mining which is considered to be not appropriate for the setting of this thesis. See Appendix G.

3.3.6 Cross channel behavior & identification

In general, tracking individuals over multiple sessions is a hard task even if proper techniques are used to create sessions from the data. E.g. people use different computers, different browsers and have different IP addresses. All these aspects make it difficult, if not impossible, to link data generated from a single person over multiple sessions. Besides, privacy issues are at stake as well.

[3] deals with a strategy for analyzing customer behavior over multiple channels. The work is elaborated in the context marketing and product purchases. A key element in the work is that it does not use a unique ID over multiple channels to identify a customer uniquely. Instead it defines so-called *channel independent behavioral attributes*

for the identification customers. A proper means to map customers over multiple channels based on the behavior is to be preferred over a cross channel customer ID.

However, the setting of the thesis does use customer IDs that uniquely identify a customer in both the web and call center channel. This makes the work less complex as these behavioral characteristics will be domain dependent. Deriving these characteristics would be a study in itself. Besides the possibility of linking each session to a customer, this customer ID is used in the call center database which allows for linking web sessions and call center calls. The clickstream data that is used in this thesis is generated by customers that visit the username/password secured section of the website and each of these sessions is tagged with a customer unique ID. As mentioned, each session has a unique session identifier. The sets of customer and session IDs relate to each other in a one-to-many relationship.

Example 9 (Privacy issue) *Customers have to authenticate themselves when they want to access the SSP. A visitor on an e-commerce site is likely to have several visits preceding the visit in which the purchase is made. In the online purchase process the anonymous visitor identifies him/herself as personal related data has to be disclosed in order to complete the transaction. For a complete picture all visits have to be joined together in the analysis for patterns. This approach would be subject of debate as far as privacy is concerned.*

The privacy is respected in the setting of this thesis. An ID is used to link other personal data that is collected (e.g. age). This does not allow for the identification of individuals. There is no personal information being disclosed. Besides, customers explicitly authenticate themselves when using the SSP and this is registered.

3.4 Clickstream data collection

In this section the potential of WA clickstream data for a data mining task is examined[7]. A first step to be taken in any data mining task is to make take appropriate preprocessing steps. In the case of WUM, the data collection phase that precedes this

[7]The assumption is that the WA solution from which the clickstream data is used is implemented properly.

step is a critical part. It determines the quality of the data that is obtained. An important aspect with respect to the clickstream data as input for a mining task is to address *to what extent the user information need is captured by the generated clickstream data.* This aspect relates to data completeness and granularity.

3.4.1 Clickstream data collection framework

Traditionally, data sessionizing and concept taxonomy definition are an essential part of the data preprocessing step [18][26][5]. In most WUM projects the clickstream data that is being used is collected by webservers and available in a server log file format. The issues related to the use of this type of data in WUM are addressed in several papers [9][15][23]. The key observation is that web servers logs are designed for debugging purposes are not for WUM. The insights were a the basis of the WA data collection architecture as we know it nowadays [12].

In this thesis the clickstream data that is used is collected by and processed from a WA tool[8]. This type of collection framework has several benefits over server side data collection. Most importantly, it increases the data quality. Besides, it is (more) clear which data is *not* collected. Figure 3.4 shows the general client-side tracking tier setup of a WA solution. As described in Chapter 2, an image request is used to send the parameter-value pairs to data collection servers.

So far, session identification and concept taxonomy construction from server logs are tasks in the data preprocessing phase that are performed *after* the data is collected. This approach can be considered to be *reconstructive.* Given the issues related to server side data collection, this is a challenging task. In this thesis the data preprocessing step is split up. Sessionizing the data and creating a concept taxonomy are brought forward. The concept taxonomy is *defined* by domain- and WA experts and created *beforehand* (before the data collection takes place). The tool used to collect the clickstream is designed for this purpose and it sessionizes data automatically. This approach can be considered to be *constructive.*

[8]Omniture HBX Analytics

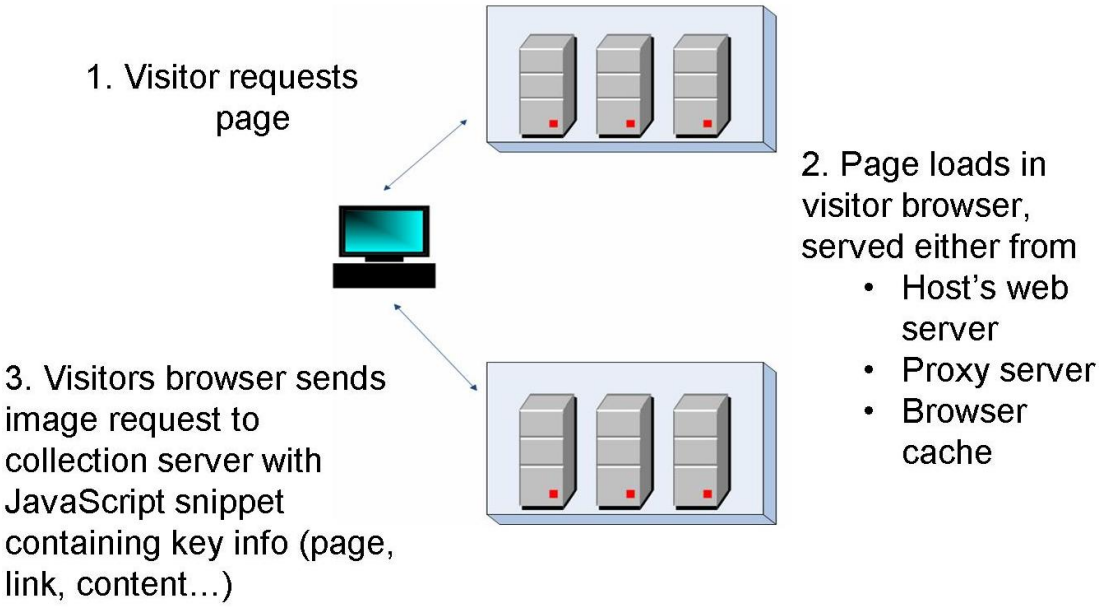


Figure 3.4 – The WA data is collected from the client side by executing JavaScript, regardless of the source that delivers the HTML after a page request by the browser.

In this work an additional step is considered that precedes the data preprocessing task (see Figures 3.5 and 3.6 for a comparison). This is considered to be the *data collection step*. It deals with the definition of a concept taxonomy and data sessionizing and increases data accuracy and completeness compared to server side tracking methods.

The clickstream data that is used in this thesis is available in XML format (with corresponding DTD)[9] (see Appendix D for an example). All requests from first and/or third-party cookie accepting browsers are collected and processed (data from third party cookie rejecting browsers are discarded). It has the following characteristics:

- The data is sessionized with respect to a unique session id and ordered with respect to time. A session is defined as a collection of one or more image requests to the website. For each pair of consecutive requests made by a website visitor holds that the time elapsed between these requests is at most 30 minutes. If inactivity following a request exceeds 30 minute the session expires.
- Each session is equipped with a customer unique id

[9]The format differs from vendor to vendor. Another common format is tab delimited files

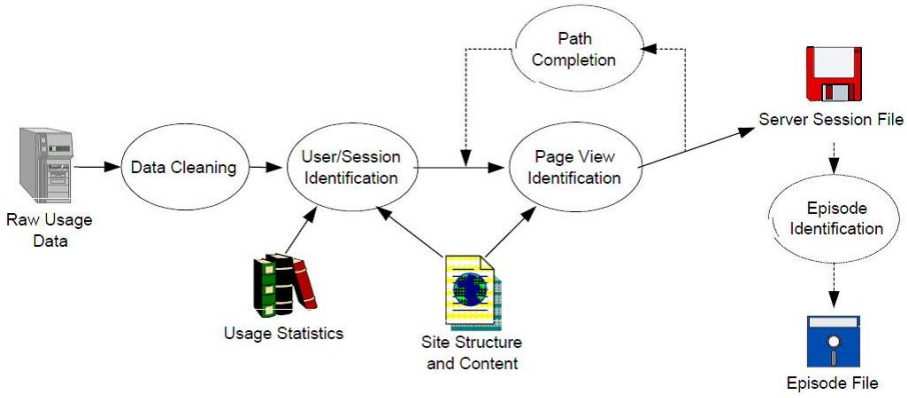


Figure 3.5 – Data *preprocessing* step in case of *server side* collected data use. The sessions have to be *reconstructed*. Figure adopted from the presentation that belongs to [27]. The image itself is taken from http://www.ieee.org.ar/downloads/Srivastava-tut-pres.pdf

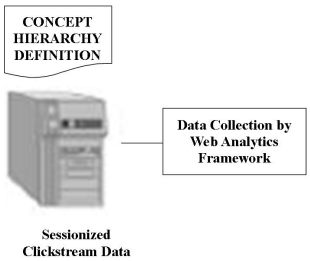


Figure 3.6 – Data *collection* step that precedes the data preprocessing step in a *WA* context. The concept taxonomy is *constructed* and the data is automatically sessionised by the data collection framework. With each page load an image request with parameters that are set based on the concept taxonomy definition is sent to the collection servers.

Several traditional (server side) data collection issues are overcome by the client-side data collection framework. Table 3.2 gives an overview of the most prominent benefits.

3.4.2 (Technical) limitations of the WA data collection architecture

A key element in a WA solution is that it uses *cookies* to store information in a clients browser. It allows for cross session identification. *Cookie deletion* degrades the value of statistics that report on (unique) visitors. The issue does not apply to the setting in this thesis. Customers have to authenticate themselves to access the SSP. This allows for recognizing multiple visits of a single customer over time. A more serious issue

Issue overcome	Benefit
• Cross session & channel identification	This is addressed in section 3.3.6
• Browser caching	Contributes to data completeness. Important since repeatedly visiting a (set of) page(s) is an interesting pattern
• Proxy servers	Contributes to data completeness
• Robot requests	Robot generated requests are excluded
• RIA applications	Relative page views *href="#"* are captured and other web 2.0 features are captured with JavaScript. These constructs are (becoming more) popular since they reduce server load

Table 3.2 – Overview of issues overcome by WA data collection in comparison with web server collected data

is *cookie rejection* that is the result of a browser settings that disallow WA tracking cookies to be installed. However, this is not a technical issue but a privacy issue. Technically, it is possible to associate each page view with the customer ID. This would allow for reconstructing the session. But from the browser setting it is indicated that capturing data from this browser for analysis is not permitted. This has to be respected.

Another issue is the phenomenon of *multi-tab browsing*. Modern browsers allow easily for multiple views at the same moment on a particular internet domain by means of separate tabs in a single view. Qualitative research is conducted to study this behavior. It affects the clickstream collection as it only registers sequences of events made at browser level, not at tab (or browser view) level. In such cases, the clickstream is likely to present sequences of page views that are physically impossible given the website topology. It is a hard task to find out in which instances this occurred. However, it has to be addressed when applying techniques such as *sequential pattern mining* [28, p.429].

3.5 Data quality evaluation

In [28, p.36] several criteria to evaluate *data quality* are defined. In the remainder of this section, five of these criteria that apply to the research context are addressed.

3.5.1 Completeness & granularity

Each page has a unique page name and the set of pages make up an taxonomy of concepts (more to the page naming in section 3.3.6). Each link click is captured according to its link text name (visible link name that is intuitive to the user). HTML features such as relative page requests (e.g. relative page views after a link click $< a\ href = "\#"... > Link < /a >$ and Web 2.0 *RIA* (*Rich Internet Application*) applications are tracked as well.

Well-known examples of Ajax[10]-based RIAs are Google Maps and Google Mail (Gmail). Both use Ajax, allowing a web page to be more interactive without having to refresh the page to serve up new content. This has several benefits. E.g. it offers web designer nice and fancy features while server load is reduced. Typically, overviews in SSP have RIA features such as combo-boxes and toggle fields. Moreover, *third party hosted content* can be tracked in visitors clickstream as well. By definition, this type of content is not registered by the web server that hosts the website as the request is made to a different server.

3.5.2 Noise and artefacts

The most prominent part of noise in clickstream data are artificial clicks generated from robots. The specification of the XML data feed states that robot generated data is not present. First of all, in general robots don't execute JavaScript [12, p.24]. Therefore they don't generate the image requests from the WA implementation. In case of a modified robot that does execute JavaScript, these are not included in the target clickstream data set. The clickstream sessions are identified by means of a customer id and robots do not have such an id.

[10] Asynchronous JavaScript and XML. See Appendix E for overview of AJAX architecture. Additional information available at http://java.sun.com/developer/technical/Articles/J2EEAJAX

Example 10 (RIA tracking) *In figure 3.7 three consecutive "page views" (from left to right) are displayed. The left page "Level 1" is loaded. This is a page request from a web or proxy server. This page contains advanced web 2.0 features like AJAX unfold boxes that do not generate a new server request when clicked. A request for page "/Level 1/Level 2" is executed by clicking the "To Level 2" button in the top right corner. Additional content is being rendered in the middle page by means of AJAX functionality. This even goes one level deeper to page "/Level 1/Level 2/Level 3". The web server will miss page views and inaccurate time spent on page values will be computed from this data. There are several reasons from several points of view to create this kind of implementation. From an IT perspective it is argued that it reduces server load and web usability engineers are advocate on fancy layouts.*

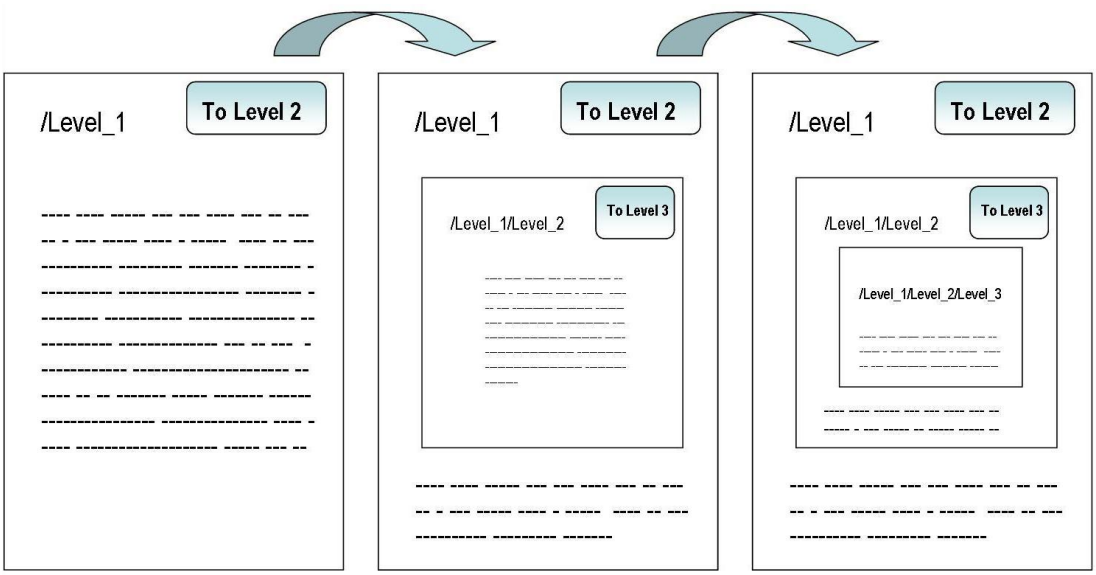


Figure 3.7 – Web 2.0 WA tracking

User action		Web Server Log		WA Suite	
Time	Page request for	Registration	Time Spent	Registration	Time Spent
T_1	"Level 1"	Y	$T_4 - T_3$	Y	$T_2 - T_1$
T_2	"Level 2"	N	-	Y	$T_3 - T_2$
T_3	"Level 3"	N	-	Y	$T_4 - T_3$
T_4	...	...	...	...	...

3.5.3 Missing values

Values might be missing due to (fast) *page hopping* (quickly moving between pages). There is no data collected in cases where the page was left before the dedicated script and image request were loaded. The strategy is to place the WA script code as close to the beginning of an HTML page as possible, since HTML is parsed from top to bottom in linear order. On average, an image request loads within 4 seconds after the page request. Moreover, page hopping does not allow a visitor to view the contents of the page. As we are dealing with payments and personal related data it is assumed that visitors need time to process the information that is displayed on the screen. This in combination with research on the relation between time and the assessment of information [10] make it that the impact of missing values is assumed to be negligible.

