TOWARDS HYBRID CLUSTERING FOR B2C CUSTOMER SEGMENTATION: CONCEPTUAL FRAMEWORK

MASTER THESIS INFORMATION SCIENCES

Almere, August 25, 2015

Author

Leon à Campo, BSc

Student nr: s4499158

Email: leonacampo@gmail.com

Supervisors

Radboud University Nijmegen: Prof. dr. ir. Th. P. van der Weide

Avanade: Sebastiaan Meloen-Noorda, MSc & Gijs Ramaker, MSc

avanade®
Results Realized

Radboud Universiteit

Abstract

An increasing amount of clients of Avanade were requesting more advanced types of analytics to analyze their customer data, like customer analytics. Currently a technique exists that is called social engagement, formerly known as social listening, that is part of Microsoft Dynamics CRM, which gives basic social insights into trends on the internet.

A more advanced type of customer analysis technique than this was required, namely, a service that would have to be able to combine data found on social networks and other online sources with existing historic customer data. Preferably this was done in such a way that marketers can focus on one type of segmentation, instead of both customer segments based on historic customer data and social trends that arise in the social engagement service.

There has been made a choice to focus on business to consumer (B2C) customer segmentation. The reason being that certain required attributes only exist for individual customers. An extensive literature research on existing customer segmentation solutions and classification techniques for social activity data was performed which resulted in a conceptual framework. This framework organizes and defines the concepts that are involved with hybrid clustering. This is further supported by mostly recent literature on segmentation and classification techniques.

So-called archetypes were created in order to create shared attributes between both forms of segmentations. They are abstractions of the information that is found in the historic customer data and social activity data. In contrast to customer profiles, archetypes consist of additional data that is the result of advanced techniques like machine learning and data mining.

Acknowledgements

This document represents my master's thesis as partial fulfilment of my master (MSc) in Information Sciences at the Radboud University Nijmegen. It proves that I can individually conduct research at master level by analyzing a complex problem and provide an innovative solution for this problem. This has been done through a six month external research project at Avanade. Meaning that the results presented in this thesis must be acceptable for both the Radboud University Nijmegen and Avanade.

First I want to thank my supervisor at Avanade, Sebastiaan Meloen-Noorda, who would reflect weekly on all the progress I had made. This close supervision really kept me motivated and resulted in new ideas and improvements every time. Sebastiaan really put in a lot of effort into giving me in-depth feedback each time.

Besides my supervisor at Avanade, I would like to thank Theo van der Weide, who was able to critically review my work in order to get my ideas more concrete and helped me to create more formalized descriptions of the various concepts.

I am also grateful that Avanade gave me the opportunity to conduct my research at their company. The positive atmosphere and fellow interns made it a very pleasant experience and I would like to thank Gijs Ramaker for his helpful workshops and reflection sessions.

Lastly, I would like to sincerely thank everyone who has reviewed my thesis or helped me during the process of concretizing my ideas.

Almere, August 25, 2015

- Leon à Campo

List of abbreviations

Abbreviation	Description
API	Application Programming Interface
B2B	Business to Business
B2C	Business to Consumer
BFS	Breadth First Search
CRM	Customer Relationship Management
DFS	Depth First Search
GIS	Geographic Information System
GPS	Generalized Sequential Pattern
IDF	Inverse Document Frequency
KDT	Knowledge Discovery from Text
LDA	Latent Dirichlet Allocation
LTV	Life Time Value
PSO	Particle Swarm Optimization
SOM	Self-Organizing Map
SVM	Support Vector Machine
TF	Term Frequency
UML	Unified Modeling Language

List of symbols

Symbol	Description
Supervised learning 2.3.3.2	
T	Training data set
x_n	Predictor variable
l_n	Label of predictor
X	Input space
Y	Output space
γ	Target labeling function
Z_n	Class of document
Q	Document
Disjoint sets 4.2.1	
A	Normal set
B	Normal set
x	Member of set
Historic 4.3.1	
C	Set of customers
c	Customer
A	Attribute
V	Value
Social 4.3.2	
S	Social activity data
D	Post
t	Terms in post
Profile 4.3.3	
P	Customer profiles
Archetypes 4.4	
H	Archetype historic customer
G	Archetype social activity
Clustering 4.5	
O	Objects
W	Subset
K	Clusters
E	Set of centroids
F	m, n-dimensional points
Dynamic segments 4.7	
Ω	Attribute space
Customer segmentation 5.3.1	
L	Output SOM
I	Input SOM
Archetype attributes 5.6.2	
ϑ	Sentiment value
ρ	Age value

List of figures and tables

Figures

Tables

Table of contents

1 Research introduction

1.1 Context

Avanade is a company that was founded in March 2000 in the US as a joint venture between Accenture and Microsoft (Avanade history, 2015). Avanade employs more than 22.000 professionals to date and has seen a yearly average growth of 20% since 2000. This business technology services firm focuses on the implementation of Microsoft enterprise solutions and services. Accenture, Avanade and Microsoft form an alliance which allows them to benefit from each other's skills, knowledge and experience (Microsoft alliance, 2015).

An increasing amount of clients of Avanade are requesting more advanced types of analytics to analyze their customer data, like customer analytics. Customer analytics meaning extensive use of data, models and fact-based management that can be used to enable more precise customer segments and help drive changes and improvements in business practices (Bijmolt et al., 2010). This allows the clients of Avanade to improve their ability to target the highest-value opportunities. Advanced analytics is becoming increasingly important to improve performance within a company and guides the facilitation of analytic tools that provide the required information to support the most important business processes (Waller & Fawcett, 2013).

1.2 Problem statement

Currently a technique exists that is called social engagement, formerly known as social listening, that is part of Microsoft Dynamics CRM, which gives basic social insights into trends on the internet (Gartner, 2014). This insight is achieved by scanning the public content of social network for specific company- or domain related keywords in multiple languages. It also analyses the context in which these keywords are used to give an indication about whether the sentiment in which these words are used is positive or negative (Microsoft Social Listening for CRM, 2015). One disadvantage of this service is that it provides the user only with limited content of the posts related to the keywords, which can lead to limited added value, when compared to manually monitoring social channels.

The aforementioned need for a more advanced type of customer analysis technique requests a service which would have to be able to combine data found on social networks and other online sources with existing historic customer data. Preferably this will be done in such a way that marketers can focus on one type of segmentation, instead of both customer segments based on historic customer data and social trends that arise in the social engagement service.

Companies generally apply some type of segmentation to their customer base. Dividing the customer base into homogenous groups enables them to deploy different marketing campaigns according to the characteristics of that group (Tsiptsis & Chorianopoulos, 2010). These segments can be formed by manually entering queries to create homogenous groups based on only one attribute. Another query can create further division, making the segments smaller and more specific. This process can also be done by computer algorithms, which allows to find homogenous groups in an automatic way. This can result in very interesting boundaries and relations between customer segments. It has also been shown that segmentation based on customer's geographic, demographic and behavioral attributes is considered to be a better approach than the use of only one attribute (Kandeil, Saad, & Youssef, 2014).

The combination of historic and social customer data would allow companies to be able to develop more elaborate and detailed profiles of their customers. The historic information would be provided by the data which the company has collected from their customers in a direct way. This can range from basic personal information like gender and birth date to data of purchased products or usage data of a certain service. In addition to this, the social activity data can be extracted from desired online sources. A selection of the typical sources of social data consist of social networks like Facebook, LinkedIn, Google+ or microblogs like Twitter (Russell, 2011). It can also include regular web pages, blogs and forums (F. Liu, Liu, & Weng, 2011).

To understand the need for more advanced customer analytics, various related solutions are listed in Table 1. These are solution that companies are providing to their clients at the moment. To get a better understanding of these solutions and the data that is used by those services, Table 1 lists the information that was found on their company websites. This list only captures current commercial implementations and does not mention any solutions that might be proposed in research papers. As seen in the table, the identified services are not able to combine different sources of data to segment customers on, let alone being able to combine different segmentations.

Name	Data source	End result
Customer Analytics Engine (Customer Analytics Engine, 2015)	Data from customer behavior	Market segments and predictions
Customer Data as a Service (IBM Unites Marketing, 2015)	Internal and external historic customer data	Complete set of historic customer data
Big Data Analytics as a Service (Big Data Analytics as a Service, 2015)	On premise data, processed by Hadoop	Better understanding of customer journey
Social Media Analytics Software as a Service (Social Media Analytics Software as a Service, 2015)	Social media posts	Customized results in configurable charts and dashboards
Analytics as a Service (Analytics as a Service, 2015)	Historical data and associated events (on premise)	Electronic feed containing forecasts and the driving factors that make up these forecasts
Predictive Analytics as a Service (Customer churn for SMBs using Predictive Analytics As a Service, 2015)	Input from questionnaire	Predictive models that can be used to segment customers and products

Table 1. Existing and related services to customer analytics

Additional to the solutions found on the internet, companies suggest the following benefits apply when customer analytics is implemented:

- "Targeting marketing optimizes market spending, drive loyalty programs and customer marketing decisions, bolster loyalty across multiple channels, develop customer experiences that promote brand affinity" (Customer Analytics Engine, 2015)
- "Identify customer trends, anticipate future behavior, suggest next-best actions" (IBM Unites Marketing, 2015)

1.3 Aim of paper

There has been made a choice to focus on business to consumer (B2C) customer segmentation. The reason being that certain required attributes only exist for individual customers. There's also the added difficulty that companies mainly use social channels to communicate with their customers in order to improve customer engagement, instead of expressing themselves regarding brands or products (Rybalko & Seltzer, 2010)From a research perspective this paper will fulfil the need for research on the concept of merging two or more customer segmentation techniques in a unitary structure (Hiziroglu, 2013).

In the review of state-of-the-art soft computing applications in customer segmentation, Hiziroglu (2013) argues that soft computing, as a family of data mining techniques, has been recently started to exploit the area of customer segmentation and that it is able to shape the future of segmentation research. He concludes that there is a need for a methodological framework that gives insight in how to combine different clustering or classification techniques in an appropriate and simplified way. This will allow business and management researchers to better understand and apply the methods that are used in this area of information and computer science. Aside from the technique itself

1.4 Approach

1.4.1 Main and sub questions

In order to provide a solution for the problem explained in the problem statement the following main question will be answered. *How should enterprises implement customer analytics to be able to segment customers in a more detailed way based on both historical customer data and social activity data in order to deploy more focused B2C campaigns?*

Because this main question touches a lot of different facets of customer analytics, the sub questions found below will be answered in order to divide the problem in logical parts.

1. How can different segmentations be combined?
2. Does the combination of segmentations based on historical customer data and social activity data provide more detailed customer segmentations?
3. How will the combination of historical customer data and social activity data enable more focused B2C campaigns?

The following information will be explored to find answers for these questions:

- Literature study on data mining and related fields e.g. text mining, information retrieval and natural language processing
- Define and explore what is meant with a campaign or marketing campaign
- Define what data is and what types of data exist
- Define the requirements for the solution
- Explore related literature to find the most suitable techniques

1.5 Related work

According to Josiah et al. (2015) customer analytics can be seen as the base analytic that is used to analyze the customer knowledge base. As a result it provides a better view on customer behavior and assesses customer values and the customer's portfolio or profiles. Josiah et al. (2015) also propose a CRM system called Analytical CRM, which in contrast to operational CRM focuses on supporting organizational back-office operations and analysis. It is designed to analyze the customer's data to better understand the customer's behavior. Two of the key features of Analytical CRM that are described in the paper consist of:

1. Utilizing customer data from various sources and integrating this into a central repository knowledge base.
2. Determining, developing and analyzing inclusive set of rules and analytical methods to scale and optimize relationship with customers.