3.5.4 Bias

The data in the data feed is collected from *third party cookie* accepting browsers. Data from non-third party cookie accepting browsers is discarded as third party cookies can considered to be malicious. For www.nuon.nl this gives a lower bound of 70% of customer visits collected. This has impact only in case where customers use both non-accepting and accepting cookie browsers. Another possibility would be that a customer would use the same browser during the window of analysis but with different levels of security configured. There is no estimate to give on this particular criterium. This issue can be overcome easily by implementing a WA solution with a so called *first party cookie*. The result of such an implementation is that internet browsers will observe that the image requests are made from the domain that is hosting the website and can be trusted.

3.5.5 Duplicate data

The clickstream data is processed and sessionized. Considering the licence structure (payment according to the amount of data collected), the data can be expected contain little duplicate data.

3.6 Framework point of departure

Based on the literature that was studied and the observations presented in the previous sections, choices are made with respect to the data representation that is considered in this framework to address the information need that underlies the clickstream. The point of departure for the framework is summarized below. Given the clickstream data that is obtained from a WA solution:

- Each unique page name from the clickstream denotes a *concept*. This name addresses both the contents of the concept as well as its position in the concept taxonomy. The full concept name is used (no generalization of concepts).
- A customer unique ID that is available in both the clickstream, call center data and additional data is used to link records
- The set of concepts present in the clickstream define the information need. For each call center topic, the set of web concepts that maps to this topic is considered to be the set of focus points.
- A web visit is considered to be a topic related visit if, given the call center topic that is subject of analysis, one or more topic related concepts were visited.
- Web visits are not considered to be independent. Multiple visits from a customer that deal with the same information need are considered as a whole.
- Information need is the input for the learning algorithm to obtain segment definitions.
- A fixed length data representation for data mining task as a means to overcome the curse of dimensionality. Moreover, it allows for the integration of data from multiple sources relatively straightforward given the common ID that is used.

Based on these principles, the actual data processing steps and mining task are carried out. This process is presented in the next chapter.

Chapter 4

KDD process

The work in this thesis is a process of extracting useful information in databases. This process is generally known as *Knowledge Discovery in Databases (KDD)* [8]. The steps to be taken in this process are depicted in figure 4.1. The input consists of customer clickstream data, incoming call calls and additional customer related data. A set of database transformation steps yields a *Customer Topic Feature Set (CTFS)* that is fit for the data mining task. Here a *rule classifier* will be used to create a rule set that discriminates between successful and unsuccessful web usage given a particular call center topic. A sub set of this rule set will be transformed into a segment definition. In the remainder of this chapter, first the approach to carry out the process is outlined. Next, terms and definitions are presented. Onwards, each step in the KDD process will is elaborated in a dedicated section.

4.1 Approach

A *relational database* is the workhouse for all data storage, retrieval and modification operations in steps 1 through 3 of the KDD process. It serves best the need for a means to integrate data from multiple sources (web clickstream, call center, demographic information) and formats (XML and relational data) in a flexible way. Moreover, open source data mining tools such as the machine learning software WEKA[1] support direct access to relational databases. The input for the analysis in this thesis is the clickstream data and the set of incoming customer calls. Depending on the WA solution,

[1]URL: www.cs.waikato.ac.nz/ ml/weka

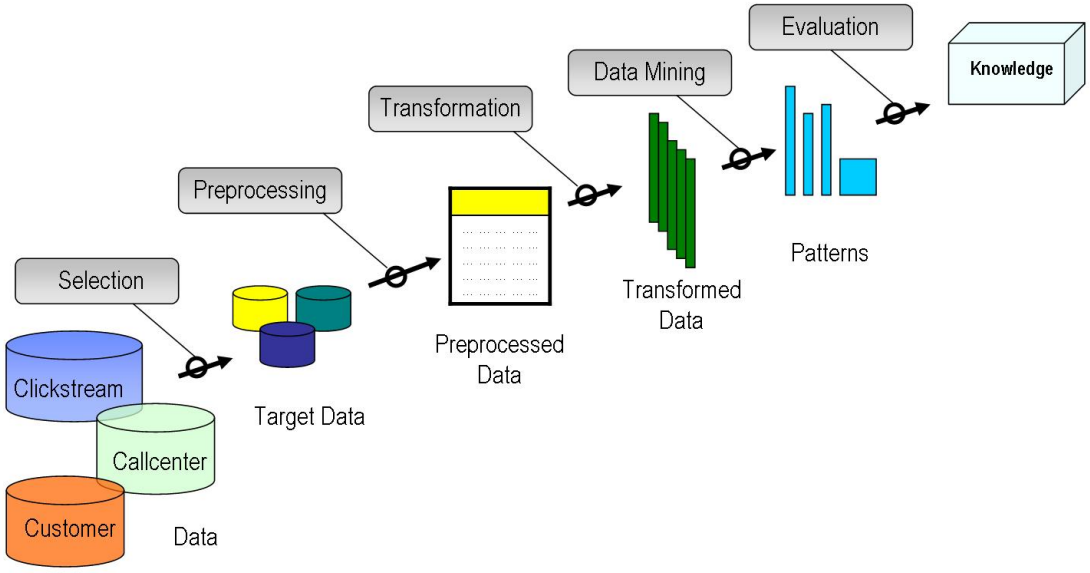


Figure 4.1 – KDD process overview.

the format in which the clickstream data is available may vary[2]. However, the steps to be taken hold in general.

There are two possible strategies to carry out the data *selection*, *preprocessing* and *transformation* steps.

1. A single topic is selected in advance. Given this topic, steps 1 though 3 are applied exclusively to the sub set of the data that corresponds to this topic. A single CTFS is obtained. A benefit is that there are no unnecessary steps carried out for topics that will not be analyzed.

2. Steps 1 through 3 are carried out for a set of self service related call center topics in order to obtain a CTFS table for each topic. Afterwards, one or more topics are selected for analysis. Each topic is mined further in isolation. A benefit

[2]The format that is used in the description is XML. This is the format in which the data is available for NUON for which a case study is presented in Chapter 5. Besides, many customers of Adversitement have an Omniture HBX Analytics implementation and their clickstream data would be available in XML as well. Another common format are tab delimited files.

of this approach is that a single iteration of the selection, preprocessing and transformation steps. From here, one can focus on the mining task.

In the elaboration of the steps of the KDD process in this thesis, the second approach is used.

4.2 Data: terms, definitions and representation

In this section a formal definition is presented of the clickstream and call center data. A general definition of the customer data is omitted. This type of data will differ from company to company and will depend on the topic that is subject of analysis.

4.2.1 Clickstream representation

A website is defined as a fine set of static and/or dynamic web pages P. The set of pages constitutes a concept taxonomy. Each page $p \in P$ is a unique string that expresses both the content of the page and the position in the concept taxonomy. The page naming strategy is the result of a WA implementation process. The set of pages makes up an taxonomy. $|P|$ denotes the cardinality of the set. V denotes the set of clickstream data of customers. A clickstream of customer c in session $s \in \mathbb{N}^+$, $x^{c,s} \in V$ is modeled as a sequence of page views $x^{c,s} =< x_1^{c,s}, x_2^{c,s}, ..., x_y^{c,s}, ..., x_{N^s}^{c,s} >$ where $x_y^{c,s}$ denotes the yth page view in time of customer c in session s and N^s denotes the total number of pages viewed in session s.

This notation is similar to the one proposed in [14]. However, the assumption that information need is persistent over multiple sessions but is also fixed to a particular item is not. This constraint is relaxed in this work. So for each session the topic of interest is determined according to particular criteria. Each session is classified according to its topics. This will be addressed in the next section. $x_y^{c,s}$ is a vector defined as $< si_y^{c,s}, r_y^{c,s}, p_y^{c,s}, dt_y^{c,s}, ts_y^{c,s} >$ with

- $si_y^{c,s}$ Unique session identifier of session s
- $r_y^{c,s}$ The order - y^{th} position - of the concept view in session s
- $p_y^{c,s}$ Name of the concept that was viewed
- $dt_y^{c,s}$ Timestamp (browser time zone) at which the concept was requested
- $ts_y^{c,s}$ Amount of time spent (in seconds) on the concept

Remark 4 *The HTTP protocol is stateless and therefore the concept of a session does not exist at protocol level. There is no connection among successive page requests made from the same browser [15] [9] [23]. The use of cookies makes it possible to compute the values for* $ts_y^{c,s}$, $(1 \leq y < N^s)$ *for all page views in the visit except for the last one,* $p_{N^s}^{c,s}$. *The value for* $ts_{N^s}^{c,s}$ *is set to unknown. The amount of time spent in seconds for all other pages is computed by subtracting the timestamp of two consecutive page views according to:*

$$ts_y^{c,s} = dt_{y+1}^{c,s} - dt_y^{c,s},\ (1 \leq y < N^s)$$

Similarly as in [14] the sequence of pages viewed by customer c in session s is represented by $p^{c,s} =< p_1^{c,s}, p_2^{c,s}, ... >$. Analogically sequences $l_y^{c,s}$, $dt_y^{c,s}$ and $ts_y^{c,s}$.

4.2.2 Call center data representation

Each call is associated with a number of attributes. Each call is represented in a vector format that is described below. C denotes the set of incoming calls that were preceded by at least one web visit. An incoming call ic made by customer c with topic t is modeled as a vector and defined as $ic^{c,t} =< c, dt, rdt, t >$ with

- c Customer unique id
- dt Timestamp of call start
- rd Timestamp of SSP registration date
- t Call center topic

4.2.3 Mapping

As stated in Chapter 1, a website self concept maps to a call center topic. The set of call center topics is denoted by $T = \{ t_1, t_2, \ldots , t_N \}$ where each element denotes a call center topic. The set of self service concepts is denoted as $C = \{ c_1, c_2, \ldots , c_M \}$. There are three concept types. C_{page} denotes a page concept. C_{FAQ} denotes an FAQ concept. C_{Form} denotes a form concept. The mapping of a concept $c_i \in C$ to a topic $t_j \in T$ is a *many-to-one* relationship. Each self service concept maps to one topic and each topic holds one or more concepts. A *Topic Definition Table* is created to list these mappings. See table 4.1 for an example.

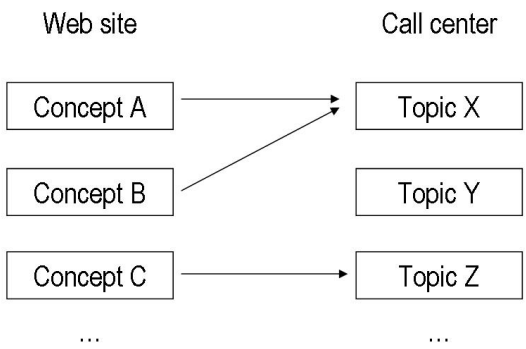


Figure 4.2 – Web concept to call center topic mapping.

Call center topic	Page concepts	FAQ concepts	Form concepts
A	$\{P_1,P_{14},P_{734}\}$	$\{Q_1,Q_{14},Q_{544}\}$	-
B	$\{P_23,P_{24}\}$	$\{Q_5,Q_{16},Q_6\}$	$\{F_1,F_2,F_3\}$
C	$\{P_93,P_{198},P_{534}\}$	$\{Q_1,Q_{18},Q_{734}\}$	-

Table 4.1 – For each call center topic the web concepts that belong to this topic are listed.

4.2.4 Topic related visit

Given a topic $t_i \in T$ that is subject of analysis and a customer clickstream $x^{c,s}$ that holds a sequence of concepts. A visit s is considered to be topic related if one or more concepts are present in visit s that map to t_i.

4.3 Step 1: Data selection

A proper sub set selection of the data in advance reduces data base computation times and saves additional efforts to remove items later on. E.g. the complete clickstream data is delivered which includes non-customer data. By selecting just the sub set of customer data for customers that used the SSP, data base performance is improved since large amounts of data that have to be discarded will be left out.

4.3.1 Clickstream data selection

The web clickstream data is stored in an XML format that is equipped with a DTD (see Appendix D for an excerpt of an XML feed). It is collected by the WA framework that is described in section 3.4. Each XML feed holds *all* the sessionized clickstream data of an hour. All 24 files of a single day are stored in a .zip file. The selection step of this data set consists of:

1. The required XML elements are selected with an XSL style sheet and transformed into comma delimited sequences
2. Export of these sequences to a relational database

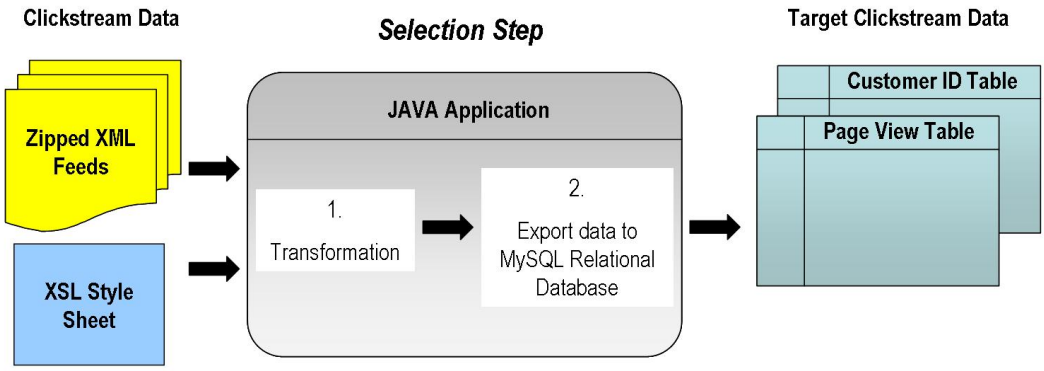


Figure 4.3 – The figure depicts the data selection step for the web clickstream data.

In Appendix H a brief description is given of the tool that is developed (as preliminary work in a separate course) to export clickstream data from the XML format to a MySQL data base.

Table	Record format
Customer ID	(Customer_ID, Session_ID))
Page View	(Session_ID, Timestamp, Concept name)

Table 4.2 – Table entry format.

The entries to be created are as follows are shown in table 4.2. Both table schemas are in *first normalized form* [24, p.257]. The records are ordered by session ID. For the *Page View* table holds that each group of records that have the same session id are ordered by timestamp in ascending order. The (Customer_ID, Session_ID) tuple is the *primary key* of the *Customer ID* table. The customer ID serves as the connection between the clickstream and call center data. The session ID serves as the identifier in the clickstream data. It is a unique identifier assigned to each web visit (see Appendix A for a definition).

4.3.2 Call center data selection

The call center data is supposed to be available in first normal form. Each record holds a customer ID, timestamp and a call topic. Besides, a record may hold additional information that classifies the call. Depending on the way in which call transactions are registered and classified, the format may vary. It is common to have a classification of each transaction that is built up hierarchically. This allows for reporting at different levels of granularity. The lowest level is most likely to address the call center topic and this level should be selected.

4.4 Step 2: Data preprocessing

In order to be able to work in a flexible and efficient way with the data, additional tables have to be created based on the *Page View* table. This is a process that consists of table definitions, insertions and updates. The image below shows the tables that are obtained. The modified tables are listed in Appendix I.