Although the paper of Josiah et al. (2015) does not go into detail in the exact data sources and analytical techniques that should be used to acquire an Analytical CRM, it contains some key ideas that can form the basis for the combination of different segmentations.

In the paper by Saarijärvi, Karjaluoto and Kuusela (2013) they claim that the attention of companies is shifting from selling products towards serving customers. This would require change, since traditional CRM activities are aimed at selling more products instead of serving customers. To solve this problem the role of customer data within the CRM framework has to be reconfigured.

Through the means of both structured and unstructured interviews the paper of Saarijärvi, Karjaluoto and Kuusela (2013) identifies that there is a need for customer data that is used for the benefit of the customers. This approach can be part of traditional CRM and will enable new opportunities for developing customer relationships.

It will also prevent customers from becoming suspicious towards the usages of customer data because of unethical behavior of certain companies in the past. Since this study is mainly preliminary, further research is needed to identify real-world applications and business concepts in which the traditional role of customer data is being challenged and will contribute to customer data sharing. This would provide better understanding about the shift in customer data usage. Empirical evidence about the problems and obstacles that will occur when implementing this reconfigured CRM can serve as a basis to determine the potential of this concept.

Focusing on current customer analytics that is being applied by companies, the paper of Germann, Lilien, Fiedler and Krausd (2014) identifies that many companies do not perceive the potential gain of customer analytics and also do not invest in it at an economically appropriate level.

They analyzed survey data from 418 top managers based in the Americas, Europe Middle East and Africa (EMEA) and Asia to get inside information on this topic. These companies think the benefits will not outweigh the costs and effort that will go into customer analytics. The paper does not provide more detailed information on why some companies think this is the case. It would be valuable to know on what information these decisions are based and how to convince them in investing more in this technique.

When looking at all the possible industries in which customer analytics can be applied, the paper of Germann et al. (2014) concludes that retailers seem to benefit the most from increases in deployment of customer analytics. They also state that an increase in customer analytics is associated with higher firm performance, but they cannot make direct causal claims regarding this relationship.

1.6 Scenario hybrid clustering

To illustrate how hybrid clustering for B2C customer segmentation could work in practice, a scenario is shown below. This scenario will show the relevance of this technique within a business context. Hereby the problem will become clearer and increases the ease of understanding the topics that will be introduced throughout this paper.

The telecommunications company TelePhone is interested in using social activity data in a structured manner during the segmentation process. Their market has become saturated, making winning customers increasingly hard. Aside from this, winning new customers is more expensive than retaining existing ones (Rehman & Ali, 2014). At the moment TelePhone mainly focuses on behavioral segmentation by utilizing the historic customer data that is available inside the company. Demographic data is used for more rough segmentation.

Behavioral data consists among others things of information on usage and loyalty status. Telco's have easy access to detailed information on the usage of their services, since most customers are charged based on the usage of monthly bundles. Demographic data consists mainly of more general information that is obtained during the registration at the company. It does, however, contain important information on age and residence. In addition to this additional demographic information like income category can be added.

Because of the ever increasing amount of personal information that people post on the internet, it is very interesting to utilize this information to further segment customers based on social activity data. Posts in social channels can contain positive, negative or neutral information on the telco, which can be used to retain customer or win new ones.

TelePhone tried to take the social activity data into account during the process of setting up new marketing campaigns. Their marketing division found it hard to make a logical combination between the observation they made in the social activity data and the customer segments they acquired from the historic customer data.

Implementing hybrid clustering for B2C customer segmentation will allow TelePhone to directly address both historic customer data and social activity data during the process of setting up marketing campaigns. As the social activity data will now be combined with the historic customer data in hybrid clusters.

Section 1. Scenario hybrid clustering

1.7 Thesis outline

Chapter 2

This chapter introduces the main concepts and relevant background literature. The main topics that are discussed are customer analytics, data sources, data mining, text mining, information retrieval and natural language processing.

Chapter 3

This chapter explains the requirements which hybrid clustering for B2C customer segmentation has to meet. The requirements are compared against an existing system that aims to give insight in social activity. In addition to this, it will also explain the general overview of the process.

Chapter 4

This chapter presents the conceptual framework that aims to organize all concepts related to the problem in order to make it more understandable and manageable. In addition to this it also maps the relationships among the concepts. It will go over topics like segmentation, customers, archetypes, clustering, representation of clusters and dynamic segments.

Chapter 5

This chapter aims to provide more detail about the concepts that are explained in chapter 4. It provides a UML diagram which shows all the individual processes that are part of the total process. All processing steps are explained on the basis of papers that have presented solutions with similar goals. These papers could be used in further research that aims to add more detail to the process.

Chapter 6

This chapter is required, as it critically compares the requirements with the presented solution. It aims to validate the results of this paper. This validation is split into three steps, which also were the three main requirements stated in chapter 3.

Chapter 7

This chapter shows the conclusion of this thesis, which consists of the findings and the presented solution. It will also show the shortcomings of this solution.

Chapter 8

This chapter is mainly a reflection on the complete process of this thesis. It also helps to point out which parts of the solution require more research or certain aspects that would be interesting to explore in a more elaborate way.

2 Customer analytics, data sources and processing

In the introduction it is discussed that Avanade identified a need for more advanced types of customer analytics. A part of advanced customer analytics can consist of a method that enables the combination of segmentations based on different sources of data. From an academic viewpoint there is a need for a methodological framework that gives insight in how to combine different clustering or classification techniques in an appropriate and simplified way. To meet these needs this chapter will provide more information on the field of customer analytics and data sources it utilizes.

Throughout this paper formalized notations of concepts will be used. This notation is used within set theory, which states that sets are made up of objects (Stoll, 1979). These objects do not have to be physically collected together in order for them to constitute a set. Objects that are a member of a set A can be represented as $a \in A$. Likewise objects that are a member of a set B can be represented as $b \in B$. Meaning that all objects b are a member of set B. In the same manner it is possible to state that an object is not a member of a set e.g. $c \notin B$, meaning that object c is no member of set B.

2.1 Customer analytics

Customer analytics is necessarily concerned with analyzing data and models, and requires standard techniques from areas like statistics, data mining, machine learning and intelligent data analysis (Nauck, Ruta, Spott, & Azvine, 2006)(Bijmolt et al., 2010). It is implemented to drive decisions and actions, where data and models are defined at the individual customer level. Nauck et al. (2006) identify the following three main areas within customer analytics:

Customer segmentation

The process of dividing the customer base of a company into smaller groups to make marketing actions more effective. Each group can be seen as a segment of the total customer base with similar characteristics. Cluster analysis can be used to create customer segments. Customer are then mapped onto the segments. It has the potential to increase targeting effectiveness and to improve response to changing needs (Canhoto, Clark, & Fennemore, 2013).

Predicting customer actions

Historic information of individual customers consisting of process data of events and interactions is used to predict future customer actions. This can only be done effectively when long customer history is available.

Understanding customer views

Survey data is used to understand how satisfied customer are about the products or services they use. Better understanding of customers can be leveraged to create an increase in customer loyalty or satisfaction.

The use of customer analytics is especially effective when the available customer data is voluminous and beyond human processing capabilities (Germann et al., 2014). With the aid of computers customer analytics-based methods can be used to help make repetitive decisions. Because of this repeated use, cost that went into development can be justified. It also provides a feedback mechanism that can be used to continuously calibrate and improve the methods.

Marketing functions within companies have been shifting towards quantitative approaches that are based on data-driven decisions instead of instinctively choosing the best decision to take (Stodder, 2012). These marketing functions are also the functions within a company that are the most suitable for customer analytics. Basing decisions on information that is retrieved from data-driven methods enables companies to become more efficient in understanding their customers' expectations and preferences which allows them to deliver services, merchandise, and promotions that fulfill these needs and wants (Corrigan, Craciun, & Powell, 2014).

2.1.1 Campaigns

A campaign can be seen as a collection of marketing activities that aims to have a certain beneficial effect on the company (Tsiptsis & Chorianopoulos, 2010). It is one of the most common customer-facing activities and aims to improve customer relationships (Yao, Sarlin, Eklund, & Back, 2014). The monetary benefit of the campaign depends on the goal of the campaign and the response from potential or already existing customers. Effective campaigns are key to leading a successful business (Chan, 2008). It is possible to make a distinction between the type of the campaign and its goal. The type defines the approach of the campaign (Patterson, 2005). This approach can include campaigns with multiple stages and the use of a range of communication channels.

In the past, campaigns haven't been tailored to specific customer groups because of a lack of understanding of their behavior (Canhoto et al., 2013). Nowadays this has completely changed due to an increase in research, availability of customer data, computing power and competition (Davenport, 2006). The direct goals for campaigns differ widely, but altogether aim to maintain good customer relationships and enhance customer value (Chan, 2008). Since the campaign goal mainly specifies the desired reaction, the main goals will be explained next according to the paper of Patterson (2005).

Name	Goal
Acquisition campaigns	Acquiring new customers and/or preventing them from buying products from competitors. Uses of acquisition models can help to optimize the marketing campaign.
Cross-/deep-/up-selling	Selling additional products, more of the same product, more financially beneficial products or a beneficial composition of complementary products
Retention	Retaining existing customers, keeping them satisfied and preventing them from moving to competitors.
Customer intimacy	Similar to retention campaigns in the sense that it is also aimed at retaining customers. Customer intimacy is however more focused on customer specific offers, instead of more general offerings.
Recovery	Get customers to re-instate an existing product or service e.g. after a period of having cancelled a contract.
Market research communications	Receiving a response from the customers to a market research survey.
Data enrichment communications	Collection additional information about the customer e.g. stimulation of filling in additional profile information
Statutory/operational communications	Rather than selling a product or service, this campaign is aimed at passing important information to the customer e.g. changes in the terms and conditions.
Calendar-driven	Improve brand recognition by conducting campaigns at fixed moments in the year e.g. during holidays or at the beginning of seasons.

Table 2. Campaign names and goals

Table 2 shows an overview of all the campaign names and their goals. It is important for campaigns to have a clear, preferably measurable goal and customer target audience in order to be effective, although poorly defined campaigns can also potentially be effective (Tsiptsis & Chorianopoulos, 2010). Another aspect of campaigns that is important to mention, is the fact that it can lead to different customer behavior (Yao et al., 2014). Meaning that the segments in which these customers are divided in can change. This increases the need for dynamic segments which can react to changes in customer data.

2.2 Data sources

In the most abstract sense, data can be seen as a collection of data objects and attributes that belong to these objects (Li, 2015). Attributes are properties or characteristics that describe the data objects. When presented with a table of data, the rows can be seen as the objects and the columns are the attributes that give information with regard to this object e.g. person as object and gender as attribute. The attributes of a single object can be represented by a single row of values, which is called a vector (Smola & Vishwanathan, 2008). The attributes within this vector can be either numeric or categorical, respectively quantitative and qualitative.

2.2.1 Data structure

The structure of data can be described in different ways. When looking at the structure of documents, it can be seen as either structured or unstructured. A structured document has a clear sematic and syntactical structure. It should be known what each word or value in the document represents e.g. values of attributes in a database. Although an unstructured document might display a sematic and syntactical structure, this structure is implicit and to some degree hidden in its textual content (Feldman & Sanger, 2007).

Humans have the ability to distinguish and apply linguistic patterns which allows them to make use of this implicit structure (Gupta & Lehal, 2009). This also includes synonyms, contextual meaning, typographic elements, and spelling variations. Computers are not able to handle these variations but can, in contrast to humans, process text in large volumes. Besides unstructured and structured documents, Feldman and Sanger (2007) also identify two other types of structuredness of documents, namely: weakly structured and semi-structured documents. These types are identified by the way the document is formatted. Examples given in this work for weakly structured documents are research papers and business reports. They lack strong typographical, layout, or markup indicators. E-mail and web pages are examples of semi-structured documents, as these have consistent format elements.

2.2.2 Nominal, ordinal, interval and ratio

There should be made a clear distinction between attribute values and the attributes themselves. Four types of different scales per attribute can be identified, namely nominal, ordinal, interval and ratio scales (Stevens, 1946). These scales define the suitable properties of the attributes values. Nominal scale attributes allow the determination of equality, meaning that it can be used to distinguish one object from another e.g. gender, bought products and residence. Ordinal scale attributes allow the determination of greater or less, meaning that it provides information that allows objects to be ordered e.g. sentiment (positive, neutral, negative), age group (young, old) or prospective and historical revenue values. Interval scale attributes allow the determination of equality of intervals or differences, meaning that the differences between attributes become relevant. An interval scale attribute has a defined zero point and can have values below zero e.g. time and temperature in Celsius. The last scale type are ratio scale attributes which allow the determination of equality of ratios. The difference with interval is that it has a non-arbitrary zero point, meaning that comparisons between values become meaningful e.g. €100,00 is twice as much as €50,00.

2.2.3 Continuous, discrete and categorical

It is also possible to describe data that fall into these scales in a more general way. Attributes that use an interval or ratio scale are categorized as being continuous attributes (Raykov & Marcoulides, 2012). A continuous attribute can take an infinite number of possible values between two points lying in its range, in contrast to a discrete variable that van only take a countable number of distinct values. Attributes on a nominal scale or an ordinal scale can be seen as categorical and mostly contain qualitative data. The distinction between continuous, discrete and categorical are important, as the types of permissible statistic operations and data mining methods will depend on them e.g. for the most typical value within a collection of numerical attributes the mean can be calculated and for categorical attributes this is the mode, which is the value that occurs the most (Brito, Almeida, Monte, & Byvoet, 2015).

2.2.4 Historic customer data

The utilization of historic customer and social activity data has already been mentioned in the introduction. The most simple form of historic customer data is generated from registering customers within the company. This data consists of basic customer socio-demographics information like: name, address, gender, birth data, contact info and sometimes also number of children, marital status, educational status and annual income (Tsiptsis & Chorianopoulos, 2010).

Additional data on transactions, purchases and product groups are common to use in addition to some of the socio-demographic fields. These are common to apply data mining techniques to, in order to find useful patterns that can be used in the customer segmentation process.

2.2.5 Social activity data

Social activity data can be seen as all external data that is freely available on the internet and involves existing and potential customers that express themselves with relation to their daily lives, products and services (Shum & Ferguson, 2011). The most interesting information that can be found would be related to markets, trends, products or services which the “listening” company can use to base decisions on when creating marketing campaigns, products or services. “Listening” is a term generally used to refer to the process of monitoring and making sense of all the information that is posted daily on social channels (Schweidel & Moe, 2014). Ideally the extracted information of products and services would also include an indication of the sentiment that is related to this post. Although the benefit of monitoring social channels has been perceived by many already, there are still plenty of difficulties in utilizing and measuring this in an effective way (Schweidel & Moe, 2014).

The first challenge of utilizing and measuring social channels is the fact that the majority differ from each other in the way comments or posts are structured. Meaning that multi-channel listening would call for a solution that can be adjusted for each channel in such a way that it leads to similar results in relation to the rest of the channels. This is either found difficult to do or just ignored, meaning that the measurements do not reflect the customer's opinion. Another problem identified by Schweidel and Moe (2014) is that aggregated data of sentiment can provide misguiding results, since the proportion of positive comments for different channels like blogs and micro blog can be very different, given the period in which the comments are observed. Meaning that independent from the content, the sentiment will be biased towards a certain value dependent on the period of the year e.g. January vs August. These so called social dynamics need to be accounted for when monitoring multiple channels in relation to brand sentiment.

2.3 Data mining and text mining

Data mining or knowledge-discovery is the process of extracting potentially useful information from raw data (Rajagopal, 2011). Although "mining" may imply that data mining allows the discovery of new facts within a collection of data, it actually allows the discovery of trends and patterns in large quantities of structured data in a (semi)automated way (M. A. Hearst, 1999). The data used by data mining is mostly stored in large databases (M. Hearst, 2003). This means, that data mining is predominantly put to practice by companies who have substantial collections of historic data of customers, their interactions and purchases.

2.3.1 Data mining

Data mining can help to find patterns among the historic data of different customers and their purchases. This knowledge discovery process does not involve any human interactions and can provide both descriptive and predictive information, depending on the techniques used (Rajagopal, 2011). Patterns found by mining data can help organizations to make better decisions (Susena, 2014). Other knowledge discovery processes are statistical analysis of data and ad hoc queries that can be performed on the customer database. Important models within data mining are classification, clustering, association analysis, sequence discovery, regression, forecasting an visualization (Ngai, Xiu, & Chau, 2009), shown in Table 3. The following explanations of the previously mentioned models are based on the papers of Rajagopal (2011), Ngai et al. (2009) and Susena (2014).

- Classification: is a supervised learning method to predict future behavior of customers by classifying database records into a number of predefined classes based on certain criteria. These predefined classes are formed by a model that is trained by training data.

- Clustering: groups objects in homogenous groups according to logical relationships. This is done in such a way that objects from the same cluster are similar and objects from different clusters are dissimilar. Objects have no class label before the clustering process starts.
- Association analysis: indicates correlations between attributes. Two common approaches are Breadth First Search approach (BFS) and Depth First Search approach (DFS). Association analysis is very similar to sequence discovery, except for the time element that is not taken into account when deriving rules.
- Sequence analysis: discovering sequential rules that identify time dependent behavior patterns and associations between attributes. This type of analysis can be utilized to problems that involve data containing a sequential nature (Ahola & Rinta-runsala, 2001).
- Regression: statistical estimation technique used that maps objects to a real value that provides prediction value.
- Forecasting: creates an estimation of future values based on patterns found in historic records. The historic records allow the creation of a model that can continuously be used to create an estimation of future values.
- Visualization: generate a visual representation of complex data in order to make patterns and relationships between attributes or objects easier to understand.

Model	Techniques	Application
Classification	Decision trees, Neural nets, Regression	Predicting customer behavior
Clustering	K-means, Neural nets, Discrimination analysis, Self-Organizing Maps (SOM)	Customer segmentation
Association analysis	Apriori, Statistics	Market basket analysis, Cross selling
Sequence discovery	Apriori	Web usage discovery
Regression	Linear regression, Logistic regression	Modelling causal relationships, Testing scientific hypotheses
Forecasting	Neural nets, Survival analysis	Forecasting model
Visualization	Hygraphs, SeeNet	Present complex data

Table 3. Data mining models and their applications

An example of a pattern which a supermarket might encounter is the simultaneous purchase of two products. Knowing the purchasing patterns of your customers can be used to an advantage of the company (Hand, Blunt, Kelly, & Adams, 2000).

This is what makes it lucrative to store data from customers and their purchases. Text mining is a field that is related to data mining and also tries to extract useful information through the discovery of patterns (Hotho, Andreas, Paaß, & Augustin, 2005). How text mining differs from data mining is the source of data, and in contrast to data mining, it can also work with unstructured and semi-structured data (Gupta & Lehal, 2009). This type of data is more likely to be found on the internet, making it an important field when looking into mining information from online open sources.

2.3.2 Text mining

Text mining uses document collections as its information source and instead of structured database records the patterns are to be found in unstructured natural language text in the documents of these collections (Feldman & Sanger, 2007). Similar to data mining, text mining relies on preprocessing routines, pattern-discovery algorithms and presentation-layer elements such as visualization tools (Feldman & Sanger, 2007). This means that text mining consists of several consecutive steps in order to find the patterns in the unstructured text. These steps can be split into two phases: the preprocessing or text refining phase and the knowledge distillation phase (A. Tan, 1999).

Because data mining assumes the data has already been stored in a structured way, the prepossessing of text mining is different. The prepossessing or text refining has to transform the unstructured data into a more structured intermediate form (A. Tan, 1999). The following challenges can be expected when analyzing big data sets, as found in the paper by Fan and Bifet (2013):

1. Volume: size continues to increase
2. Variety: many different types of data: text, sensor, audio, video, graph etc.
3. Velocity: data arrives continuously, real-time information is interesting
4. Variability: changes in structure and how users interpret data
5. Value: being able to answer otherwise impossible questions

2.3.3 Machine learning

Sometimes machine learning is confused for data mining, as they both aim to find interesting regularities, patterns or concepts in empirical data (Mannila, 1996). There are however as few differences between them. With data mining, the emphasis lies on the complete knowledge-discovery process and for machine learning this seems to lie on the learning step of the process (Hand et al., 2000). Since the emphasis with machine learning is on this learning step, there mostly is an underlying interest in the mechanism that produces the data (Mannila, 1996). This is also because the nature of the mechanism can be very complex. This isn't necessarily the case for data mining.

Machine learning knows many applications e.g. page ranking, collaborative filtering, automatic translation, classification, named entity recognition and speech recognition (Smola & Vishwanathan, 2008). Of these applications, classification will be the most relevant for this paper.

2.3.3.1 Unsupervised learning

Unsupervised learning methods can be applied without prior information on the amount or type of classes within the data, meaning that the algorithms do not need to trained on labeled training data (Guerra et al., 2011). Unsupervised learning is essentially an algorithm that will work with any interference of a human and does not require any prior training to work. An example of unsupervised learning is clustering. This algorithm will identify and define its classes, instead of manually having to setting the boundaries for the clusters (Dhandayudam, 2012). The following types of clustering algorithms can be identified:

- Exclusive clustering: k-means, Self-Organizing Feature Mapping and Particle Swarm Optimization (Deng, Jin, Higuchi, & Han, 2011). Exclusive indicates the fact that an object can either be a member of a cluster or not, there is nothing in between. It can also only be a member of only one cluster in total.
- Overlapping clustering: Fuzzy C-means (Peters, Crespo, Lingras, & Weber, 2013). Overlapping in the sense that objects can be a member of more than one cluster at a time. In addition to this there is also a measurement of how strongly a members belongs in a clusters, instead of either being part of it or not.
- Hierarchical clustering: Dendograms (Hiziroglu, 2013). As hierarchical indicates, there is the added dimension of a hierarchical structure between the clusters. Clusters can be part of a bigger cluster. This allows details to be hidden when this is required. Usually the results of this type of clustering are presented as a dendogram.

2.3.3.2 Supervised learning

Supervised learning differs from unsupervised learning in the sense that it automatically creates a model based on labeled training data (Guerra et al., 2011). The process of labeling data has to be done manually by people. This makes it a more labor-intensive process than unsupervised learning, but it can lead to better results in some situations.

Classification is an example of a supervised learning algorithm (Smola & Vishwanathan, 2008). From the model that leads from the training data, the classifier is able to label new unlabeled data based on the labels that were assigned to the values of the training data (Guerra et al., 2011). Classification can be divided into 3 steps: feature vector creation, training a classifier and label prediction (Timonen, 2013).