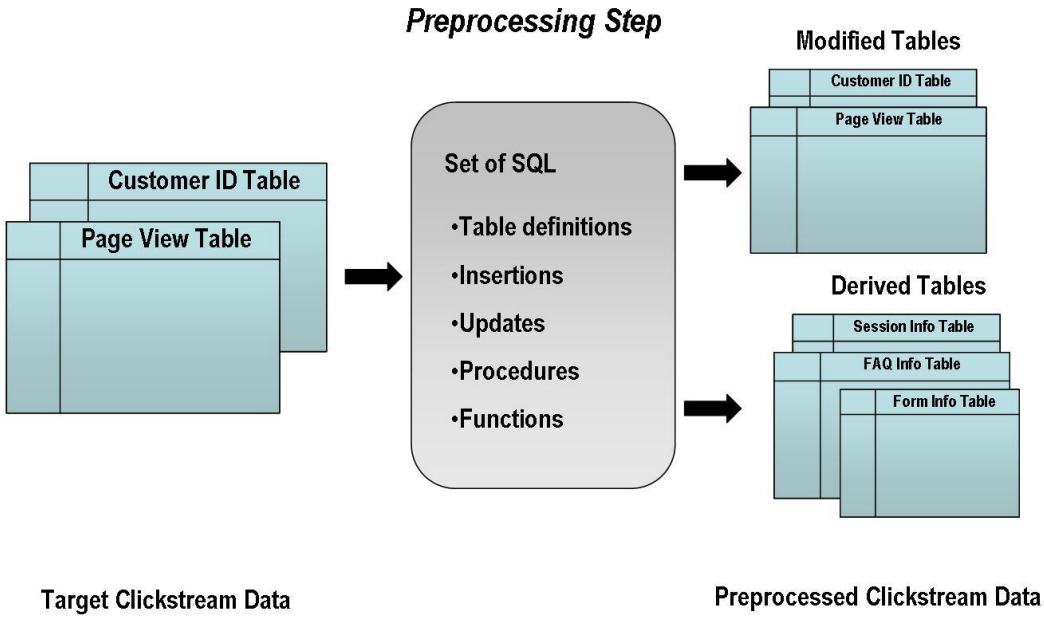


Figure 4.4 – The Figure depicts the data preprocessing step for the web clickstream data.

4.4.1 Modified tables

The *Page View* and *Customer ID* tables are inspected for noisy data records. This includes inaccurate concept names as a result from testing. A general approach is hard to define. Besides a relative ordering of page views in a session is added and the time spent on page is calculated.

4.4.2 Derived tables

Table 4.3 shows example entries for the *FAQ info* table that is constructed from the *Page View* table. A *Form* table is created in a similar way. These tables are created to provide a flexibility with respect to selecting data from these concepts. It introduces a certain amount of data redundancy. This is a trade-off flexibility and data storage.

4.4.2.1 FAQ info table

All concepts in the taxonomy that correspond to an FAQ answer have a discriminating prefix. The remainder of the name addresses the subject of the answer. Besides, each FAQ answer maps uniquely to a call center topic. For each session, a summary table

is created related to the FAQ concept. Column **Answer** holds the concept name and **Topic** is the corresponding call center topic.

Session ID	Answer	Topic	Times Viewed
S1	/faq/answer m	Topic X	2
S1	/faq/answer g	Topic Z	1
S4	/faq/answer m	Topic Y	1

Table 4.3 – *FAQ info* table example

4.4.2.2 Form info table

All concepts in the taxonomy that correspond to an form step have a discriminating prefix. The remainder of the name addresses its intended use. Each form maps uniquely to a call center topic. For each session, a summary table is created related to the form concept. Column **Answer** holds the concept name and **Topic** is the corresponding call center topic.

Session ID	Form	Topic	# Steps	Steps completed
S1	/form/topic x/step 1	X	3	1
S1	/form/topic y/step 3	Y	4	4
S2	/form/topic z/step 1	Z	3	3

Table 4.4 – *Form info* table example

4.4.2.3 Session info table

Based on the tables obtained by applying the steps in figure 4.4, for each session a session info record can be created that characterizes this session. The strategy to create a record is as follows. Website events (page views, form fills) are mapped to a self service topic. Given a session, for each self service topic it is determined whether this session holds events that relate to this topic - see Figure 4.5.

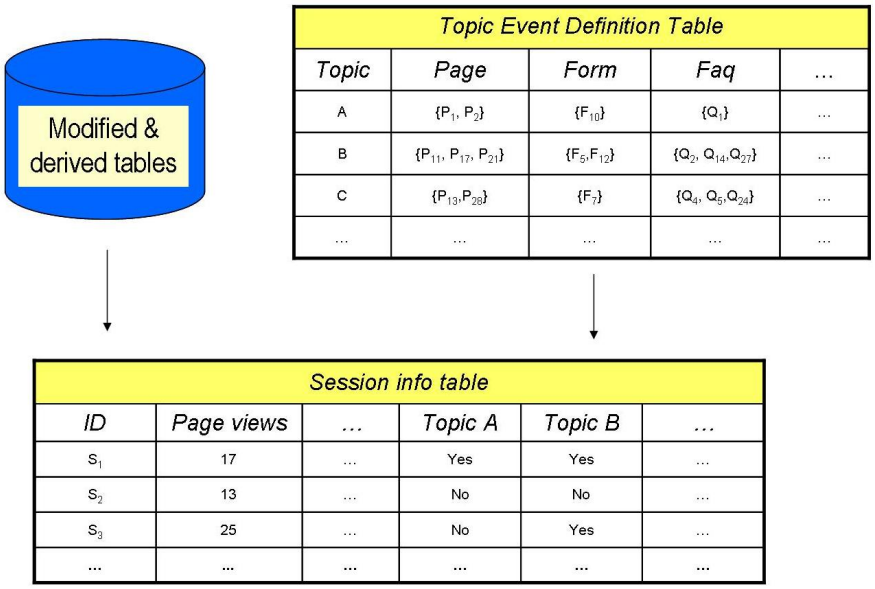


Figure 4.5 – Data transformation step: session info table creation

4.4.3 Customer selection

In the analysis, a particular interval of time has to be selected. This is e.g. due to the data availability. In this interval, both the clickstream and call data have to be available. Using the complete data set would be pass over the fact that each customer has her/his own interval of interaction. Therefore an interval has to be determined from which all customers are chosen to ensure that the complete set of interaction of events related to a particular information need is available in both the clickstream data as well as the call center data. The situation is depicted in Figure 4.6. All customer IDs from the interval $[T_s + \delta \; ; \; T_f - \theta]$ are selected. Onwards all the data that belongs to these customer IDs is selected. Appropriate values for δ and θ have to be determined based on the distribution of the of the average amount of time that exceeds between the first topic related visit. In case of multiple calls made by a customer for the same topic all visits prior to the call but after the previous call are selected.

4.4.4 Web usage selection

Given a particular topic, the set of the clickstream data has to be labeled in order to obtain two classes. The one holds customers with successful web usage, the other

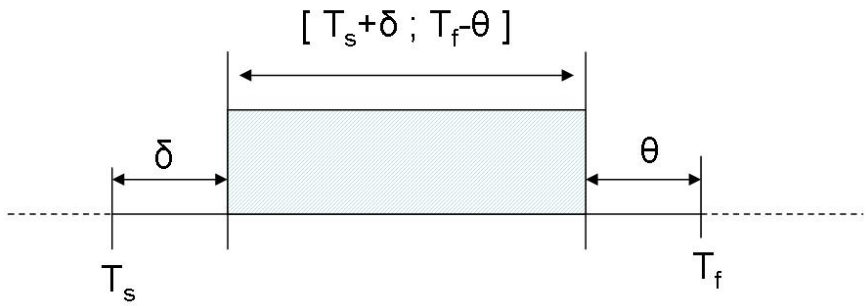


Figure 4.6 – Time frame customer selection. Appropriate values for delta and theta have to be determined based on the distribution of the of the average amount of time that exceeds between the first topic related visit. E.g. if in 90% of the cases an incoming call is within a two days of the first topic related visit, then two days would be a reasonable value for both.

customers with unsuccessful web usage. This step intrinsically contains *candidate instance pruning* as only topic related visits are considered. This is based on the following assumptions:

Assumption 1 *Customers that did not visit topic related concepts and did not make a call (currently) don't have an interest in this particular topic.*

Assumption 2 *Customers that did visit topic related concepts and did not make a call have an interest in this particular topic and were able to find the information they were looking for.*

Assumption 3 *For customers that made a call for a particular topic, the actual usage of the related concepts in the web channel has to be evaluated as it deals with SSP performance.*

4.4.4.1 Unsuccessful web usage selection

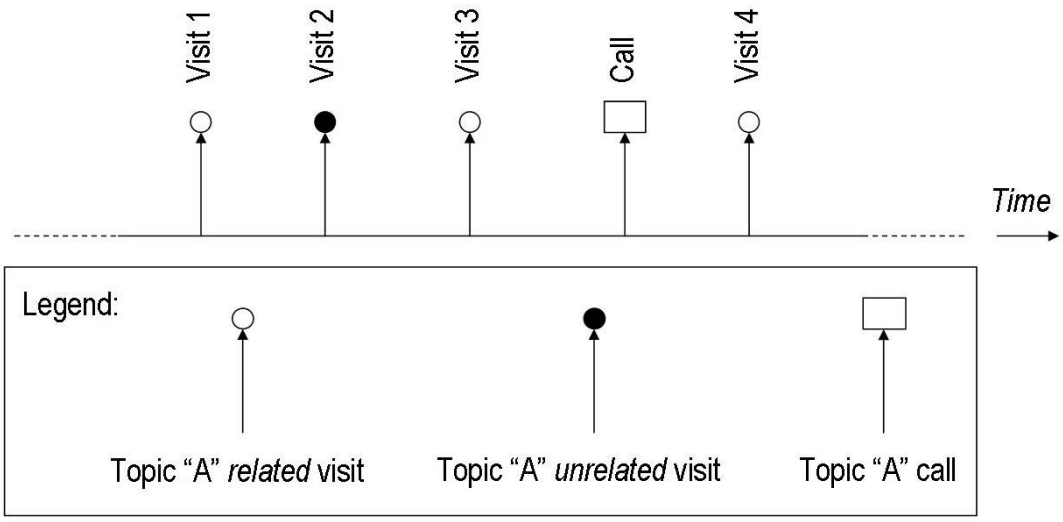
The web usage of customers that made a call is considered to be unsuccessful. This sub set of customers is selected from the call center data. The call topic assignment is assumed to be accurate. The call topic is determined by a human agent that addresses the topic of the call based on the contents of the conversation with the customer. Measures to evaluate this accuracy are hard to define. Yet, call center agents are trained and expected to register each call as carefully as possible as this increases the

value of the information for the company. Figure 4.7 shows an example of the selection procedure that is based on two criteria:

1. The visit has to be *topic related*

2. The visit has to be *prior* to the call.

The first criterium follows directly from the research question. The rationale behind the second criterium is that is assumed that the initial information need that is fulfilled with the call. A visit after the call deals with information need that is different from the initial one.



Visit	Topic related	Occurrence	Include in set?
Visit 1	yes	prior to call	yes
Visit 2	no	prior to call	no
Visit 3	yes	prior to call	yes
Visit 4	yes	after call	no

Figure 4.7 – Entries to illustrate the selection process corresponding of unsuccessful web visits.

4.4.4.2 Successful web usage selection

Given the session info table, the sub set of successful web usage, there is a single criterium for a visit to be included in this set:

1. The customer is not included in the set of customers call made for the topic that is being considered

4.4.5 Customer Topic Event Set

Next, a *Customer Topic Event Set* (CTES) is created. This record holds for each customer all relevant data with respect to the particular self service topic to be evaluated. This data representation addresses two of the principles stated in section 3.6. In the first place, data is processed at the level of a customer in relation to a topic. This in order to deal with session dependency. Besides, order preservation of events in the web usage is considered to be of little importance. Instead the registration of relevant events is of importance. The result is a record for each customer-topic pair.

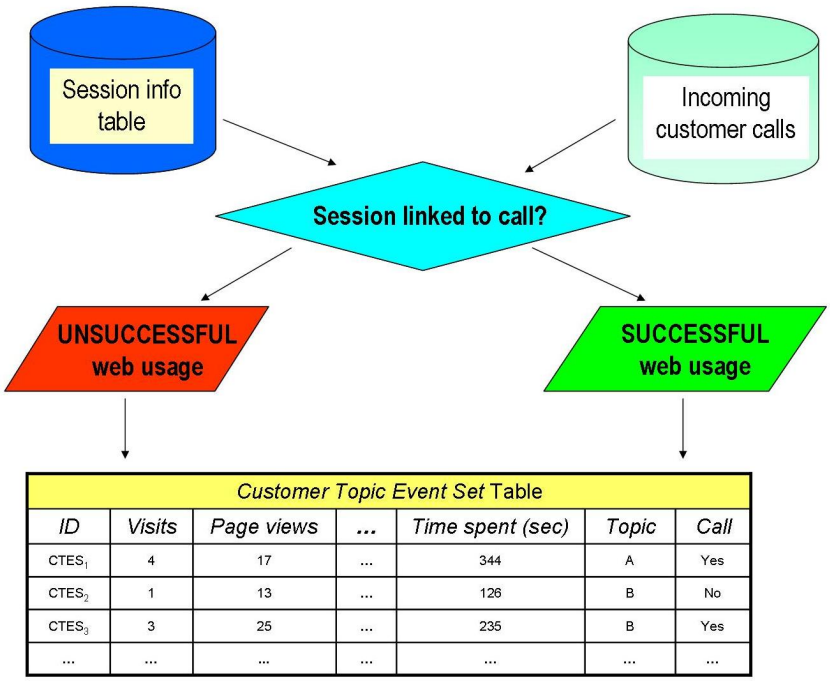


Figure 4.8 – Data transformation step: Customer Topic Event Set (CTES) creation

Attribute	Description
$CTES_{ID}$	Customer-topic record ID
CV	The total number of concepts viewed that hold self service content
DCV	The total number of distinct concepts viewed that hold self service content
CV_{t_i}	The total number of concepts viewed that hold self service content related to topic t_i
DCV_{t_i}	The total number of distinct concepts viewed that hold self service content related to topic t_i
TS_{t_i}	The total amount time spent on concepts that hold self service content that map to t_i
C_{t_i}	Boolean attribute to indicate whether there was a topic related call t_i

Table 4.5 – CTES attribute overview

4.5 Step 3: Data transformation

As mentioned in Chapter 3, clickstream data mining is clearly in its infancy with respect to this aspect. In this thesis the feature construction and selection task is the result of knowledge from WA best practices in segment definitions and common sense reasoning. From the clickstream data metrics are available. Metrics are by default numbers and counts. Features derived from these metrics will be numeric. In order to reduce the dimensionality of the data, techniques are used as described in [28] to map these numeric attributes to categorical. These categories are in a supervised way based on domain knowledge. Onwards, a *Customer Topic Feature Set* (CTFS) table is created for each topic holding instances that are homogenous with respect to this topic. A CTFS is created from the CTES table by applying two types of actions:

1. Numeric attributes will be mapped to a categorical value according to the techniques presented in [28, p.418] to discretize numeric attributes. This technique is used to reduce the dimensionality of the attribute value set as high dimensionality degrades classifier performance.

2. Customer related data will be will be added to enrich the data (e.g. demographic data). The additional data that is entered will be domain and topic specific. On the on hand it depends on the attributes that available. On the other hand, the topic will determine which attributes are relevant. E.g. in case a topic is considered that is related to changing contact data, it is of little relevance to include a customers payment history. These considerations and choices have to be made for each topic.

For matters of convenience and readability the feature construction is presented in the case study. Here an overview is given of the features. A distinction is made between *general* and *specific* features. The set of general features apply in general. The set of specific features depends on the topic for which the segment definition has to be derived. It mainly depends on the information that is available and that is relevant.