During the first step of classification, feature vectors are created. A feature vector is like a normal vector, but contains a set of values, features, that describe an object e.g. weights for words within corpora (Geffet & Dagan, 2009). The first step involves some preprocessing depending on the requirements of the classification algorithm. In addition to this, the feature vectors are weighted in order to put more emphasize on the important feature vectors. The second step involves the actual training of the classifier. The feature vectors are required to poses labels in order to train to classifier. These labels could be the result of manual labeling, which would require a person to manually assign each feature vector to predefined classes (Smola & Vishwanathan, 2008). After the training process the last step takes place, the resulting classifier can be used to predict the class of an unlabeled feature.

Within a training data set T, each instance is represented by a vector (x_n, l_n) such that $T = \{(x_1, l_1), \dots, (x_n, l_n)\}$ where x_n depends on the n-th predictor variable and l_n expresses its label (Mohri, Rostamizadeh, & Talwalkar, 2012). For a new instance x, the supervised algorithm builds the following target labeling function γ with the input space X and output space Y,

$$\gamma : X \to Y$$

A specific example of a classification method is Naive Bayes, which is a classification method based on probability theory (Jackson & Moulinier, 2007). This method uses prior probability of training data to classify unlabeled data. Meaning that the probability P of a term t occurring in a class Z is known e.g.

$$P(t \mid \mathrm{Z})$$

To calculate the probability that a document that is represented as a term vector with m components, $Q = (t_1, \dots, t_m)$, belongs to a certain class, the probability of each individual term t that occurs in document Q can be calculated in the following way:

$$P(Q \mid \mathrm{Z}_n) = \prod_{j=1}^{j=m} P\left(t_j \mid \mathrm{Z}_n\right) = P(t_1, \dots, t_j \mid \mathrm{Z}_n) = P(t_1 \mid \mathrm{Z}_n) \cdot P(t_2 \mid \mathrm{Z}_n) \cdot \dots \cdot P\left(t_j \mid \mathrm{Z}_n\right)$$

Naïve Bayes takes the assumption that the presence or absence of a word in a document is independent of the presence or absence of any other word (Manning, Raghavan, & Schütze, 2008). This is called the conditional independence assumption.

2.3.4 Preprocessing

Techniques like text and data mining require the data to be in a certain format in order to perform the algorithms that can find useful patterns (Hotho et al., 2005). Data mining requires highly structured data and text mining can also work with semi-structured and unstructured data. This does not mean that text mining can directly be applied to raw data.

Preprocessing can help to remove unwanted data from the data collection. This can include empty records, but also the removal of outline (Seret, Maldonado, & Baesens, 2015). A lot of preprocessing is targeted specifically at finding the most relevant and descriptive information for each document, so the most relevant document for a certain word is found in the most efficient way (Hotho et al., 2005).

At the end of the preprocessing phase, the data is converted into a more manageable presentation called a feature vector, which is an entity without internal structure (Feldman & Sanger, 2007). Each document will be represented as a vector in this feature space. The feature model can be seen as an intermediate form which can be used to in the second step of text mining, namely the knowledge distillation.

2.3.5 Precision and recall

The precision and recall problem is something that is always encountered when trying to extract relevant information from data sources. When performing a query or extracting information, the precision of the results indicate the amount of relevant information that is found (Manning et al., 2008). If precision is increased, recall decreases. This means that results which might have been relevant won't be shown, in this way higher precision is achieved and thus relevant, but less results are shown (Lewis, 1994). The same holds for recall; when recall increases, precision will drop. Meaning that more results will be shown, but they won't be as relevant.

2.4 Information Extraction

Data in the form of natural language text contains information which is not directly suitable for automatic analysis by a computer (Hotho et al., 2005). This is because automatic analysis requires structured data. This problem generated the need for a method or technique to transform this unstructured data into more structured data. Information extraction can be seen as a method which tackles this problem. It allows the transformation of unstructured text into a structured database, called an intermediate form (Sirsat & Chavan, 2014). This intermediate form can be a collection of documents. When applied to this paper, each post within the social channels can be treated as a document. A collection of natural language text document is also called a corpus (Feldman & Sanger, 2007). These corpora can be subjected to quantitative analysis techniques that identify patterns or frequent non-fixed combinations of words and distinctive words called keywords (Baker, 2010).

Information extraction can be seen as a precursor to data mining (McCallum, 2005). From a set of unstructured data it is possible to extract relevant and valuable information in otherwise worthless data, but it is also possible to find relevant information between other valuable but irrelevant information (Kroeze, Matthee, & Bothma, 2003).

2.5 Natural Language Processing

Natural language processing refers to the method of analyzing or synthesizing spoken or written natural language in an automatic way (Jackson & Moulinier, 2007). By applying natural language processing a better understanding of human text can be achieved (Hotho et al., 2005). This can be reached through the use of linguistic concepts like part-of-speech and grammatical structure (Kao & Poteet, 2007). This is very useful since the amount of information in the form of natural language only keeps increasing. Apart from this, there already exists an enormous amount of information which is only available in natural language form (Grishman, 1997). Natural language processing would make it possible to structure this information which would make the individual facts accessible (Collobert et al., 2011). It also has a symbolic and an empirical approach (Jackson & Moulinier, 2007). The symbolic approach relies on hand-coded rules, grammar and gazetteers (Bellot, Bonnefoy, Bouvier, Duvert, & Kim, 2014) and the empirical approach derives language data from text corpora, by using statistical analysis of language.

Named Entity Recognition can be performed to automatically extract facts from information. These facts can include the names of entities such as persons, objects, products, brands, companies or locations (Bellot et al., 2014). Named Entity Recognition produces the best results on more structured texts with symbolic natural language processing while statistic natural language processing reaches the best results when the texts are less structured (Bellot et al., 2014).

Knowledge Discovery from Text (KDT) is a text mining technique which is deeply rooted within natural language processing. This technique allows the extraction of explicit and implicit concepts and semantic relations between concepts (Gupta & Lehal, 2009). It is an important technique for emerging applications like text understanding.

3 Requirements hybrid clustering

To make the idea of combining different segmentations more concrete, this chapter will formulate more specific requirements regarding the main goal of this method.

Social engagement shown in Figure 1, which is currently being used to get insights in social trends, does not provide detailed information (1) on these potential customers expressing themselves in social channels. It also lacks integration with the information that is currently available (2) about customers, the historic customer data, and it only monitors trends that are related to manually set terms (3) (Meloen, 2015). These 3 weaknesses should be resolved by the proposed conceptual framework in the following way:

- More detailed insight in social activity data besides sentiments and frequency of manually set terms
- A method for combining or utilizing historic and social activity data in a uniform way during the segmentation of the customer base
- Terms suggestion or automatic method for deciding what terms to use

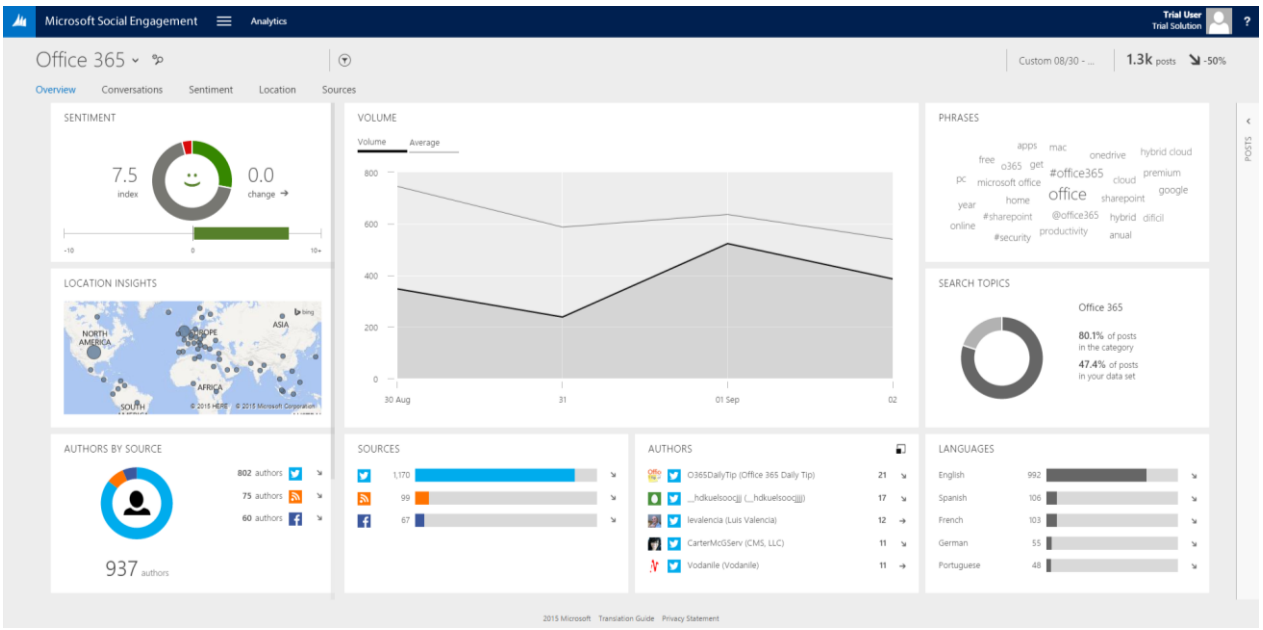

Figure 1. Microsoft Social Engagement Dashboard

The solution would call for the creation of an intermediary customer profile, called an archetype. These archetypes would be formed for both data sources. Since the archetypes would be filled in with more abstract and preferably also overlapping data, the comparison between the archetypes of both data sources would be possible.

Similarity between archetypes could create matches between customers from historic customer data and people expressing themselves in social channels. The overlap in information would create the links between the archetypes. Additional information contained in them could be used to create a more complete view of that type of customer.

In this way marketers would only have to focus on these archetypes, which contain data from both social and historic sources. After the formation of the types, the need for keeping them up to date would arise, meaning that it's an ongoing process.

3.1 Visual representation

Figure 2 gives a visual representation of hybrid clustering. The detailed working of this technique will be explained in the following chapters. The idea is that historic customer data will be used to create customer segmentations with the aid of a certain clustering algorithm. This will also be done with the social activity data, assuming that it is available in an already preprocessed form. Clustering this data will also create segments.

The archetypes that will be formed from these segments will need to show some similarities in order to decide which social archetypes match the best with the historic archetypes. It is possible that multiple social archetypes show relevant similarity with a historic archetype, meaning that multiple social trends are associated with that archetype. Because the archetypes originate from a historic and social segment, the original segments can be linked to each other.