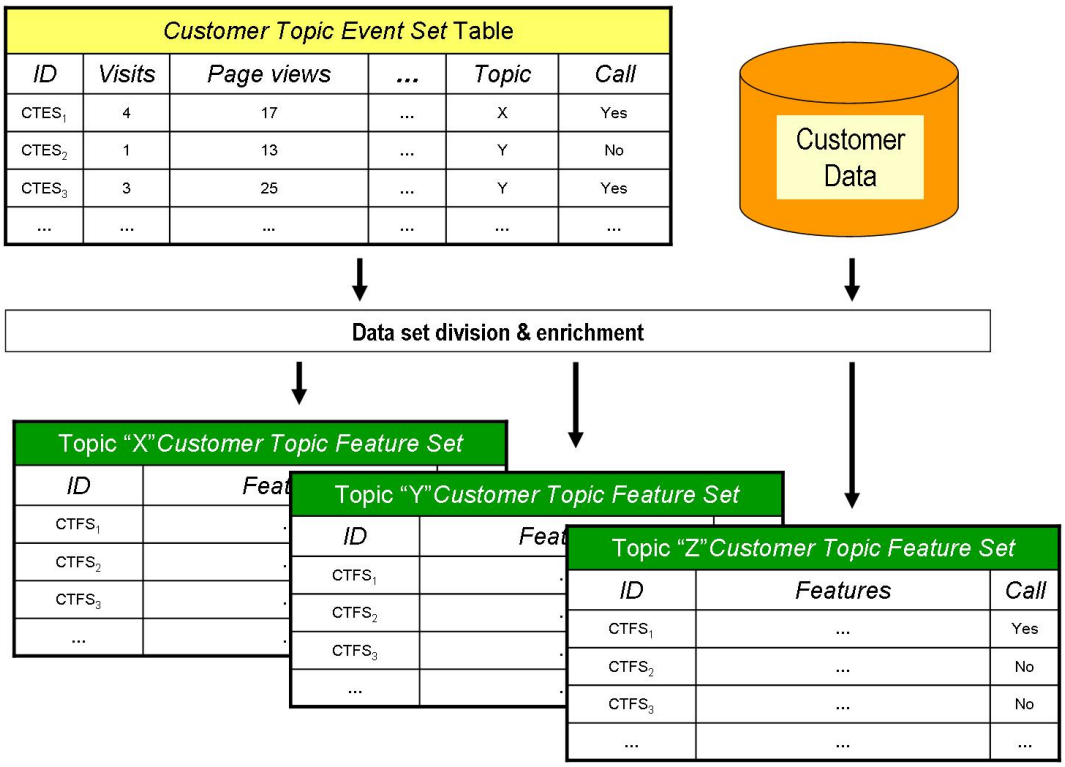


Figure 4.9 – Data transformation step: Customer Topic Feature Set creation

4.6 Step 4: Data mining

In this section first an description is given of the place of the mining task in the framework. Next, the motivations for the use of a *rule classifier* in the data mining task are presented. Onwards, the characteristics of the *RIPPER* rule classifier are described and the measures to evaluate the quality of the model obtained. Onwards, the setup for the classification task is presented. Lastly, an evaluation method for the performance of the segment definition is elaborated.

Figure 4.10 depicts the contribution of the mining task in this framework. Given a particular topic that is being analyzed and its corresponding CTFS, the role of the rule classifier is to output a set of rules that discriminates between successful and unsuccessful web usage. The CTFS serves as the training set. The output of the rule classifier has to be *transformed* into first order boolean logic which is the formalism of a WA suite to define segments.

4.6.1 Motivation for using a rule classifier

The classifier has to address/overcome three important issues that are characteristic for the context of this thesis:

1. In a WA suite a segment has to be defined in first order boolean logic.
2. A segment evaluation can be considered to be a *two-class* classification task. Unsuccessful web usage that satisfies the segment definition can be considered to be *yes* instances, whereas successful web usage are *no* instances.
3. Imbalanced class distributions

Segmentation can be considered to be a binary partitioning task. Given segment a condition, each instance is tested against this condition. There are two possible outcomes. Either the instance is included in the segment or not. This allows for placing clickstream in a separated segment as well in order to explore the web usage behavior if this customer group to determine the differences. Therefore the classification task is

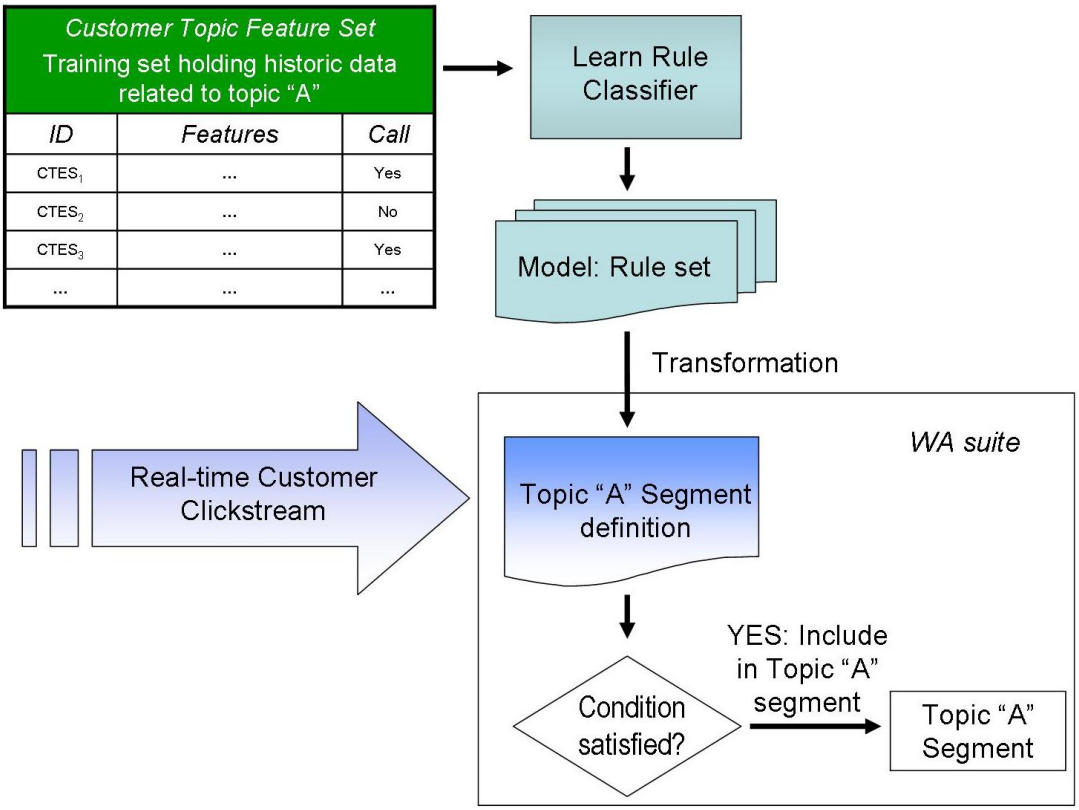


Figure 4.10 – Training a rule classifier in order to discriminate between successful and unsuccessful web usage with respect to a particular topic. All data is by definition placed in a *global* segment that holds all the clickstream data that is collected.

represented as a 2-class problem. One of the classes is chosen as positive class (unsuccessful web usage), and the other as negative class (successful web usage).

In [28, p.223] an overview is presented of the characteristics of a rule classifier:

- Outputs (relatively) descriptive models that are easy to interpret
- Performance comparable to decision trees
- Handle imbalanced class distribution relatively well (RIPPER algorithm)

These characteristics join the requirements to a large extent. Therefore this type of classifier suites best in this framework.

4.6.2 Alternative approach

Rule classifiers use the sequential covering algorithm [32, 105] to extract rules *directly* from the data. This is considered to be a *direct method for rule extraction* [28, p.213]. An alternative approach would be to learn a *decision tree* from the data and to generate rules from this tree [28, p.221]. This is considered to be an *indirect method for rule extraction*. However, as pointed out, a rule classifier has a performance that is comparable to that of a decision tree without the additional task to generate rules.

4.6.3 Motivation for the use of the RIPPER algorithm

One of the popular rule classifiers that are described in literature is the *RIPPER* (Repeated Incremental Pruning to Produce Error Reduction) algorithm [6]. This algorithm is proposed to be used in the context of this thesis. Besides its classification qualities and characteristics, the WEKA has an implementation of the RIPPER algorithm so it is directly available. Below is a summary of the main characteristics of this algorithm [28, p.220].

- For two-class classification, majority class is set to *default* class and rules are learned for detecting the minority class
- The sequential covering algorithm used: rules directly extracted from the data
- General-to-specific rule growing strategy
- Class based ordering scheme for the rule set that is outputted

Especially the aspect of the class based ordering scheme is a nice feature. It allows for a cut of the rules that is considered to be sufficient by selecting a subset of the rules. E.g. in case there are several rules but the first rule has high accuracy and does not require the implementation of additional parameters in the clickstream, this might be considered to be best for the situation.

4.6.4 Classification model evaluation

Two measures are proposed to evaluate the performance of a classification rule [28, p.208]:

- *Coverage* of a rule: Fraction of records that satisfy the antecedent of a rule
- *Accuracy* of a rule: Fraction of records that satisfy both the antecedent and consequent of a rule

Ideally a rule set is obtained that is 100% accurate that has 100% coverage. In practice, this is unlikely to be the case. In the consideration of the trade-off between accuracy and coverage, it is best to strive for maximal accuracy with a reasonable coverage. The requirement for proper segment analysis is that the data is homogeneous with respect to the cross channel usage behavior. Increased coverage is likely to result in the assignment of membership to clickstream that is related to successful web usage which is undesired.

In the evaluation of the accuracy to be used model to be used a *confusion matrix* is used. This is a metric that is traditionally used to summarize the number of correctly and incorrectly predicted instances by a classification model [28, p.296]. There are four possible verdicts given the label assigned to an CTFS entry by the model.

1. Unsuccessful web usage is correctly included in the segment → *True Positive*
2. Unsuccessful web usage is incorrectly not included in the segment → *False Negative*
3. Successful web usage is correctly not included in the segment → *True Negative*
4. Successful web usage is incorrectly included in the segment → *False Positive*

Tables 4.6, 4.7 and 4.8 are used in the classification model evaluation.

	Predicted class		
	UWU	*SWU*	
Actual class	**TP**	**FN**	*UWU*
	FP	**TN**	*SWU*

Table 4.6 – Confusion matrix with class labels for unsuccessful web usage (UWU) and successful web usage (SWU)

Entry	Include in segment?
True Positives (TP)	Yes
False Negatives (FN)	Yes
True Negatives (TN)	No
False Positives (FP)	No

Table 4.7 – The confusion matrix entry interpretation in relation to the web usage.

Type	Set
UWU	$\mathbf{TP} \bigcup \mathbf{FN}$
SWU	$\mathbf{FP} \bigcup \mathbf{TN}$

Table 4.8 – Desired segment web usage selection

The crucial aspect is that the segment definition will select entries from $\mathbf{TP} \bigcup \mathbf{FP}$. Apart from selecting correct entries (FP), entries are selected that should have not been included (FP) and entries are discarded that should have been included (FN). The goal with respect to segment definitions is to have a sub set of the data that holds most unsuccessful web usage while minimizing the number of successful web usage instances. The motivation is that a WA suite reports on aggregated data. The actual class label of the clickstream cannot be discovered within the report suite. In terms of the confusing matrix this gives a requirement to the model that:

- Maximizes the number of True Positives
- Minimizes the number of False Positives

4.6.5 Classification issues

There is a set of well-known issues related to learning a classification model. Two of these issues apply to the setting of this thesis. These are addressed in the next paragraphs.

4.6.5.1 Class imbalance

The setting of this context deals with class distributions that are inherently imbalanced. Class imbalance [28, p.294]is a heavily studied research topic in machine learning clas-

sification and performance. It is dealt with in a variety of ways [16] [22] [11] [20] [31]. The method proposed in the setting of this thesis is to used *cost sensitive learning*[32, p.164].

Cost sensitive learning The rule classifier is embedded in a *meta* classifier to make it *cost sensitive*. Given the confusion matrix, a penalty is given for misclassified instances. False positives will be assigned a higher penalty than false negatives. The rationale is that WA suites report on aggregated data. Once in the segment it is impossible to discriminate between successful and unsuccessful web usage. Given the cost for misclassification, the classifier is learned to reduce the overall cost of misclassified instances and it will predict the class with the least expected misclassification cost.

4.6.5.2 Adressing model overfitting

Another well-known issue is *model overfitting*. This can be divided into two types. A *training error* is a misclassification of a training record - the historic data from which the model is learned. In learning a classification model, the set of historic data is divided into *training data* and *validation data*. The training data is used by one or more learning methods to come up with classifiers. The validation data is used to optimize parameters of those classifiers, or to select a particular one. A *generalization error* is the expected error of the model on previously unseen records. To overcome this issue, *cross validation* will be used in the configuration of the model training.

Cross validation Cross validation is a technique to reduce the generalization error a classifier [32]. In general, it is best to have as much training instances as possible to learn a model. However, there is only a limited amount of data available. In $k - fold$ cross-validation a number of k folds or partitions of the data is created. Each of these k partitions holds approximately an equal number of instances. Onwards, k models are learned. Each partition serves once as a the validation data to determine the accuracy of model that is obtained from the training data that consists of the other $k - 1$ partitions. Besides, *stratification* is required. The random sampling of the data to obtain k partitions has to be achieved in such a way that each class is properly represented in both the training and validation set.

4.7 Step 5: Evaluation

Once a segment definition is implemented in the WA suite, it will test clickstream data for segment membership real-time. To evaluate the how well the segment definition performs a confusion matrix is used as well. Figure 4.11 depicts the situation. S_{global} denotes the segment (without condition) that holds all clickstream data. S_A denotes the segment that is defined for topic A. Numbers 1 through 4 denote the sets that are considered. The quality of the definition - which is in fact the accuracy and coverage of the rule classifier model built on the training data - is evaluated in the following way. Given a particular time frame that is used to evaluate the segment definition for topic A, the following steps have to be performed.

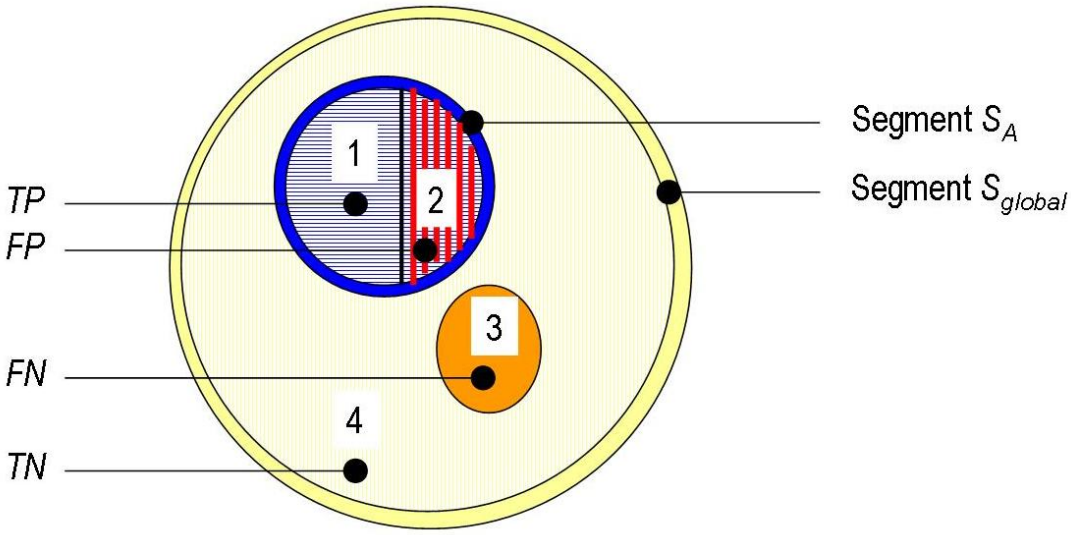


Figure 4.11 – Venn diagram of segment membership possibilities

- Create set $Customer_{global}$ by selecting all customer IDs from the WA clickstream. Depending on the implementation, the customer IDs that are assigned to the segment of topic A are marked.
- Create set $Customer_{SegmentA}$ by selecting all customer IDs from $Customer_{global}$ that are included in the segment for topic A.
- Create set $Customer_{NON-SegmentA}$ by selecting all customer IDs from $Customer_{global} \setminus Customer_{SegmentA}$.

Measure	Formula
True positives - **TP**	$Customer_{SegmentA} \cap TopicA_{calls}$
False positives - **FP**	$Customer_{SegmentA} \setminus$ **TP**
False negatives - **FN**	$Customer_{NON-SegmentA} \cap TopicA_{calls}$
True negatives - **TN**	$Customer_{NON-SegmentA} \setminus$ **FN**

Table 4.9 – Segment evaluation ratios.

- Create set $TopicA_{calls}$ by extracting all customer IDs of incoming calls for topic A from customers that had at least one topic related visit prior to the call.

Table 4.9 lists the ratios can be computed in order to evaluate the accuracy and coverage of the model. As mentioned before, it is best to have a segment definition that holds a reasonable number of true positives with a minimum number of false negatives.

Chapter 5

Case study: MijnNUON

5.1 MijnNUON characterization

MijnNUON is the username/password secured customer SSP at www.nuon.nl. It allows registered customers to view personal data and perform what are considered to be *routine tasks*. The assumption is that MijnNUON registered have to a certain extent the awareness of the services available in this portal. Customers have to sign-up to have access to the portal. Moreover, since only authenticated visits are considered in this thesis, customers have to actually login. They know of the existence of the portal and they are likely to know the available overviews. The exact level of familiarity is hard to define, yet it allows for reasoning about particular usage behavior.

5.1.1 Time window

The time window that is subject of analysis and for which data from both the call center and the clickstream overlap ranges from 01-06-2008 through 31-08-2008. For each customer that made a visit during this time frame, the complete history of channel use (calls, e-mails etc.) for interactions with NUON (other than the web visits) is available.

5.1.2 Population

The following conjecture is made: *SSPs are no fun or entertainment sites that are listed in peoples top-10 favorite web sites. SSPs are used because of a particular necessity.* The focus is on customers that had at least one MijnNUON visit prior to a self service

related call. They are assumed to be familiar with the concept of MijnNUON and self service opportunities.

5.1.3 Web site characteristics and statistics

Figure 5.1 provides an high level overview of the NUON web site. Table 5.1 gives a brief description of each of these concepts. The set of *personal*, *forms* and *FAQ* concepts make up the set of self service concepts. The contactpage is included as well as it holds the number of the call center.

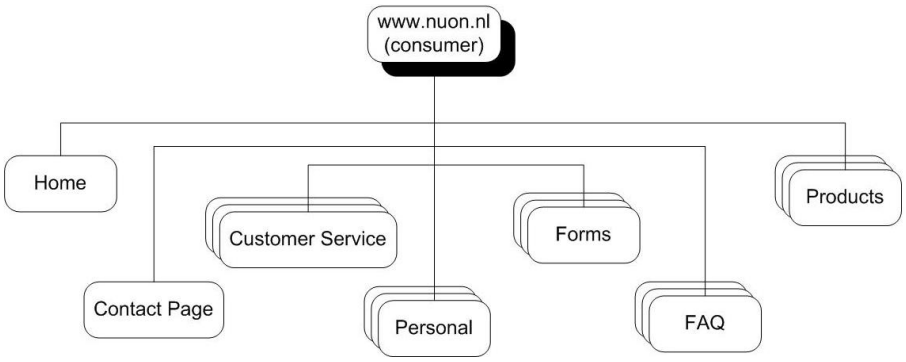


Figure 5.1 – MijnNUON consists of the personal and the forms section. Both the FAQ section and the contact page are supportive to this portal. These concepts make up what is considered to be the set of self service concepts

5.2 Topic *payment amount change*

Nuon customers pay there bill for energy in advance on a monthly basis. The amount to be paid is estimated based on several parameters. These parameters are both personal - estimated consumption based historic consumption - and more general - expected price fluctuations. The total sum is divided by twelve and this makes up the amount to be paid on a monthly basis. After twelve payments in advance, the annual account is made up. This bill shows the actual consumption en the amount of money to be paid or to receive. Modifying the payment amount is by definition with respect to the frequency not a routine task as is only allowed once per consumption year. Yet it should be a routine task with respect to the complexity.

Item	Explanation
Home	Index or hub page. The entry page in 95% of the visits. Contains links to all concepts.
Contactpage	Page holding all contact to NUON related data including the phone number of the NUON call center
FAQ	List of Frequently Asked Questions and Answers. List is drawn up from on the basis of the questions that are posed in the call center
Forms	All forms that are available to change personal data
Personal	The actual MijnNUON environment. Accessed through customer unique username/password combination
Customer Service	Set of pages serves as guide to the user. Points user to a particular section or page based on the information need
Products	Set of pages with product info
Other	The set of remaining pages

Table 5.1 – www.nuon.nl concept outline

There is a three step form available to accomplish this task. Many customers succeed, yet there is a substantial number that does not. Insights in their characteristics can be valuable in order to address the motivation for this group of customers to make a call instead of completing this task over the internet.

5.2.1 Increase periodic amount

If a customer expects that the energy consumption will be *higher* than estimated, the periodic amount can be increased. This spreads out the cost over a period instead of a payment at once at the end of the year.

5.2.2 Decrease periodic amount

If a customer expects that the energy consumption will be *less* than estimated, the periodic amount can be decreased. Each customer can decrease the periodic amount at most once a year with a lowering of a most 10% of the current amount. Moreover the meter readings must haven been taken instead of estimated[1]

5.3 Data exploration

Data exploration is a preliminary investigation in order to better understand its specific characteristics. Especially in the case of the clickstream data this is an important aspect. Visualization of the data can be helpful to discover particular characteristics based on which features can be constructed. A comprehensive overview of techniques for visualization in data mining is available in [25]. The WA suite can also be used to inspect the data as it provides many visualization options. Data from three different sources is used. The clickstream data, logs from the call center and additional customer related data. The data from of these sources will be addressed in a dedicated section.

5.3.1 General

In this section several facts are outlined that apply to the setting of NUON, www.nuon.nl and the SSP. They are meant to illustrate the aspects that are taken into account when

[1]There is no data available whether the meter readings actually have been taken or not at the moment a customer has a web visit.

Window of analysis related facts that have to be taken into account
• All timestamps in summertime • Holidays in the Netherlands • No changes made to the web site elements that are analyzed • General rate raise from the 1st of July 2008 • Clickstream data missing from 29-06-2008, 02-08-2008 and 03-08-2008 due to server maintenance. SSP was unavailable during these days • The contact page holds a general number. No means to discriminated/determine if contact page number is used. • If customer visits the site but does not login then this visit is not tracked. This does affect the analysis. These visits do not particularly contribute to *fulfilling information need* as this information is only available in the portal. • No active linking between SSP and product concept. E.g. in case of a risen expenses, a customer environment is adaptive by offering a energy saving product.

elaborating this case study. It is not possible to give a general strategy as these elements are domain specific. However, each domain will have its own characteristics and peculiarities that contribute to the way customers interact with the SSP over time and that underlie the data sets that are used in this context.

5.3.2 Phenomenon of *Monday morning office hour browsing*

Increased activity in web and call center channel usage on Monday morning. This holds for all concepts and topics, authenticated and anonymous visits. NUON is familiar with this aspect, yet there is no clear explanation. However, it affects the data that is collected and how it should be interpreted. E.g. people are at the office. This means that there is not much time to visit www.nuon.nl and to resolve the issue themselves. Phone calls, colleagues might interrupt, multiple browser tabs will be open at the same time etc. However, these factors will influence the data that is collected.

Table 5.2 – Overall MijnNUON customer traffic related statistics

Metric	Number
Page views	1.797.128
Visits	120.539
Distinct customers	51.363

Metric	Overall views		Concept views		Topic related views	
	S	U	S	U	S	U
#Pages viewed	14.8	16.5	10.9	12.3	2.6	2.8
#Distinct pages viewed	10.1	11.0	7.5	6.2	1.6	1.8
$\frac{Time\ spent\ (seconds)}{Visit}$	518	594	379	518	76	109

Table 5.3 – The table gives an overview of the average number of views divided into successful (S) and unsuccessful web usage (U). *Overall views* includes all page views including an index pages such as the homepage. *Concept views* includes all self service related concept views. *Topic views* includes all topic related concept views.

5.3.3 Clickstream data

In the customer selection $\delta = \theta = 4$ days is used (see Figure 4.6). In this way for 93% of the customers the complete sequence of web transactions for this topic is selected. This results in a number of 1153 unsuccessful web usage instances for this topic. The number of successful web usage instances for this topic is 50210.

5.3.4 Concept drift

As mentioned, throughout the window of analysis there were no radical changes made to both the content of the web site as well as to its structure. The contents of the SSP is dynamic and customized for each user. This can be considered to be a form of concept drift[19]. SSP typically customized. From time to time updated new data is uploaded: new term account, amount, annual account. with respect to these items, a time stamped table is available from which can be derived which numbers were shown at which moment in time.

#Visits	#Customers	Percentage
1	17852	77,7%
2	3220	14,0%
3	889	3,9%
4	419	1,8%
5	219	1,0%
6	121	0,5%
7	71	0,3%
8	56	0,2%
9	26	0,1%
10	33	0,1%
10+	68	0,3%

(a) Successful web usage

#Visits	#Customers	Percentage
1	779	68,6%
2	217	19,1%
3	68	6,0%
4	30	2,6%
5	18	1,6%
6	9	0,8%
7	4	0,4%
8	3	0,3%
9	1	0,1%
10	1	0,1%
10+	6	0,5%

(b) Unsuccessful web usage

Figure 5.2 – MijnNUON Visit distribution divided for successful and unsuccessful web usage

5.3.5 Temporal aspects: heat maps

In order to get an understanding of the use of MijnNUON three heat maps show different characteristics (see Figures 5.3, 5.4 and 5.5). The first shows the visit distribution. The second shows the distribution over time of the number of pages viewed. The third one relates the amount of time spent to time slots. From heat maps on the next two pages, there is no clear difference w.r.t to visits, page views and time spent over time that discriminates unsuccessful from successful web usage. Yet, this is informative in itself and complies to the what is known by NUON with respect to the peaks and lows in channel usage.

5.3.6 Call center data exploration

- No question asked "Did you use the web site to resolve this issue yourself?"
- Active push to have customers to sign up for MijnNUON. Non-registered customers that make a call are brought to their attention about MijnNUON. This contributes to the pattern of customers that make a call and that sign up for the SPP afterwards.

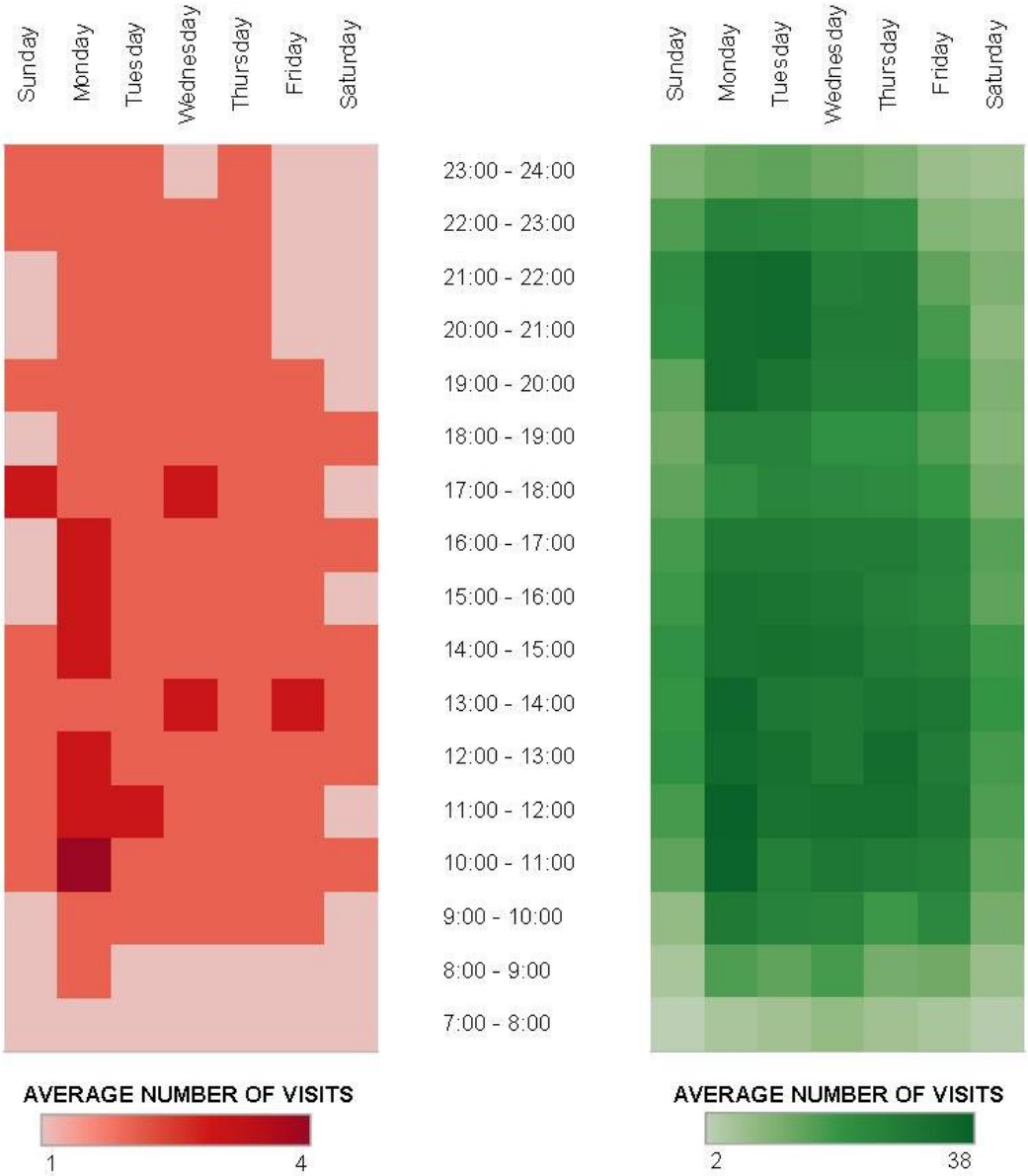


Figure 5.3 – The heat maps show the distribution of the average number of payment amount related visits. The image on the left is based on visits from customers that made a call on "change payment amount". The image on the right is based on visits from customers that did not make a call on "change payment amount".

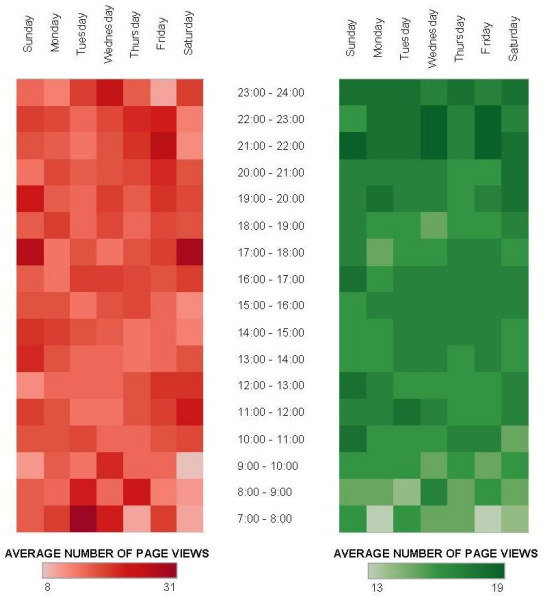


Figure 5.4 – The heat maps show the distribution of the average number of pages viewed in payment amount related visits. The left image is based on visits from customers that made a call on "change payment amount". The image on the right is based on visits from customers that did not make a call on "change payment amount".

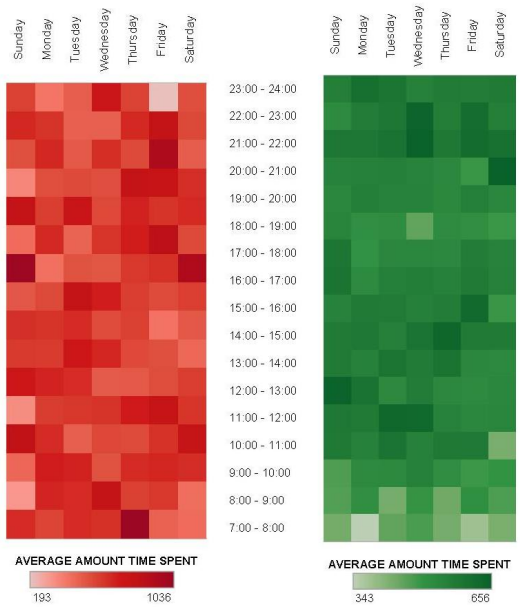


Figure 5.5 – The heat maps show the distribution of the average amount time spent per visit in payment amount related visits. The left image is based on visits from customers that made a call on "change payment amount". The image on the right is based on visits from customers that did not make a call on "change payment amount".