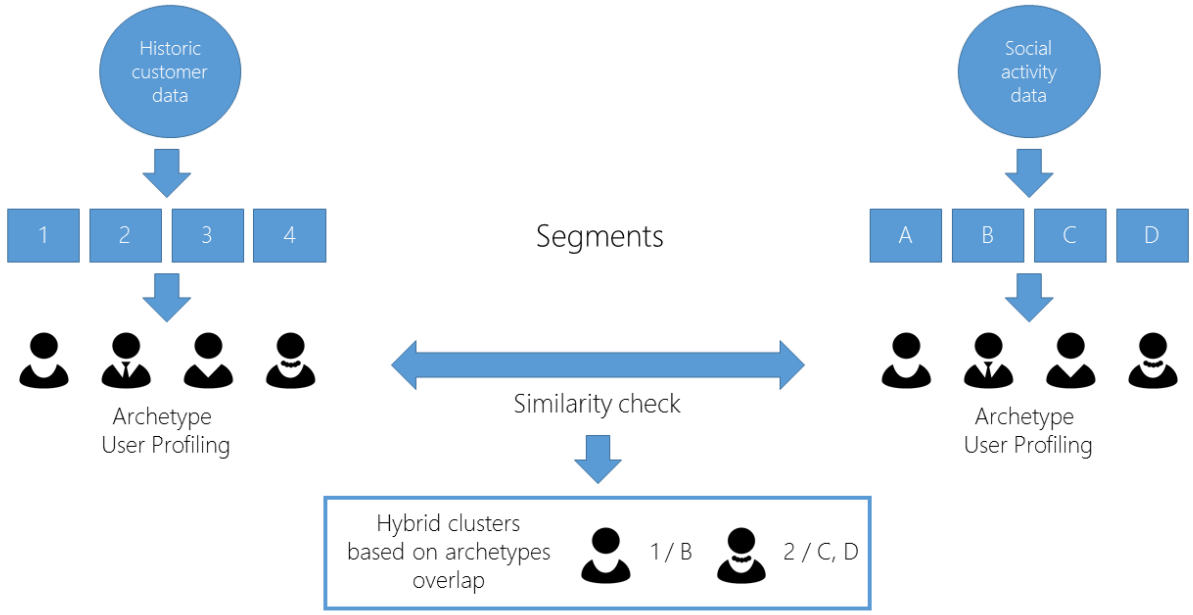

Figure 2. Visualization hybrid clustering concept

4 Proposed conceptual framework

In this chapter, the conceptual framework for hybrid clustering will be presented. The main aim of this framework is to organize and define the concepts that are involved with hybrid clustering. This will serve as a basis for the coming chapter that will provide more in-depth information on hybrid clustering.

4.1 Relations between concepts

The following concepts will be explained in this paragraph: segmentation, customers, archetypes, clustering, representation of segments and dynamic segments. Each segmentation can contain many customers. Customers will only belong to one segment based on the clustering technique that will be used. Archetypes are a type of customer profile that contains data from different sources and results from analysis and classification techniques. The results should be representable in a graphical form and since the data used to create the segments contains ever changing data, the dynamics of the segments should be taken into account. Figure 3 gives a visualization including the cardinality of the relations between the concepts.

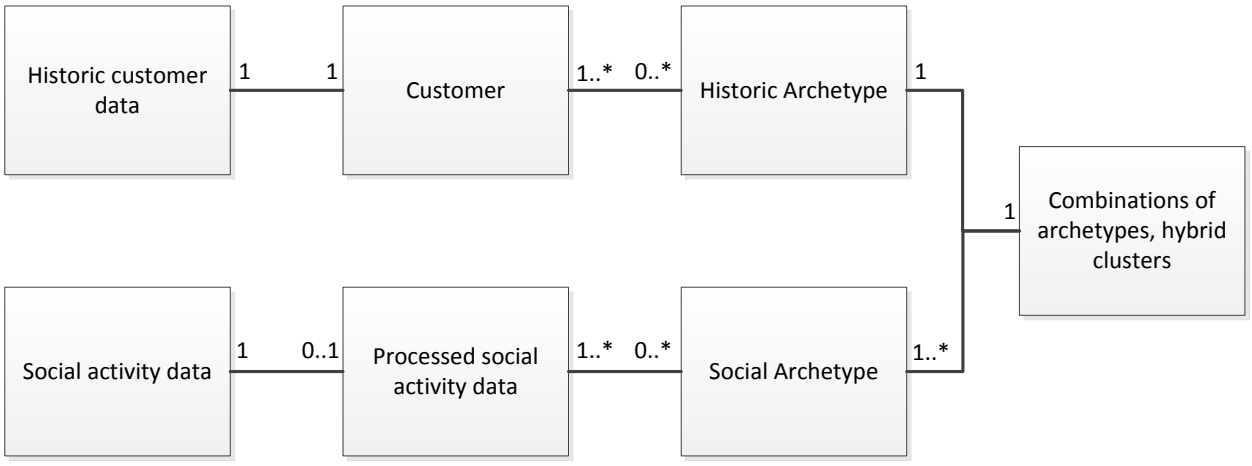

Figure 3. Relations between concepts

4.2 Segmentation

This first step in understanding the concept of combining different segmentations is to define what a segmentation is and how different segmentations can behave when they are combined. To illustrate this it is possible to use Venn diagrams (Grunbaum, 1984). Each of the segments will be presented as a circle containing objects with a corresponding variable.

4.2.1 Disjoint segments

When we treat the sets A and B as segments, they can be described as being disjoint when the intersection of the two results in an empty set, which means that not a single object is a member of both segment A and B, as seen in Figure 4 (Stoll, 1979). To put this in a more mathematical way, set theory can be utilized. The operations that are used to create new segments, or so called sets, resemble the operations that are used with normal integers. This can be symbolized by $A \cap B = \emptyset$. A collection of sets is a disjoint collection if each distinct pair of its member sets is disjoint. For example,

$$\{1, 2, 3\} \cap \{4, 5, 6\} = \emptyset$$

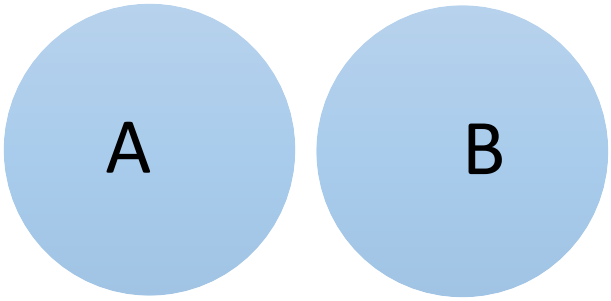

Figure 4. Independent segments

4.2.2 Dependent segments

Segments can be described as dependent when the intersection of the two results in objects that are both member of segment A and B, as seen in Figure 5 by the darker section in the middle of segment A and B. The intersection of the sets A and B can be symbolized by $A \cap B$ and reads "A intersection B". The resulting set consists of objects that are both a member of A and B, $A \cap B = \{x | x \in A \wedge x \in B\}$. The following example illustrates this principle,

$$\{1, 2, 3\} \cap \{1, 3, 4\} = \{1, 3\}$$

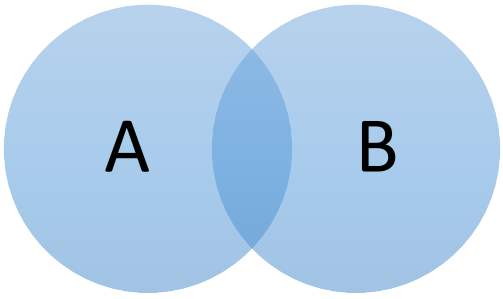

Figure 5. Dependent segments

Another dependency is shown in Figure 7. Segment B is completely contained within A, meaning that all members of B are also part of A. Therefore B is a subset of A. Figure 6 shows an almost complete overlap between the segments A and B. This means that segment A and B contain almost identical members. The parts that are not overlapping show the only unique members for both segments. Meaning that the set of A is near identical to B.

$B \subseteq A$

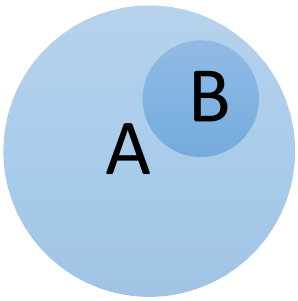

Figure 7. Segment B contained in A

$A \cong B$

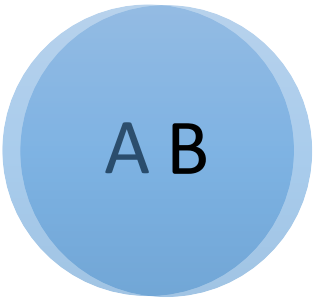

Figure 6. Segment A and B almost identical

4.2.3 Customer segmentation

Customer segmentation is the process of dividing the customer base into homogenous groups, which enables to conduct different marketing campaigns according to the characteristics of that group (Tsiptsis & Chorianopoulos, 2010).

A customer segment consists of a group of customers that have similar characteristics and is a subset of the group to which all customers belong (Dhandayudam, 2012). These segments can be obtained by dividing them based on certain characteristics or by applying the data-driven approach.

This last approach applies clustering algorithms to the complete customer base, this will create smaller groups of customer with similarities that can be relevant to the company and have a more rich segmentation scheme (Tsiptsis & Chorianopoulos, 2010).

Targeting segments instead of the whole market enables companies to deliver better value to consumers resulting in maximum rewards due to close attention to consumer needs (Sun, 2009). It is important to recognize the continuous changes that certain customer segments undergo.

The following goals of customer segementation are identified by Tsiptsis and Chorianopoulos (2010):

- Greater understanding of customers to support the identification of new marketing opportunities
- Design of new products/services for each segment's characteristics rather than the mass market

- Design of customized product offering strategies to existing customers
- Offering tailored rewards and incentives
- Selecting the appropriate advertising and communication message and channel
- Selecting the appropriate sales and service channel
- Determining the brand image and the key product benefits to be communicated based on the specific characteristics of each segment
- Differentiation in customer service according to each segment's importance
- More effective resource allocation according to the potential return from each segment
- Prioritization of the marketing initiatives which aim at customer retention and development according to each segment's importance and value

4.2.4 Segment requirements

In the most extreme case continuous changes in customer data and the segments this data creates, could lead to the disappearance of some segments and the occurrence of new ones (Blythe, 2008). To be of applicable use, such segments must meet the following requirements:

- Members of the segment must be identifiable and measurable
- A segment must be accessible as a group that can be communicated with
- A segment must be substantially big
- Members of the segment must have a close agreement on their needs. It must be congruent.
- The members and nature of a segment must be reasonably stable

The following bases are mainly used for segmenting a market (Sun, 2009):

- Geographic segmentation: dividing the market on the basis of different geographical units.
- Demographic segmentation: dividing the market into different groups based on age, sex, family life circle, income, occupation, education, religion, race, generation gap and nationality.
- Psychographic segmentation: dividing the market into different groups based on life style, personality or values.
- Behavioral segmentation: dividing the market into different groups based on knowledge of, attitude toward, usage of, or response to a product.

Each of the bases on which segmentation is possible contains different variables or attributes of the customer on which segmentation is possible. The simplest variant of segmentation can be based on only one variable, called single-variable segmentation (Blythe, 2008).

It is however more likely that segmentation on more variables is desirable, multivariable segmentation, since it provides more accurate segments. Increasing the variables also leads to smaller segments and therefore smaller markets.

4.3 Customers

A customer can be seen as an entity within the database that describes a real life person through attributes and corresponding values. That is, for an existing customer, meaning a person that has previously been registered and purchased a product or service. Because this paper will only focus on B2C transactions, customers in this context cannot consist of other companies. The reason being that certain attributes only exist for individuals. There's also the added difficulty that companies mainly use social channels to communicate commercially with their customers in order to improve customer engagement, instead of expressing themselves regarding brands or products (Rybalko & Seltzer, 2010). This means that the term customers can be used to describe already existing customers that exist within the historic customer data of the company or potential and existing customers that express themselves within social channels.

4.3.1 Historic customer data

There exists a set of customers $C = \{c_1, \dots, c_n\}$ and a set of attributes $A = \{a_1, \dots, a_m\}$ that are possible for each customer. In addition to this there exists a set of values V, that consists of the values that are assigned to the attributes by the historical customer data. Every attribute has a value that belongs to the domain of that attribute.