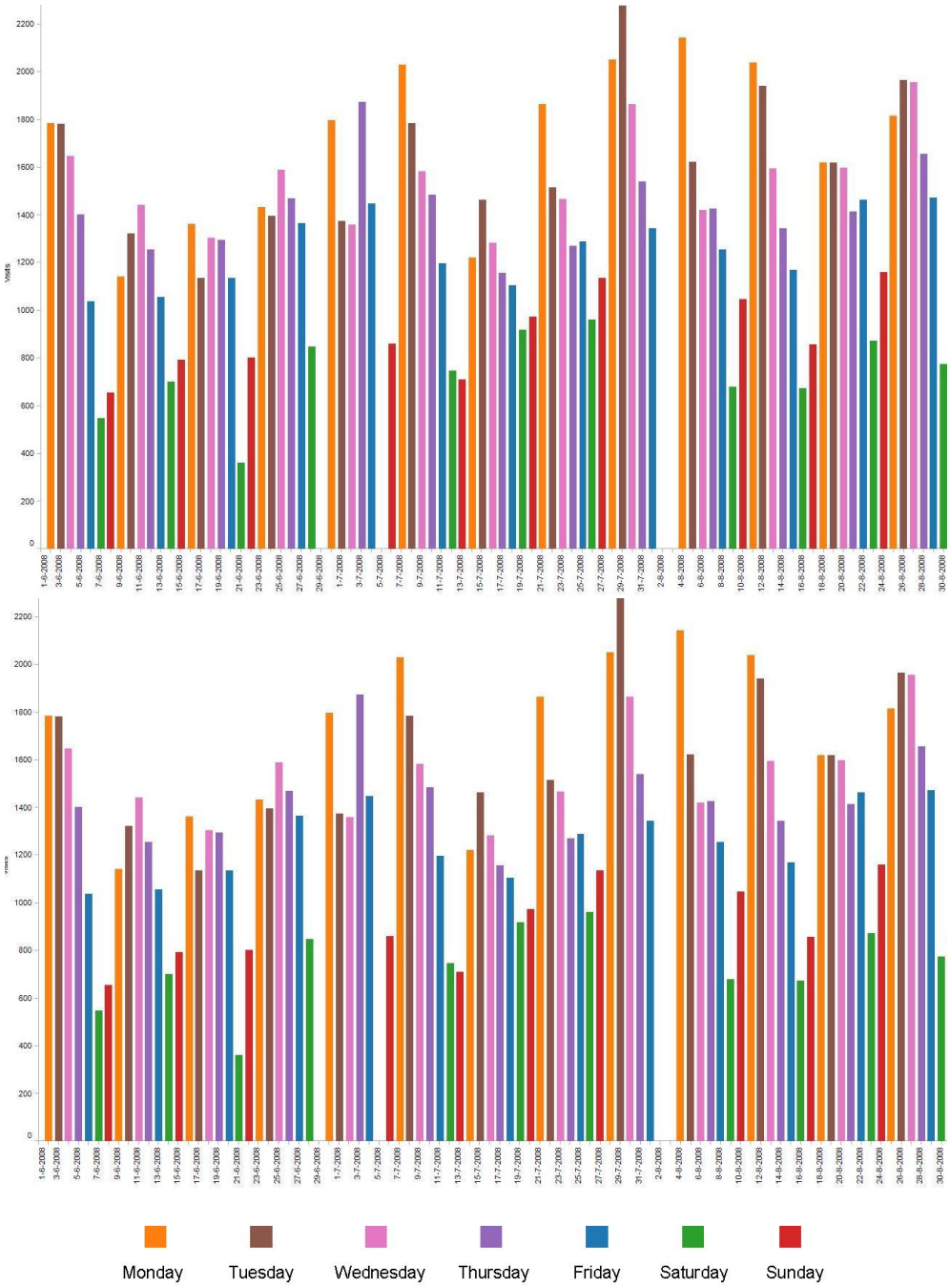


Figure 5.6 – The top figure shows the visit trend for successful web usage, the bottom figure for unsuccessful web usage. The vertical axis holds the total number of visits. There are no clear discrepancies among the trend lines. Within each view the trend is consistent. Peaks at the beginning of the week and lows in te weekends. There are no breaches in the trend the day after SSP unavailability.

A random check in the call center transcripts did not bring abnormalities to light. Customers simply called to increase or decrease the amount. The call center classification by the human agents is assumed to be accurate. They are instructed to classify each call according from selection fields that hold values from a predefined set of classification labels. These labels are entered in form. These fields do no contain a default value. This in order to force the agent to make a proper choice.

5.3.7 Customer data

Additional data is used in the CTFS to enrich the record with demographic and payment related data. The demographic attributes contained *null* values. Since the RIPPER algorithm uses the sequential covering algorithm, missing values are treated as though they dont match any of the tests [32, p.201].

5.4 Topic event definition table

Visits to be identified as *topic related visits* based on the pages that were viewed during the visit:

Concept	Page name
Form	/form/online.paymentamountchange/step 1 /form/online.paymentamountchange/step 2 /form/online.paymentamountchange/step 3
Faq	/faq/5502 - why is the current term amount different from that in my previous house? /faq/How do I change the term amount? /faq/What is the term amount?
Customer service	/cs/pass on change/change term amount
MijnNUON	/personal/change/my term amount /personal/view/online annual account/term amounts /personal/view/mijn invoices/term invoice

Table 5.4 – Case study topic event definition table

5.5 Feature construction

Based on the insights obtained from the data exploration, the features are constructed. There is on important requirement:

- Have to be expressible in boolean logic as this is the formalism to define segments. Moreover, in cases where additional scripting is required to capture data elements, these to be implementable/expressible in script terms

Table 5.5 gives an overview of the division of features into *general* and *specific*. General features are proposed independent of the specific topic that is subject of analysis. Specific features are constructed based on the topic that is subject of analysis in combination with the type of the additional information that is available. Besides a division is made based on the source. *Clickstream* refers to the features that are currently available in or that can be directly computed from the clickstream. *Call center* refers to additional attributes that are available in the call classification. This will differ from company to company. *Customer* refers to additional attributes that are customer related and that are available in a database. The possibilities to capture data from this source will depend on the specific configuration of a the databases in relation to the web server.

		Generic	**Specific**
Source	Clickstream	I	IV
	Call center	II	V
	Customer	III	VI

Table 5.5 – The table shows the division of features according to their underlying data source and the generality.

5.5.1 Feature definition table

Feature TRV: TRV denotes the total number of topic related visits. For callers, only the number of topic related visits *prior* to the call is counted. For all other customers all topic related visits are counted.

Feature	Based on	Data type	Category
TRV	Web usage	numeric	I
SSC_{Views}	Web usage	numeric	I
$SSC_{DistinctViews}$	Web usage	numeric	I
$SSC_{TimeSpent}$	Web usage	numeric	I
SST_{Views}	Web usage	numeric	I
$SST_{DistinctViews}$	Web usage	numeric	I
$SST_{TimeSpent}$	Web usage	numeric	I
FU	Web usage	numeric	I
FAQ	Web usage	numeric	I
$ContactPage$	Web usage	boolean	I
$CC_{history}$	Call center use	numeric	II
$SSP_{history}$	Call center use	numeric	III
Sex	Demographic	categorical	III
$AgeGroup$	Demographic	categorical	III
$SSP_{registration}$	MijnNUON	categorical	III
$AmountRange$	Payment	categorical	VI
$Relocation$	Demographic	boolean	VI

Table 5.6 – Feature definition table

Motivation: Indication of the interest in this particular topic over time.

Concept view related features:

- SSC_{Views} denotes the total number of self service concepts viewed in all visits.
- $SSC_{DistinctViews}$ denotes the total number of distinct self service concepts viewed in all visits.
- $SSC_{TimeSpent}$ denotes the total amount time spent on self service concepts viewed in all visits.
- SST_{Views} denotes the total number of distinct topic related concepts viewed in all visits.
- $SST_{DistinctViews}$ denotes the total number of topic related concepts viewed in all visits.
- $SST_{TimeSpent}$ denotes the total amount time spent on topic related concepts in all visits.

Motivation: Attributes to express the site and SSP exploration.

Feature FU: By definition, all instances labeled n have used all 3 distinct form pages (and viewed at least 3 form pages) as these are required to successfully complete the transaction. There is no assumption to be made on the number of form pages viewed. Therefore a ratio is used that expresses the average amount of time spent on a form page.

- $FormPageViews$ denotes the total number of form pages viewed
- $Form_{ts}$ denotes the total amount time spent on form pages

$$FormPageView_{ratio} = ROUND((\frac{Form_{ts}}{FormPageViews}) * 100)$$

In case of absence of this concept in the clickstream, there a two extraordinary cases:

- 0 form pages viewed: set value to 0
- 1 form page view, the last in the visit: no time spent. Set value to 0

Motivation: Attribute to indicate the concept usage. Captures the absence of this concept in the clickstream as well.

Feature FAQ: FAQ denotes the total number of faq answers related to periodic amount change that were viewed

Motivation: Attribute to express to what extent the customer tried to resolve the issue by using the dedicated FAQ answers.
that are available.

Feature $C_{history}$: $Call_{history}$ denotes the total number of incoming calls generated by the customer for the given topic. Differentiate on the basis of call frequency.

Motivation: Attribute to express the history of a customer with respect to call center channel usage.

Feature $SSP_{history}$: $SSP_{history}$ denotes the total number of self service transactions generated by the customer for the given topic.

Motivation: Attribute to express the history of a customer with respect to self service usage.

Feature $ContactPage$: $ContactPage$ denotes whether the customer viewed the page that holds the number of the call center. This page is not viewed *by accident*. The web site structure and topology are such that always two unique pages have to be visit consecutively in order to obtain the number.

Motivation: Attribute to express whether the customer actively search for the call center phone number on the web.

Feature *Sex*: *Sex* of the customer. This is a binary valued attribute. Either M (male) or F (female).

Feature *AgeGroup*: Feature that takes the age of the customer. It is plausible that the customer that signed the contract will use the SSP. However, an SSP account is related to a contract, not to the person that actually uses the portal. An estimate on the number of customers that use the portal without being the signer of the contract cannot be given.

Agegroup	
Number of months registered	**Nominal value**
age ≤ 30	30-
$31 \leq$ age ≤ 35	[31-35]
$36 \leq$ age ≤ 40	[36-40]
$41 \leq$ age ≤ 50	[41-50]
age > 50	50+

Feature $SSP_{registration}$ Feature to express (to a certain extent) the familiarity with MijnNUON. Taken as the number of months elapsed between MijnNUON registration date and transaction date. A manual discretization method was used to map the numerical values to categorical values.

SSP experience group	
Number of months registered	**Nominal value**
$SSP_{registration} = 0$	0
$1 \leq SSP_{registration} \leq 3$	[1-3]
$4 \leq SSP_{registration} \leq 6$	[4-6]
$7 \leq SSP_{registration} \leq 12$	[7-12]
$SSP_{registration} > 12$	12+

Feature *Relocation*: Feature to indicate whether there was there was a relocation at most 3 months prior to the call/first visit in the window of analysis. The period

of three months was chosen because within this range this attribute was available for each customer. Besides, in general customers are confronted with a different periodic amount after a relocation. This is because the periodic amount is based on the energy consumption history of a house and not of a customer. Many customers are unaware of this.

Feature *AmountRange*: For each customer the periodic amount at the moment of visiting is known. A manual discretization method was used to map the numerical values to categorical values.

Amount group	
Amount	**Nominal value**
$0 \leq$ amount ≤ 49	[0-50)
$50 \leq$ amount ≤ 99	[50-100)
$100 \leq$ amount ≤ 149	[100-150)
$150 \leq$ amount ≤ 199	[150-200)
$200 \leq$ amount ≤ 249	[200-250)
$250 \leq$ amount ≤ 299	[250-300)
amount ≥ 300	300+

Chapter 6

Experiments

In this chapter an experiment with respect to the payment amount change topic is elaborated. First, the number of instances is reduced based on the characteristics of the customers with respect to changing the periodic amount. Onwards the setup used in WEKA to run the test is described. Finally, the model that is outputted is presented.

6.1 Instance selection

Given the customers that were selected in Chapter 5, two instance types can be excluded beforehand from the set of instances. This reduces the number of web usage instances for the classification task.

1. Customers that successfully used the form to modify the payment amount. This can be verified easily: the form process is completed successfully in the self service environment.

2. Customers that changed the payment amount within a year prior to visit, since it is only allowed to decrease the term amount once a year. This can be verified by checking the transaction history of a customer with respect to the payment amount change.

This reduces the number of instances to 36734 holding 1136 instances for unsuccessful web usage and 35598 for successful web usage.

6.2 WEKA classification setup

The *Explorer* module of WEKA (version 3.5.8) was used to configure the test setup (see [32] for full documentation of using and configuring WEKA). The *CostSensitiveClassifier* was selected from the set of available meta-classifiers. This allows for making its base-classifier that will be embedded cost sensitive. The base classifier to be embedded is JRip which is the WEKA implementation of the RIPPER algorithm. Besides, 10-fold cross validation was used to address model overfitting. The number of 10 chosen as is considered to be a good value on average [32, p.150]. Several runs were made with different configurations of the penalty assignment to false positives and false negatives. One matrix resulted in output that can be considered to be reasonable. The output is presented below. The accuracy of the model was 98.1.

Cost matrix	
0	1
2	0

Confusion Matrix		
a	b	← classified as
602	534	a = y
159	35439	b = n

This configuration maximizes the number of true positives while keeping the number of false positives to a minimum. This was the requirement for a reasonable classification model as indicated in Chapter 4.

6.3 Rule set evaluation

Below is the model show that is obtained. This rule set can be transformed in a segment definition. Given the current implementation of the NUON WA solution it would be required to capture additional parameters. These include *Relocation*, $SSP_{history}$ and $CC_{history}$. Besides, a boolean value has to be captured in the clickstream that excludes customers that are not allowed to change their payment amount (due to the restriction that this is only allowed once in a year).

```
1) (FU >= 6) and (SSP_{history} >= 1) => label=y
2) (CC_{history} >= 1) => label=y
```

```
3) (Relocation = y) and (FU >= 238) => label=y
4) (ContactPage = y) and (FU >= 50) and (Relocation = y) => label=y
5) (FU >= 15) and (SST_{DistinctViews} = 4) and (ContactPage = y) => label=y
6) (Relocation = y) and (FU >= 130) and (FU <= 155) => label=y
7) (SST_{DistinctViews} = 2) and (ContactPage = y) and (SSP_{history} >= 1) => label=y
8) (ConctactPage = y) and (SST_{DistinctViews} = 3) and (FU >= 57) => label=y
9) (Relocation = y) and (FU >= 45) and (FU <= 122) => label=y
10) (SST_{DistinctViews} = 2) and (ContactPage = y) and (FU >= 211) => label=y
11) (FU >= 72) and (SST_{DistinctViews} = 5) => label=y
12) => label=n
```

It is likely that this is not an optimal model for this particular topic. However, the focus of this thesis is not to obtain the best classification model but to show how to derive a segment definition from a classification model. Besides implementing the segment definition and analysing the data in the WA suite might yield insights that can be used to enhance the classification task to obtain the segment definition. Yet, this rule set shows the connection between web and cross channel usage. Several rules contain cross channel elements such as *Relocation* and $CC_{history}$.

Chapter 7

Conclusions

7.1 Contribution

Web Analytics[1] (WA) suites provide advanced visualization and reporting options on web site visitor behavior. Besides, certain forms of tracking cross channel usage behavior of the use of other channels *prior* to the use of the web channel. The required functionality for reporting on related call center use *after* web channel use are currently lacking. In this thesis a framework is presented to derive segment definitions for a WA suite based on customer cross channel usage behavior with respect to internet self service and call center use. This allows for a WA suite to serve as the central platform in the analysis of customer web usage behavior in relation to call center use.

The framework contributes to the business model of Adversitement B.V. (WA consultancy) for online business optimization. By deploying the framework, a model (in the form of a set of rules) is obtained that discriminates between successful and unsuccessful web usage. This model can be transformed into first order boolean logic (the formalism of WA suites to define segments). Besides, given this model, it is clear *which* additional data might have to be captured in the clickstream in order to be able to implement the segment definitions that holds data for unsuccessful web usage. These segments can be analyzed in isolation to determine why customers use the SSP and consecutively make a call. Understanding this behavior allows for taking appropriate

[1]WA: The objective tracking, collection, measurement, reporting, and analysis of quantitative internet data to optimize websites and web marketing initiatives.

actions to improve the quality of the SSP in order to reduce the volume of incoming self service related calls. As a result of this costs are saved. This framework can be deployed in any web site environment that hosts an SSP and from which clickstream data is collected by means of a WA solution.

7.2 Methodology summary

The framework proposed in this thesis provides the complete roadmap to derive segment definitions based on WA clickstream, call center and customer related data. The thesis points out the importance of the quality of the clickstream data that is used. In the setting of this thesis, the information need of a customer that underlies the clickstream is determined based on the page concepts that were viewed. For the context of this thesis, it is shown that WA clickstream data is more suitable for deriving the information need of web visitors (customers) compared to server-side collected data that is traditionally used in WUM.

The required preprocessing steps that are essential in order to obtain a data representation that is fit for this classification task are described in detail, step by step. This thesis shows the large number of aspects and issues that have to be taken into account with respect to data preprocessing. It involves labeling visits according to the information need that underlies the clickstream, mapping web concepts to a call center topic and strategies to aggregate data from a customer over multiple web visits and to link web usage to incoming calls. Moreover, features are derived and constructed from the clickstream data. This set of features is enriched with customer related data that is relevant to the specific self service related call center topic that is being analyzed. As a result a channel independent data representation is obtained that captures the information need a customer with respect to a topic and that is fit for the classification task.

Given the fact that first order boolean logic is the formalism to express segment definitions in a WA suite, the use of a rule classifier is motivated. A confusion matrix is used to evaluate the accuracy of the obtained model. The configuration of the classification task is focused at maximizing the number of true positives while at the same time minimizing the number of false positives. This is motivated by the fact that

a WA suite reports on aggregated data. Therefore it is best to have a segment that holds only data from unsuccessful web usage, although this might result in a data set that is not complete. To achieve this goal cost sensitive learning is used. A method to evaluate the performance of the segment that is implemented in a WA suite is proposed by means of a confusion matrix as well.

7.3 Future work

The framework addresses all required steps and alternative techniques are mentioned. However, it is highly unlikely that each of these steps is optimal for every arbitrary domain. Each step has to be fine-tuned to the specific setting of the domain to which it is applied. The actual implementation and evaluation of segment definitions in a WA suite are outside the scope of this thesis. Based on the evaluation of the performance of the segment it has to be determined whether to make adjustments to the segment definition or the deploy the framework again.

Web visits are considered to be related to a call center topic based on the presence of topic-related concepts. There is no weighting applied. Besides, for several customized concepts only elements from the actual content that is presented in the concept view of the customer are taken into account. E.g. for the case study that relates to payments and invoices, it is known for each customer what the numbers and amounts are that are being shown when a customer views the related concepts. However, this is not further addressed by the concept name. It is left for further research to address to what extent the information need is captured by the clickstream that is being generated and how it can be derived in combination with additional data. This includes concept relevancy weighing. Instead of the binary method that is used in this thesis, a function to compute and assign weights to a (set of) concept(s) that was/were visited to address the information need of customer in a particular topic more specific is desired.

The clickstream features proposed are constructed based on the little work that is available in literature on this particular topic. It is hard to indicate how and if web usage constructed features relate to e.g. demographics. No web site usability aspects are taken into account in the feature construction. Besides, features will be domain

and topic dependent as well as the values that are used in for replacing numerical attributes by categorical ones. Additional research on clickstream feature construction and extraction is required to evaluate the quality of the features that are proposed.

To achieve a cost saving, the expenses made to implement the collection of additional parameters in the clickstream that a required for a particular segment definition have to be in proportion to the costs saved (due to a reduced number of incoming calls). It is left for future work to develop a cost utility function that can be used in determining the *overall best* classification model to choose from which a segment definition is derived. Given a particular issue that has to be tackled, this function has to take into account the analytical insight gained and potential cost saving achieved on the one hand and the implementation cost of additional parameters in the clickstream on the other hand.

References

[1] Azreen Azman and Iadh Ounis. Discovery of aggregate usage profiles based on clustering information needs. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 470–471, New York, NY, USA, 2004. ACM. 44

[2] Arindam Banerjee and Joydeep Ghosh. Clickstream clustering using weighted longest common subsequences. In *in: Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining*, pages 33–40, 2001. 48

[3] Jayanta Basak and Sunil Goyal. Cross-channel customer mapping. In *ICIS '08: Proceedings of the Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008)*, pages 119–126, Washington, DC, USA, 2008. IEEE Computer Society. 49

[4] Ed H. Chi, Peter Pirolli, Kim Chen, and James Pitkow. Using information scent to model user information needs and actions and the web. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 490–497, New York, NY, USA, 2001. ACM. 44

[5] Vic Ciesielski and Anand Lalani. *Data mining of web access logs from an academic web site*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 2003. 49, 51

[6] William W. Cohen. Fast effective rule induction. In *In Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995. 76

[7] Robert Cooley. The use of web structure and content to identify subjectively interesting web usage patterns. *ACM Trans. Interet Technol.*, 3(2):93–116, 2003. 41, 46

[8] Smyth Padhraic Fayyad Usama, Piatetsky-Shapiro Gregory. The kdd process for extracting useful knowledge from volumes of data. In *Communication of the ACM*, volume 39, pages 27–34, November 1996. 59

[9] Kurt Fenstermacher and Mark Ginsburg. Client-side monitoring for web mining. *Journal of the American Society for Information Science and Technology*, 54:625–637, 2003. 51, 62

[10] Peter I. Hofgesang. Methodology for preprocessing and evaluating the time spent on web pages. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 218–225, Washington, DC, USA, 2006. IEEE Computer Society. 47, 49, 57

[11] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, 2002. 79

[12] Avinash Kaushik. *WA: An Hour a Day.* SYBEX Inc., Alameda, CA, USA, 2007. 14, 51, 55

[13] Michael Khoo, Joe Pagano, Anne L. Washington, Mimi Recker, Bart Palmer, and Robert A. Donahue. Using web metrics to analyze digital libraries. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 375–384, New York, NY, USA, 2008. ACM. 34

[14] Tomáš Kliegr. Representation and dimensionality reduction of semantically enriched clickstreams. In *Ph.D. '08: Proceedings of the 2008 EDBT Ph.D. workshop*, pages 29–38, New York, NY, USA, 2008. ACM. 42, 47, 48, 49, 61, 62

[15] Ron Kohavi. Mining e-commerce data: the good, the bad, and the ugly. In *KDD*, pages 8–13, 2001. 45, 51, 62

[16] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997. 79

[17] Chaofeng Li and Yansheng Lu. Similarity measurement of web sessions by sequence alignment. In *NPC '07: Proceedings of the 2007 IFIP International Conference on Network and Parallel Computing Workshops*, pages 716–720, Washington, DC, USA, 2007. IEEE Computer Society. 48

[18] Lin Lu, Margaret H. Dunham, and Yu Meng. Mining significant usage patterns from clickstream data. In *WEBKDD*, pages 1–17, 2005. 45, 46, 48, 51

[19] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008. 87

[20] Kate McCarthy, Bibi Zabar, and Gary Weiss. Does cost-sensitive learning beat sampling for classifying rare classes? In *UBDM '05: Proceedings of the 1st international workshop on Utility-based data mining*, pages 69–77, New York, NY, USA, 2005. ACM. 79

[21] Caren Moraes Nichele and Karin Becker. Clustering web sessions by levels of page similarity. In *PAKDD*, pages 346–350, 2006. 48

[22] Albert Orriols and Ester Bernadó-Mansilla. The class imbalance problem in learning classifier systems: a preliminary study. In *GECCO '05: Proceedings of the 2005 workshops on Genetic and evolutionary computation*, pages 74–78, New York, NY, USA, 2005. ACM. 79

[23] Cyrus Shahabi and Farnoush Banaei Kashani. A framework for efficient and anonymous web usage mining based on client-side tracking. In *WEBKDD '01: Revised Papers from the Third International Workshop on Mining Web Log Data Across All Customers Touch Points*, pages 113–144, London, UK, 2002. Springer-Verlag. 51, 62

[24] Abraham Silberschatz, Henry F. Korth, and S. Sudarshan. *Database Systems Concepts*. McGraw-Hill Science/Engineering/Math, fourth edition, October 2001. 65

[25] Tom Soukup and Ian Davidson. *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*. John Wiley & Sons, Inc., New York, NY, USA, 2002. 85

[26] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000. 46, 51

[27] Jaideep Srivastava, Prasanna Desikan, and Vipin Kumar. Design of distributed data mining applications on the knowledge grid. In *Proc. of the National Science Foundation Workshop on Next Generation Data Mining (NGDM'02)*, pages 191–195, Baltimore, Maryland, November 2002. 45, 53

[28] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining.* Addison Wesley, May 2005. 19, 25, 26, 27, 36, 46, 47, 54, 55, 72, 75, 76, 77, 78, 120

[29] Jianyong Wang, Yuzhou Zhang, Lizhu Zhou, George Karypis, and Charu C. Aggarwal. Discriminating subsequence discovery for sequence clustering. In *SDM*, 2007. 48

[30] Weinan Wang and Osmar R. Zaïane. Clustering web sessions by sequence alignment. In *In Proceedings of the 13th international workshop on database and expert systems applications (DEXA 2002). Aix-en-Provence*, pages 394–398. Springer-Verlag, 2002. 48

[31] Gary M. Weiss and Foster Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003. 79

[32] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005. 19, 42, 76, 79, 92, 100

Appendix A

Web analytics terms and definitions

A.1 Definitions

Web analytics is a rapidly evolving area. Recently definitions and standards were proposed by the Web Analytics Association. Below is an overview of terms and definitions used in this thesis (source: http://www.webanalyticsassociation.org)

There are two types of Web analytics metrics counts and ratios:

- **Count:** The most basic unit of measure; a single number, not a ratio. Often a whole number (Visits = 12,398), but not necessarily (Total Sales= $52,126.37.). Some metrics cannot be summed across time and/or within a report. See metric definitions for specific limitations.
- **Ratio:** A derived metric, obtained by dividing one number by another. The result is usually not a whole number. Because its a ratio, per is typically in the name, such as Page Views per Visit. Most ratios used in web analytics are not summable. Another type of definition is included for terms that describe concepts instead of numbers.
- **Dimension:** A component or category of data. Metrics (counts and ratios) are measured across dimensions.

All metrics can apply to three different universes:

- **Aggregate:**Representative of the entire site.
- **Segmented:** A subset of the site traffic for a defined period of time, filtered in some way to gain greater analytical insight: e.g., by campaign (e-mail, banner, PPC, affiliate), by visitor type (new vs. returning, repeat buyers, high value), by referrer.
- **Individual:** Activity of a single Web visitor for a defined period of time.

Building Block Terms: Building block terms include four main metrics, Unique Visitors, Visit/Sessions, Page Views, and Events that make up the foundation for all web measures. These measures can be used either as a unique value by themselves or as the denominator within various formulas. The following definitions are provided as infrastructure on which to build upon.

Page

- Type: Dimension
- Calculation: An analyst definable unit of content.

Page View

- Type: Count
- Calculation: The number of times a page was viewed.

Visit (or Session)

- Type: Count
- Calculation: A visit is an interaction, by an individual, with a web site consisting of one or more requests for a page. If an individual has not taken another action (typically additional page views) on the site within a specified time period (by default 30 minutes), the visit will terminate by timing out.

Unique Visitors

- Type: Count item Calculation: The number of inferred individual people (filtered for spiders and robots), within a designated reporting timeframe, with activity consisting of one or more visits to a site. Each individual is counted only once in the unique visitor measure for the reporting period.

- Note: Authentication, either active or passive, is the most accurate way to track unique visitors. However, because most sites do not require a user login, the most predominant method of identifying unique visitors is via a persistent cookie that stores and returns a unique id value, introducing inaccuracies from cookie deletion, shared computers, browsing from multiple browsers or computers, etc.

Event

- Type: Dimension and/or Count
- Calculation: Any logged or recorded action that has a specific date and time assigned to it by either the browser or server.
- Note: Events are activities that happen within a page, for example: ad impressions, starting and completing transactions, changing form fields, starting multimedia views, etc. Events can also be associated with advanced web technologies, such as Ajax and Flash. Because an Event can be both a dimension and a count, a web analytics report may show number of events (event as a count), or it may show specific events and how many page views, visits, or unique visitors were associated with the events (event as a dimension).

Appendix B

Adversitement B.V.

Adversitement is a full service e-marketing consultancy specializing in web analytics and customer intelligence. With a proven experience in over 100 implementations, Adversitement assists organizations in the entire process, from technical implementation to optimization, supported by training (Omniture Authorized Training Centre) with the goal to increase turnover. Besides the quantitative research we have the tools and knowledge to do qualitative research. This combination of highly developed knowledge, experience, technical innovation and specifically selected (self-developed) products such as KPI portals, make Adversitement unique in its market and a web analytics partner in Europe.

B.1 Core activities

- Increase visitor retention and turnover
- Optimize online business and search engine performance
- Improve campaign performance
- Test website usability
- Manage keyword campaigns
- Web analytics training
- Integrate external data sources and make data more comprehensible

Appendix C

Web analytics page script excerpt

Below is an excerpt from the Omniture HBX Analytics script to collect the required data elements from each web page.

```
<script>
var _hbEC=0,_hbE=new Array;
function _hbEvent(a,b){b=_hbE[_hbEC++]=new Object();b._N=a;b._C=0;return b;}
var hbx=_hbEvent("pv");hbx.vpc="HBX0200u";hbx.gn="GATEWAYNAME";

//BEGIN EDITABLE SECTION - CONFIGURATION VARIABLES
hbx.acct="HBX+ACCOUNT+NUMBER";      //ACCOUNT NUMBER(S)
hbx.pn="PUT+PAGE+NAME+HERE";        //PAGE NAME(S)
hbx.mlc="CONTENT+CATEGORY";         //MULTI-LEVEL CONTENT CATEGORY

//OPTIONAL PAGE VARIABLES - ACTION SETTINGS
hbx.fv="";                  //FORM VALIDATION MINIMUM ELEMENTS OR SUBMIT FUNCTION NAME
hbx.lt="auto";              //LINK TRACKING
hbx.dlf="n";                //DOWNLOAD FILTER
hbx.dft="n";                //DOWNLOAD FILE NAMING

//SEGMENTS AND FUNNELS
hbx.seg="";             //VISITOR SEGMENTATION
hbx.fnl="";             //FUNNELS
:

//CUSTOM VARIABLES
hbx.ci="AB-45-FTG-321";     //CUSTOMER ID
hbx.hrf="";                 //CUSTOM REFERRER
hbx.pec="";                 //ERROR CODES
:
</script>
<script language="javascript1.1" src="hbx.js"></script>
```