Customer $c \in C$ can have the attributes $a_1, a_2, a_3, a_4 \in A$ and the values $v_1, v_2, v_3, v_4 \in V$ an example of a customer object could be, $c = \{a_1 : [v_1], a_2 : [v_2], a_3 : [\,], a_4 : [v_3, v_4]\}$. To give a more clear idea of how this would look in practice, attribute and value will be substituted by real world examples.

$$c = \{age : [24], gender : [female], service : [\,], products : [product1, product2]\}$$

4.3.2 Social activity data

Besides a set C containing all the historic customer data, there also exists a set S with all the social activity data. Each object $s \in S$ will contain a post D containing a set of terms $D = \{t_1, \dots, t_n\}$ and a set attributes A. Meaning that $s = \{D, A\}$.

4.3.3 Customer profiling

Customer profiling is the process in which customer profiles are created based on their attributes. This covers all types of information, including age, gender, life cycle, preferences, hobby, location, bio, and time period to use social media (Dunk, 2004). This information will enhance the customer information management in CRM (Karna, Supriana, & Maulidevi, 2014)(Ahola & Rinta-runsala, 2001).

Customer profiles enable companies to get a better understanding of the different types of customer they are dealing with. This is useful when a company wants to directly contact a current customer or is exploring if a new customer fits to any of their current profiles (Dunk, 2004).

A customer profile $p \in P$ is a partial mapping of customer attributes to values V, meaning that $P \subseteq C: A \mapsto V$. This can be illustrated in the following way. Profile $p \in P$ has the attributes $a_1, a_2, a_3, a_4 \in A$ and the values $v_1, v_2, v_3, v_4, v_5 \in V$ in the following way $p = \{a_1 : [v_1], a_2 : [v_2], a_3 : [v_3, v_4], a_4 : [v_5]\}$.. To give a more clear idea of how this would look in practice, attribute and value will be substituted by real world examples.

$$p = \{age\_group : [20 - 25], gender : [female], services : [1,2], cluster : [3]\}$$

A customer profile is a formula in terms of attributes which can be true or false. A customer can either be part of a profile or not, based on its attributes. For example, a customer c will be part of a customer profile p when it's attributes are within the boundaries of the specification of that profile. This formula can take the following form for some profile p_2

$$a_1 > 0 \wedge a_2 < 5 \rightarrow p_2.$$

4.4 Archetypes

Archetypes can be seen as a kind of customer profile (Keupp & Gassmann, 2009). They are abstractions of the information that is found in the historic customer data and social activity data. In contrast to customer profiles, archetypes consist of additional data that is the result of advanced techniques like machine learning and data mining. Examples of these are sentiment, segments or clusters, gender approximation and age approximation. The cluster information per archetype will be the same, meaning that there will exist an archetype for each segment or cluster.

Since archetypes only include additional information and thus extra attributes, they can be treated the same as profiles. There should be made a clear distinction between archetypes based on historical customer data and social activity data. We define the set of archetypes for historic customer data as $H = \{h_1, \dots, h_n\}$ and the set of archetypes for social activity data as $G = \{g_1, \dots, g_m\}$. φ is used to formalize all attributes that exist in the sets C, S, H and G. In order for an attribute it to exist in C it should be part of H. The population for each archetype $h \in H$ and $g \in G$ can be formalized as the following,

$$\rho(h) = \{\varphi \in H | \varphi \in C\} \qquad \rho(g) = \{\varphi \in G \mid \varphi \models S\}$$

The reason archetypes have an additional attributes that result from advanced techniques, is to make the output from the segmentations based on historic customer and social activity data comparable. Without this set of additional attributes there wouldn't exist any attributes of the same type. Meaning that the union between any segment created from historic customer data with a social activity data segment will result in an empty set e.g. $A \cap B = \emptyset$. How these techniques work will be discussed in chapter 5. For now it is sufficient to assume that the techniques will allow the creation of additional attributes for the social activity , which results in a non-empty set when intersecting both segmentations.

The archetypes from the historic customer data h and the social activity data g will both contain similar attributes that allow an intersection that should not always result in an empty set. The attributes are likely not to be a complete match, which means that a more fuzzy intersection is required. Based on the level of similarity θ between the attributes from h and g that are intersected, it is possible to determine if the archetypes are similar enough to be joined (Wang, Xu, & Wu, 2009). The formula for this comparison is the following, where $H = \{h_1, \dots, h_n\}$ and $G = \{g_1, \dots, g_m\}$ and of which h_i and g_j are the individual archetypes:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(h_i \cap g_j) \geq \theta$$

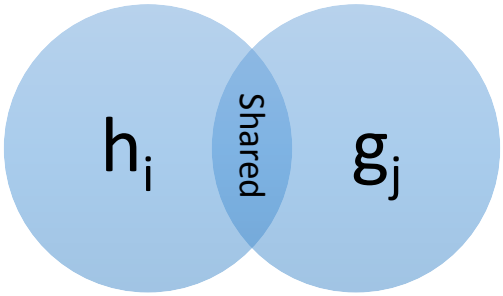

Figure 8. Archetype shared attributes

When the shared attributes are similar enough, see Figure 8, historic customer data that was used to create the archetype will be linked to the social activity data though a union. The similarity measure and the union of the individual archetypes can be formalized in the following way:

$$M(H,G) = \{\, h_i \cup g_j \mid \sum_{i=1}^{n}\sum_{j=1}^{m}(h_i \cap g_j) \geq \theta\}$$

This step involves a certain degree of generalization. It is needed to make such a generalization because it is not possible to directly link social activity data to existing customers. The real name belonging to the owner of a post is hard to find or validate (Rao, Yarowsky, Shreevats, & Gupta, 2010), but it is even more likely that it isn't a customer of the products or services of the company that is scanning the social activity data. As seen in Figure 9, this shows how the sets will be combined when the intersection of the shared attributes, shown in Figure 8, are similar enough. This set is the result of the union between the archetypes h_i and g_j.

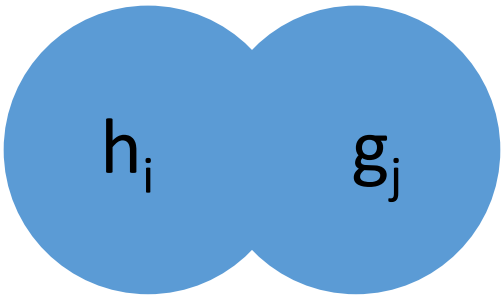

Figure 9. Combined archetypes

4.5 Clustering

Within the field of data mining, clustering is an example of an unsupervised learning algorithm, meaning that it requires no predefined classes. Unsupervised essentially means that an algorithm will work without any interference of a human and does not require any prior training to work (Dhandayudam, 2012).

There are various clustering algorithms, of which only k-means will be explained in this paragraph since it is one of the most popular hard clustering algorithms (AlFalahi, Atif, & Abraham, 2014)(Hiziroglu, 2013)(Kandeil et al., 2014). In the following chapter the most suitable clustering algorithm will be selected for the problem of customer segmentation.

The k-means clustering algorithm will identify and define the classes, instead of manually setting the boundaries for the clusters. It aims to partition objects $O = \{o_1, \dots, o_n\}$ into disjoint subsets W_i with $i = \{1, \dots, k\}$ such that there is high similarity inside the clusters and low similarity among the clusters (Peters et al., 2013).

The clusters partition the data for applications like market or customer segmentation (Luo, Cai, Xi, Liu, & Zhu, 2013). This allows the discovery of customers with similar attributes. It is required to specify the maximum and minimum amount of clusters prior to the clustering process in order to create clusters that are of a size that are preferable to work with (Gullo, Domeniconi, & Tagarelli, 2013).

This is more of a trial and error process, as the size of the clusters is a result of the clustering algorithm. Figure 10 will show the basic working of the k-means algorithm. As seen in Figure 10, the k-means clustering algorithm needs a predefined number of clusters K to start. These clusters initially start with a set of centroids E that are randomly defined. This set is represented as $E = \{e_1, \dots, e_n\}$ where $c_j = (c_{j1}, \dots, c_{jn})^T$ (Deng et al., 2011). The distance between the centroid and each n-dimensional point will then be calculated. The centroid to which the object is nearest will then be assigned to it. Figure 11 gives a visual representation of the k-means algorithm given K=2.

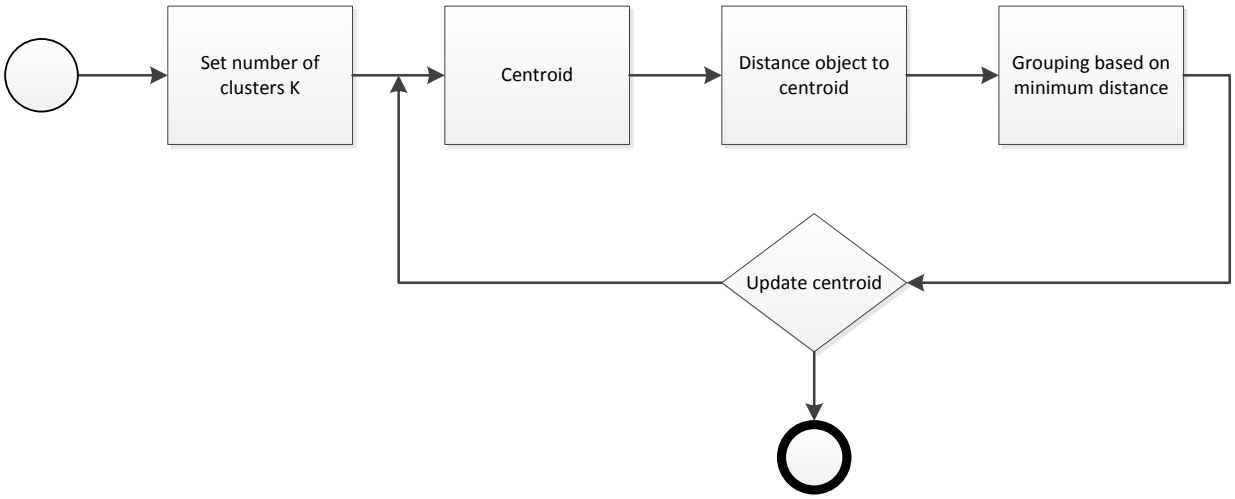

Figure 10. K-means clustering process

A set of m n -dimensional points $\{x_i, \dots, x_m\}$ can be described as $F = \{x_1, \dots, x_m\}$, containing the clustered objects x_i, represented by the transposed feature vector $x_i = (x_{i1}, , \dots, x_{in})^T$. After assigning each object to a cluster, re-calculate the centroid for the cluster receiving the new object or for the cluster losing the object (Koudehi, Rajeh, Farazmand, & Seyedhosseini, 2014). This can be repeated until the centroids are stable and aren't adjusted anymore. The method of calculating the difference also knows some variations, but for this explanation the Euclidean distance will be used (Berry & Linoff, 2004). This distance for a n-dimensional data point x_i and a centroid c_j can be calculated in the following way:

$$d(x_i, c_j) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \cdots + (x_n - c_n)^2}$$

The membership of an object within a cluster can be formulated as $\Lambda_i = (\lambda_{i1}, \dots, \lambda_{iK})$ with λ_{i1} = 1 for $h \in \{1 \dots, K\}$ and $\lambda_{il} = 0$ for $l = 1, \dots, K \wedge l \neq h$ (Peters et al., 2013).