Appendix D

Clickstream XML feed excerpt

Below is an excerpt from the XML feed that holds the clickstream data elements.


```
<!DOCTYPE SESSIONS SYSTEM "xml-datafeed-v1-2.dtd">
<SESSIONS>
    <SESSION>
        <INFO>
            <SESSION_ID>1191212355530448</SESSION_ID>
            <VISITOR_ID>1191212355530448</VISITOR_ID>
        </INFO>
        <EVENTS>
            <EVENT>
                <TIMESTAMP>2007-10-01 04:19:15 GMT0</TIMESTAMP>
                <VIEW>
                    <PAGENAME>home</PAGENAME>
                    <CONTENT>/</CONTENT>
                </VIEW>
            </EVENT>
            <EVENT>
                <TIMESTAMP>2007-10-01 04:19:24 GMT0</TIMESTAMP>
                <LINK_CLICKED>
                    <ID>Meterstanden doorgeven</ID>
                    <VIEW>
                        <PAGENAME>home</PAGENAME>
                        <CONTENT>/</CONTENT>
                    </VIEW>
                </LINK_CLICKED>
            </EVENT>
            <EVENT>
                <TIMESTAMP>2007-10-01 04:19:25 GMT0</TIMESTAMP>
                <VIEW>
                    <PAGENAME>meterstanden doorgeven</PAGENAME>
```

```
                <CONTENT>/klantenservice/meterstandendoorgeven</CONTENT>
            </VIEW>
        </EVENT>
        :
        :
        :
        <EVENT>
            <TIMESTAMP>2007-10-01 04:23:21 GMT0</TIMESTAMP>
            <EXIT_LINK>http://www.nuon.nl/klantenservice</EXIT_LINK>
        </EVENT>
        <EVENT>
            <TIMESTAMP>2007-10-01 04:23:24 GMT0</TIMESTAMP>
            <VIEW>
                <PAGENAME>klantenservice</PAGENAME>
                <CONTENT>/klantenservice</CONTENT>
            </VIEW>
        </EVENT>
    </EVENTS>
    <SUMMARY>
        <LAST_HIT_TIMESTAMP>2007-10-01 04:23:24 GMT0</LAST_HIT_TIMESTAMP>
        <FIRST_HIT_TIMESTAMP>2007-10-01 04:19:15 GMT0</FIRST_HIT_TIMESTAMP>
        <REFERRING_URL>^b</REFERRING_URL>
        <BROWSER>MSIE 7.0</BROWSER>
        <OPERATING_SYSTEM>Windows XP</OPERATING_SYSTEM>
        <JS_VERSION>15</JS_VERSION>
        <BROWSER_HOUR>6</BROWSER_HOUR>
        <TIME_ZONE>-120</TIME_ZONE>
        <LANGUAGE>nl</LANGUAGE>
        <SCREEN_RESOLUTION>1024*768</SCREEN_RESOLUTION>
        <JAVA_ENABLED>y</JAVA_ENABLED>
        <CONNECTION_TYPE>lan</CONNECTION_TYPE>
        <HOST_DOMAIN>ip.tiscali.nl</HOST_DOMAIN>
        <CITY>Amsterdam</CITY>
        <CUSTOMER_ID>AB-45-FTG-321<CUSTOMER_ID>
        <COUNTRY>NL</COUNTRY>
    </SUMMARY>
    <VISITOR_PROFILE>
        <SITE_VISITS>1</SITE_VISITS>
        <FIRST_VISIT_TIME>2007-10-01 04:19:15 GMT0</FIRST_VISIT_TIME>
    </VISITOR_PROFILE>
</SESSION>
:
:
</SESSIONS>
```

Appendix E

Short note to AJAX technology

Figure E.1 shows an overview of a AJAX based implementation for validating a user authentication request. In case of an invalid input, the user is notified. In such case, the page layout is changed. Depending on the level of granularity, this could be considered to be a *new* page. However, there is no request sent to the web server, so no new page request is registered. The implementation of these AJAX techniques is common practice in web development. With a web analytics solution, a dedicated image request can be sent with a specific page name. As a results, a richer set of page views is obtained that addresses the actual web usage more specifically.

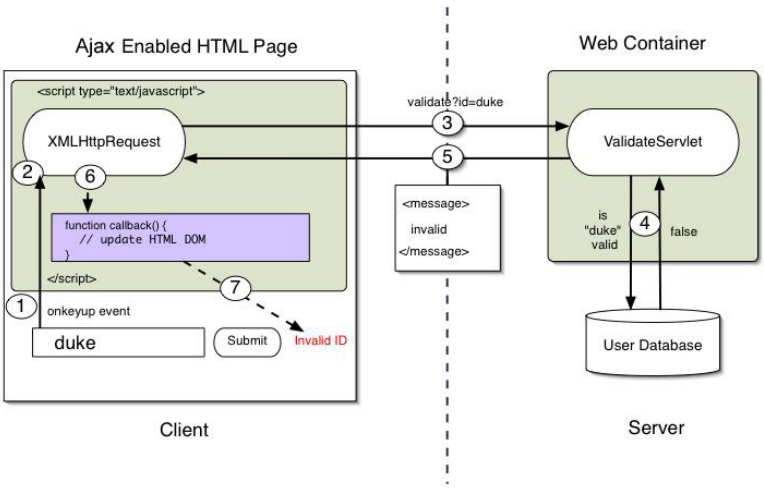


Figure E.1 – AJAX request overview. Figure adopted from http://java.sun.com/developer/technicalArticles/J2EE/AJAX/

Appendix F

Comparison: e-commerce versus self service context

Several works on Web Usage Mining in literature are set in an e-commerce contexts. Data mining is used to discover the differences in web usage behavior between *buyers* and *non-buyers*. Certain aspects apply to the context of thesis. The class imbalance problem holds for both. In an e-commerce setting, *buyers* that form the majority class versus *non-buyers* that form the minority class. In the self service context of this thesis, customers with *successful web usage* form the majority class versus customers with *unsuccessful web usage* that form the minority class. However, there are important differences make it impossible to directly apply WUM strategies developed in one context to the other. Figure F.1 gives an overview of the flow in both contexts of the minority class.

Below is an overview of the differences:

- In the e-commerce environment, all actions occur in the same channel. All actions are recorded in the same format. As pointed out throughout the thesis, actions are related over multiple channels.

- In the e-commerce environment, there is a unique *conversion point* (= order confirmation page) that demarcates the completion of the trajectory from *searching & browsing* to *making a purchase*. This conversion point is *predefined* by the host of the site. To a certain extent, a phone call can be considered to be the conversion point. However, to relate the web *searching & browsing* to a call is a hard task. A customer determines the conversion point (= point in time at which

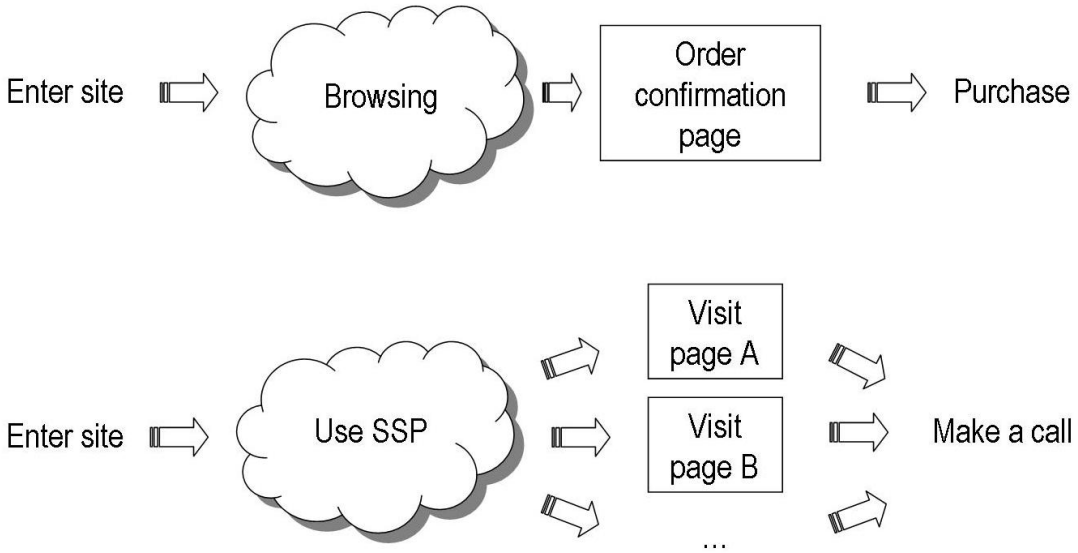


Figure F.1 – The top flow depicts a purchase in an e-commerce environment. The bottom flow depicts a call after SSP usage.

is decided to make a call) her/himself.

- A purchase is the result of completing all required steps within the scope of a single visit. A visit is a clear defined concept that allows for reasoning about concepts that were related to the purchase. It is likely that the sequence of concepts that were viewed prior to the purchase contributed to this process. In the self service environment, time elapses between browsing and calling.

- The e-commerce environment deals with *expenses* while the self service environment deals with *costs*. This means a difference in the motivation people have to use the related services.

Appendix G

Note to sequence mining

As this thesis is set in the context of WUM, it would seem fit for *sequential pattern mining* [28]. However, experiments with this technique, in which the sequences of page views were the input, did not yield useful results. This is probably due to characteristics of the environment that is being analyzed. Visits are characterized by a relatively low number of (distinct) pages viewed per visit. Besides, customized content is being presented in concepts. Customers that visit a particular concept are shown different overviews. Algorithms currently do not allow for parameters to be configured to adjust for a particular setting. Therefore it is required to offer the data such that the output of the algorithm is useful. There are no clear strategies formulated yet, given a particular domain. The example below is meant to illustrate the complexity of such a task given clickstream data generation.

Example 11 (Sequence mining) *The table shows a characteristic pattern for web usage that is related to the annual account. The concept* /personal/my annual account *shows for each customer different content. This specific sequence of page views that constitutes a single visit and repetition of this visit in time is contained in both successful as well as unsuccessful web usage with respect to call center topic* annual account*. This means that there are identical sequences with an completely opposite outcome as far as the call center channel usage is considered.*

Moreover, the output obtained from sequence mining would have to be transformed in an appropriate format as segments have to formulated in first order boolean logic. It would require implementation of enhanced scripting that is currently not supported as a standard.

Page view sequence	
Time	**Page viewed**
T_1	/home
T_2	/login
T_3	/personal/mijn nuon overview
T_4	/personal/my annual account

Appendix H

MySQL data import application

In the preparation for this thesis, I developed a application that automatically exports relevant data elements from the XML feed to MySQL data base. The motivation to use MySQL was because of two reasons:

1. A relational data base was going to be used to process and transform the input data.
2. A MySQL data base is freely available and its functionality supported well by the open-source community.

The image below shows a screenshot of the application. It allows for specifying the required data base parameters to establish the connection and the export. A dedicated table has to be created separately in the data base prior to the use of this application. Further on, it allows for the selection of specific data entries. Currently, it supports five different types of entries. These are all used in the experiment section in the thesis. The progress bar and the output window provide feedback on the process to the user. It includes error reporting to hint the user in case of malformed input.

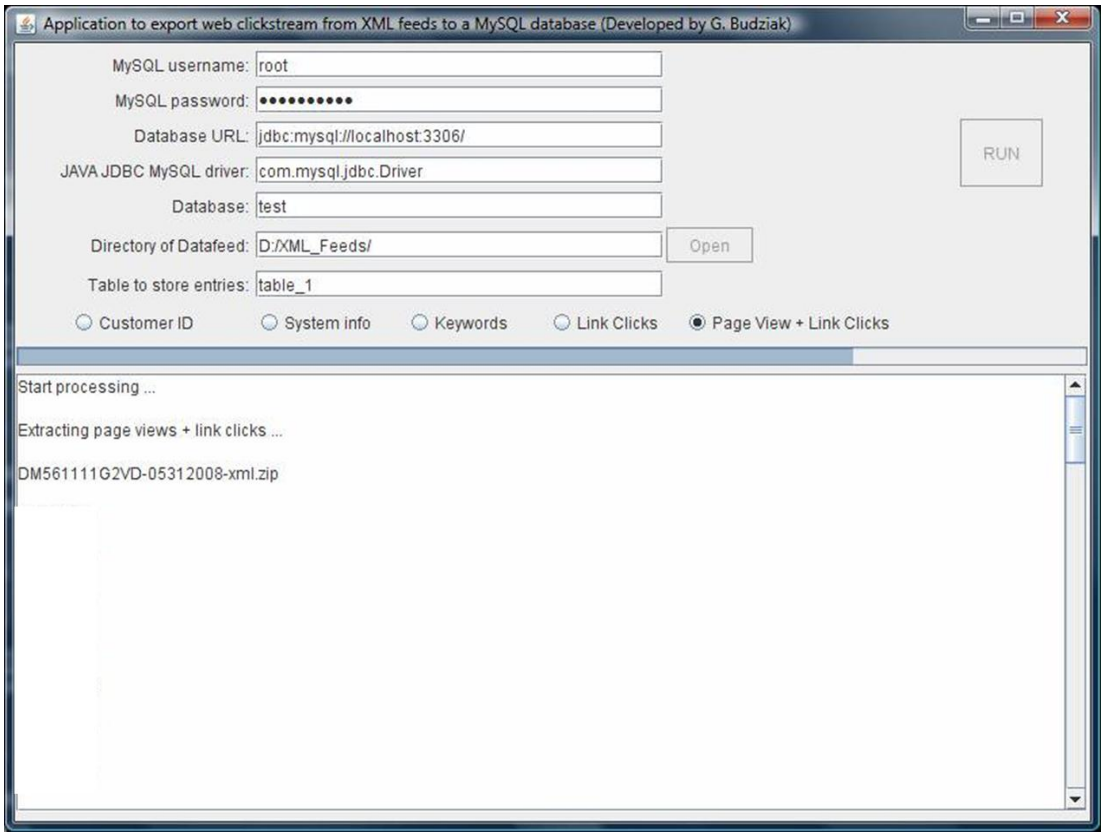


Figure H.1 – Java applet to export clickstream data stored in XML format to a MySQL data base.

Appendix I

Tables

This Appendix shows the table structures that are used in this framework and that are described in Chapter 4.

I.1 Page View Table

- Normalized table
- *(Session ID, Timestamp)* is the primary key
- Holds data related each page view event
- The table is ordered by session ID
- The sequence of page views is resembled by the table implementation such that for each pair of consecutive records r_i, r_{i+1} with the same session ID holds $T_i < T_{i+1}$

Page View Table		
Entity	Encoding	NULL
Session ID	VARCHAR(16)	NO
Session Row ID	SMALLINT	NO
Customer ID	VARCHAR(14)	NO
Page Name	VARCHAR(200)	NO
Page ID	SMALLINT	NO
Timestamp	DATETIME	NO
Time Spent	SMALLINT	YES

I.2 Page info table

- Normalized table
- Holds data related to pages
- *(Page Name)* is the primary key

Page Info Table		
Entity	Encoding	NULL
Page ID	SMALLINT	NO
Page Name	VARCHAR(200)	NO

I.3 Topic event definition table

- Maps call center topics to web concepts

Topic Event Definition Table		
Entity	Encoding	NULL
Topic	VARCHAR(14)	NO
Pages	VARCHAR(50)	NO
Forms	VARCHAR(50)	NO
FAQ	VARCHAR(50)	NO

I.4 Call center table

- Normalized table
- Holds data related incoming MijnNUON registered customer calls

Call Center Table		
Entity	Encoding	NULL
Customer ID	VARCHAR(14)	NO
Topic	VARCHAR(50)	NO
Issue	VARCHAR(50)	NO
Channel	VARCHAR(50)	NO
Timestamp	DATETIME	NO

I.5 FAQ Table

- Holds data related to the FAQ links clicked
- Links are mapped to a topic
- *(Session ID)* is the primary key

FAQ Info Table		
Entity	Encoding	NULL
Session ID	VARCHAR(16)	NO
Topic	VARCHAR(30)	NO
FAQ Answer	VARCHAR(200)	NO
Times Viewed	SMALLINT	NO

I.6 Form info table

- Summary table
- Holds data related to the form usage during a session. It gives the number of steps in the form and the level of completion during the visit
- *(Session ID)* is the primary key

Form Info Table		
Entity	Encoding	NULL
Session ID	VARCHAR(16)	NO
Form	VARCHAR(30)	NO
Number of Steps	SMALLINT	NO
Maximum Step reached	SMALLINT	NO

I.7 Payment info table

- Normalized table
- Holds data related to payment history (term account, annual account) of each customer

Payment Info Table		
Entity	Encoding	NULL
Customer ID	VARCHAR(14)	NO
Topic	VARCHAR(30)	NO
Amount	DOUBLE	NO
Arrival Date	DATE	NO
Payment Due Date	DATE	YES

I.8 Demographic table

- Normalized table
- Holds data related to customers demographic characteristics
- Many NULL values unfortunately
- *(Customer ID)* is the primary key

Demographic Info Table		
Entity	Encoding	NULL
Customer ID	VARCHAR(14)	NO
Sex	VARCHAR(1)	YES
Date of Birth	DATE	YES
MijnNUON registration date	TIMESTAMP	NO

Declaration

I herewith declare that I have produced this paper without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This paper has not previously been presented in identical or similar form to any other Dutch or foreign examination board.

The thesis work was conducted from February 2008 to November 2008 under the supervision of Dr. M. Pechenizkiy from the Department of Mathematics and Computer Science, Eindhoven University of Technology.

Eindhoven, November 2008