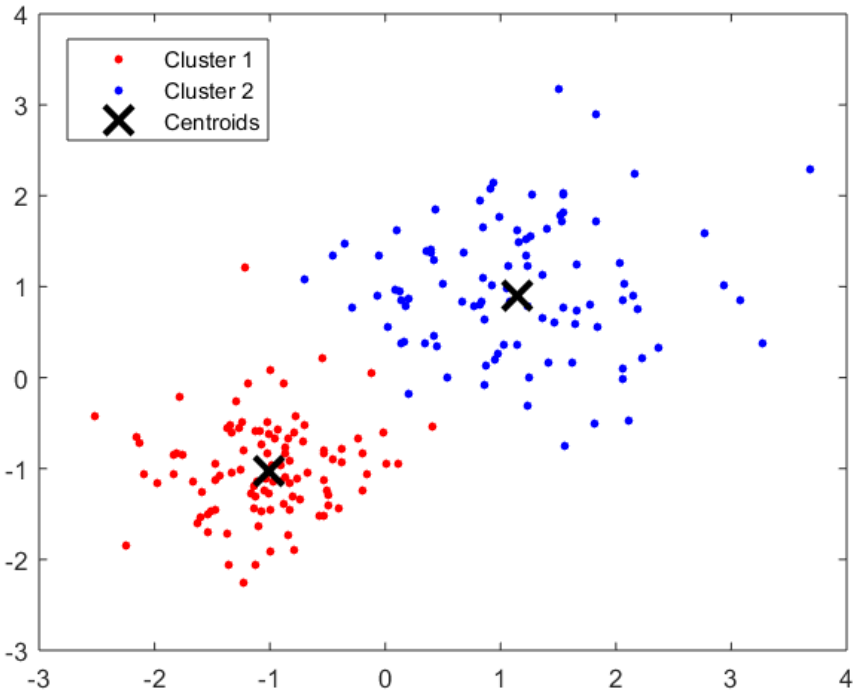

Figure 11. K-means (K=2) in MATLAB R2015a with random data

In order for the k-means algorithms to be applied the data points have to be numeric (quantitative data) (Berry & Linoff, 2004). The is no point in trying to calculate the similarity between two purchases, since there is no measurable distance between them. In order to be able to cluster data there needs to be a form of natural association. This means that all the attributes of a customer that are to be clustered should be translated into numeric values, so they can be treated as points in a space.

Post-hoc segmentation, is the classification job in the segmentation process that is based on clustering (Hiziroglu, 2013). In contrast to a priori segmentation, which chooses some variables of interests and then classifies consumers based on that designation (Hiziroglu, 2013). Most studies linking language with psychological variables rely on a priori fixed sets of words, such as the LIWC categories (Schwartz et al., 2013).

There also is a distinction between hard or crisp and fuzzy or soft clustering. In fuzzy clustering, the classes themselves are not necessarily well-defined or mutually-exclusive (Wu & Chou, 2011)(Peters et al., 2013). This means that entities within a cluster can overlap and thus become part of other clusters. This can be explained by looking at hard clustering as assigning either a 1 or 0 to a particular member and when clustering in a soft or fuzzy way this value could be between 1 and 0, so 0.4 for instance.

4.6 Representation of segments

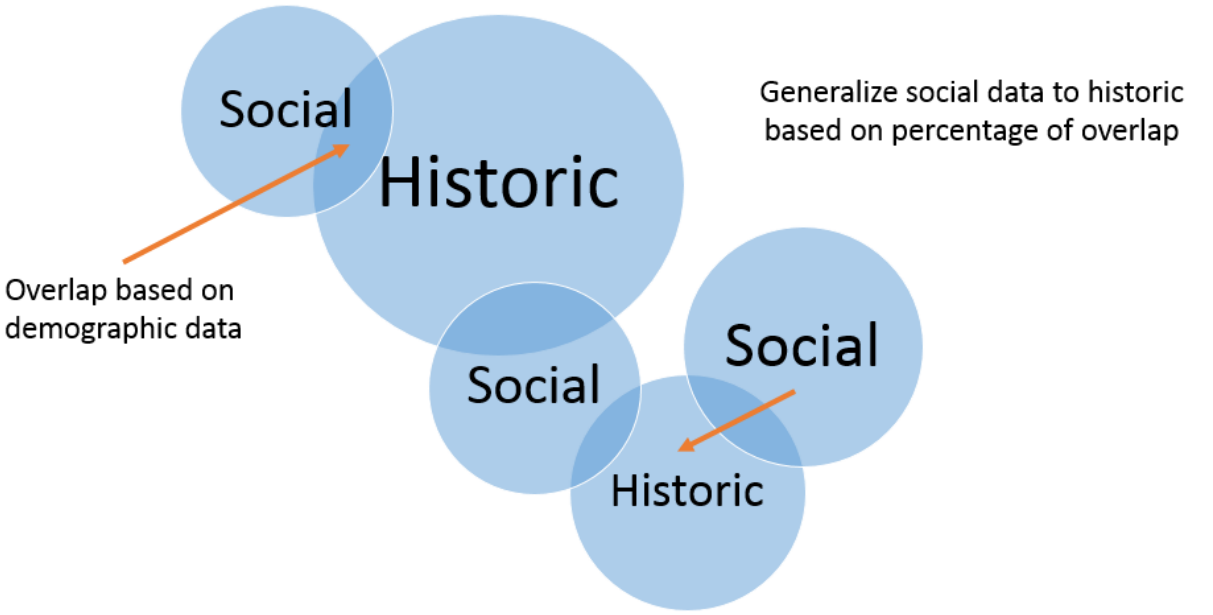

Figure 12. Representation of segments

As seen in Figure 12, the representation of the segments and their overlap would ideally be in shown as circles (Yao et al., 2014). The size of the circle indicates the amount of objects contained in them and the amount of overlap between the circles indicates the similarity based on the attributes in both circles. It is possible that a historic segments shows overlap with multiple social segments. Based on a variable threshold the amount of similarity required to show any amount of overlap can be defined.

4.7 Dynamic segments

A clustering technique like k-means that is applied to a single data set only gives a static idea of the patterns that are present in the data set (Seret et al., 2015). When customer segmentation techniques are it can be expected that customer will move between segments, based on their changing behavior. Capturing this movement is only possible with the aid of comprehensible techniques, such as described by Seret et al. (2014). The techniques used comprise of a modified version of the traditional SOM algorithm, k-means clustering and the generalized sequential pattern algorithm (GSP).

It is however more relevant to keep the clusters that are created based on the historic customer data H and the social activity data G up to date, as the data is constantly changing. Techniques like data stream clustering that are used to discover knowledge from large amounts of continuously generated data are less suitable given the nature of the data (Silva et al., 2013).

Data streams are typically provided by sensor networks, meteorological analysis or computer network traffic monitoring and can be seen as a massive sequence of data objects . Each data object can be described by an n-dimensional attribute vector belonging to an attribute space Ω w can be continuous, categorical, or mixed. The characteristics of these data steams differ in the sense that the data can be used for clustering directly and the quantity of new data is way higher than with customer data.

The most sensible solution would be to recreate the archetypes, and thus the clusters, each time the clusters need to be up to date. Although social media posts can be seen as a data stream, the quantity and speed are not near that of the data streams when searching for very specific information, which makes it an unsuitable technique.

5 Hybrid clustering for B2C customer segmentation

This chapter will explain the steps within the process of hybrid clustering for B2C customer segmentation. During this process, both historic customer and social activity data are manipulated and analyzed in multiple ways. To be able to combine segmentations based on historic customer and social activity data there exists a data problem that needs to be solved. In addition to this an explanation of relevant data mining techniques will be evaluated, after which the most suitable ones will be chosen to apply to counter this problem. Aside from data mining there will also be some machine learning techniques will be used to classify certain attributes of a post.

5.1 Combining historic customer data with social activity data

There are multiple ways to interpret the way in which historic and social segmentations can be combined. Each of these methods brings its own complications. To explore the possibilities and explain what makes some approaches complex, each of these approaches will be explained briefly.

5.1.1 Method to combine segmentations into new segments

Identify clustering algorithms to create customer segments for historic customer data and social activity data. Define a method to combine these segmentations in order to create new segments. It is however not possible to directly combine segmentations based on different data sources, since the sets will not contain any data that has similar attributes, meaning that an intersection will result in an empty set, $H \cap S = \emptyset$, as seen in Figure 13 (Stoll, 1979).

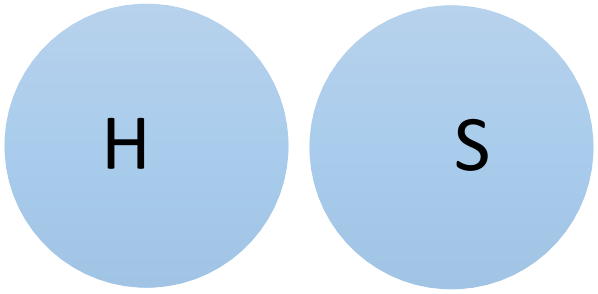

Figure 13. Intersection of H and S leading to an empty set

5.1.2 Method for mapping very specific historic segments onto bigger social segments

Identify clustering algorithms to create very small social activity segments that can later mapped to the segments created with the historic customer data. Mapping method needs to be defined. There is however no information that can be used to combine the very specific historic segments onto the bigger social segments (Gullo et al., 2013).

As seen in Figure 14. The smaller subsets of the social segments, displayed in blue, are linked to subsets in the historic segment. Creating small segments for social activity data allows more detail to be maintained, which would increase the value of adding the social data to the historic data. But because of the lack of a shared attribute, no sensible automatic combination is possible.

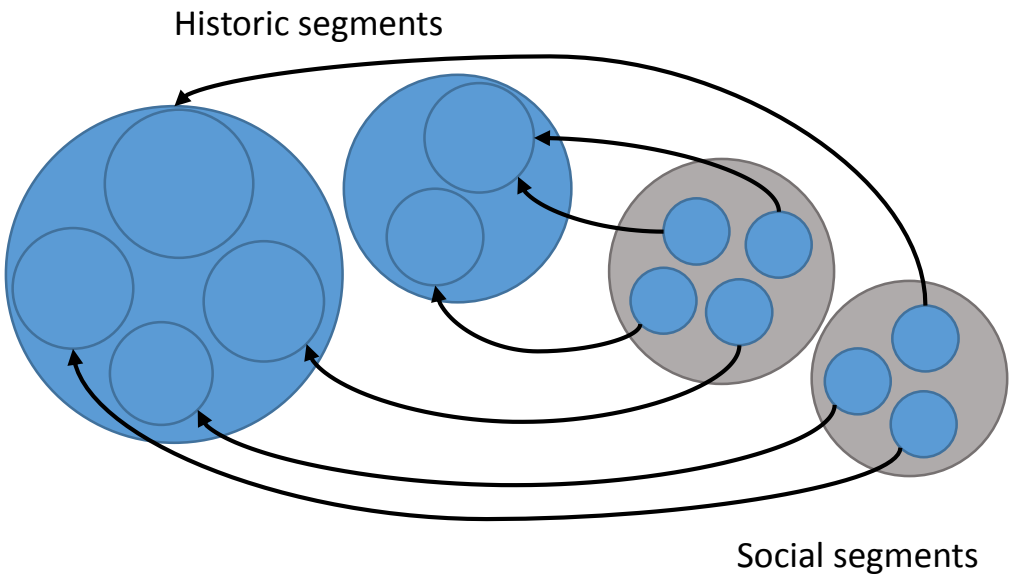

Figure 14. Mapping of social activity data

5.1.3 Method to directly combine historic and social activity data

Method to combine historic and social activity data before clustering and creating segments, in order to create more detailed segments. Ideally the social activity data would be automatically combined with the suitable historic customer data, so every existing customer also has social data. There is however no suitable method of extracting relevant social information of each existing customer, with the added problem that it might not even exist.

The methods described above a not suitable to approach the problem of combining segmentations from different data sources. The process in paragraph 5.2 describes the overall process of preparing the data and applying the right methods, in order to be able to combine the segments based on historic customer data and social activity data in a sensible manner. This process is also shown in a Unified Modeling Language (UML) diagram that will be presented in paragraph 5.2.1 in Figure 15. UML is a visual, object-oriented modeling language that is commonly used to model business processes and software systems (Engels, Förster, Heckel, & Thöne, 2005). The diagram presented in Figure 15 was created according to the rules described in the paper by Engels et al. (2005).

5.2 Process overview

Historic customer data

- Data preparation tasks such as filling missing data, removing outliers, feature extraction, and feature selection (Koudehi et al., 2014)
- Select domain relevant data e.g. customer life time value (LTV) or recency, frequency and monetary (Koudehi et al., 2014), preferably numeric data
- Apply Self organizing map method (Deng et al., 2011)
- Apply k-means clustering algorithm, exploratory data mining, to historic numeric customer data in order to create clusters of customers (Wang et al., 2009)(Seret et al., 2015)

Social activity data

- Extract posts/data from desired online sources
- Apply preprocessing techniques to extracted posts/data in order to prepare posts for further analysis (Gupta & Lehal, 2009)
- Select relevant posts/data according to information retrieval and extraction techniques (Karna et al., 2014)
- Apply sentiment analysis to relevant posts/data (Schweidel & Moe, 2014)
- Apply age/gender analysis to relevant posts/data (Schwartz et al., 2013), (Miller, 2012)
- Extract location data if relevant to online source (Cao et al., 2015)
- Assign topic/summary to each post (Sharifi, Hutton, & Kalita, 2010b)
- Apply clustering algorithm, exploratory data mining, to social activity data in order to create clusters of people based on most relevant terms or sentiment (Aggarwal & Zhai, 2012)

Hybrid clusters

- Archetypes have been formed during the first two processes of clustering historic customer data and social activity data, these will be used to combine the data from both sources
- Set a value for θ, which defines the required amount of similarity between historic and social archetype attributes, which results in an non-empty intersection.
- Comparing all historic and social archetypes will result in a set of combined archetypes, depending on the value of θ. These combined archetypes will be called hybrid clusters.
- These hybrid clusters will enable marketing campaigns to be based on customer segments that include both historic customer data and social activity data that is relevant to that specific cluster, based on the attributes that they share.
- Compare marketing campaigns based on regular clusters with hybrid cluster through an AB-test (Siegel, 2013)
- Continue renewing both clusters and archetypes

5.2.1 UML diagrams of hybrid clustering for B2C customer segmentation

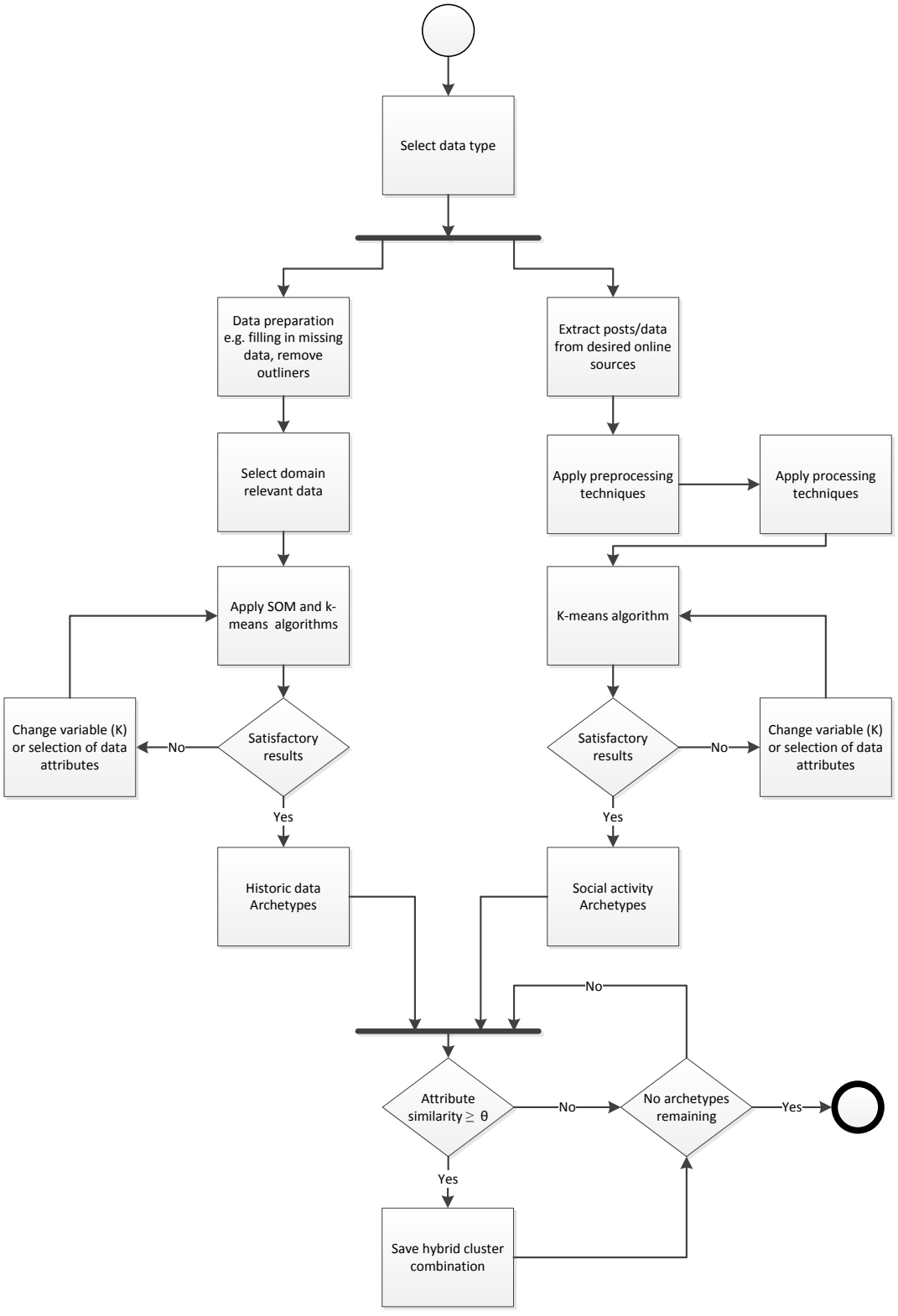

Figure 15. UML Diagram of the Process of Hybrid Clustering for B2C Customer Segmentation

Figure 15 displays the UML diagram of the process of hybrid clustering for B2C customer segmentation. Creating archetypes based on historic customer data and social activity data can be seen as 2 parallel processes. They don't have a specific order in which they should be executed. Besides this, three iterations can be observed. The first two consist of the process of getting satisfactory results from the SOM and k-means algorithm, both for historic and social data. Since the number of clusters is set manually, although there are techniques to find the optimal amount automatically (Şchiopu, 2010). Lastly, there is an iteration to try all archetype combinations, which are combined based on the level of similarity θ between the intersected attributes from h and g.

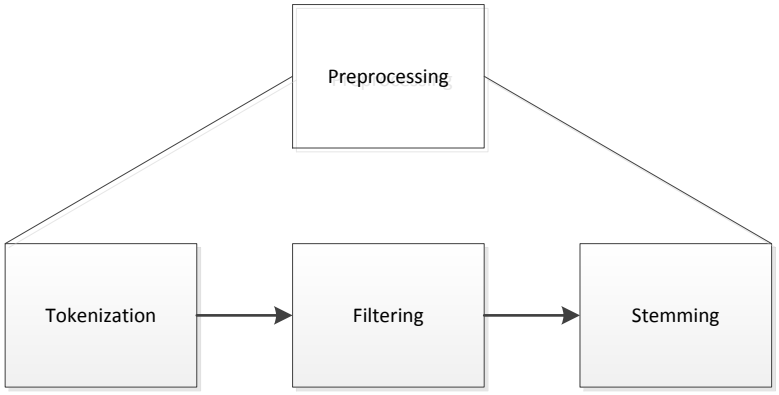

Figure 16. Preprocessing steps

Figure 16 and Figure 17 display two processes from Figure 15 in more detail. It concerns the preprocessing and processing components within the process of forming a social activity archetype. Depending on the extent in which the classification techniques can handle non-text characters, the filtering and tokenization processes have to be adjusted in order to keep this type of information in the processed post. Stemming should only be applied when the size of the dictionary needs to be reduced, since a lot of information can be lost during this process (Hotho et al., 2005).

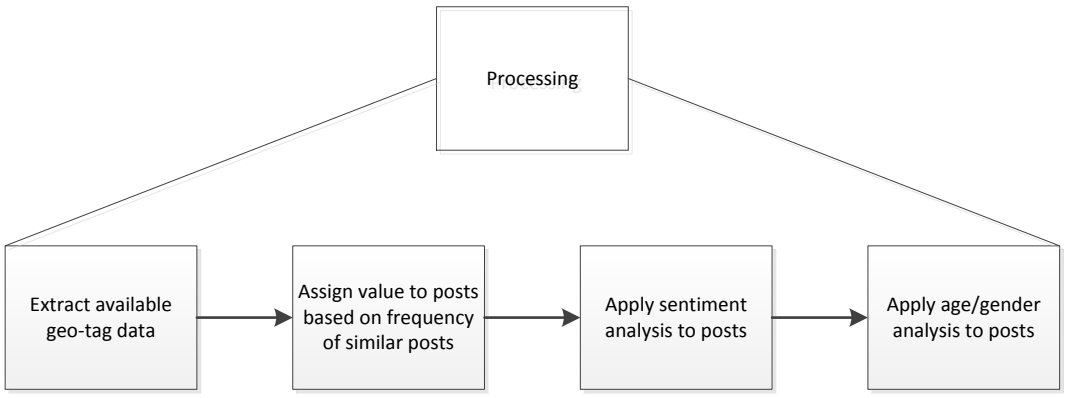

Figure 17. Processing steps social

5.3 Existing data analysis techniques

5.3.1 Customer segmentation

This paragraph will give an overview of existing techniques that are used to establish the segmentation of customers. According to the paper of Hiziroglu (2013), there can be made a distinction between clustering and classification techniques. Clustering techniques follow a post-hoc approach and discover the characteristics of the segments without any prior training. This contrasts with classification techniques which follow an a-priori approach and require the characteristics of the segments e.g. number of segments, dimensions and descriptions, in order to be applied. As the automation of the clustering process for customers is the most important and the majority of papers focus on clustering techniques, the classification techniques are out of scope for the paragraph. Table 4 will provide a non-exhaustive list of clustering techniques that focus on customer segmentation.

Used techniques	Focus	Application	Literature
K-means	Customer segmentation	Understand consumption characteristics of different customer groups within the telecom sector	(Luo et al., 2013)
K-means	Customer segmentation	Categorize B2B customer records into several groups within consumer goods company	(Kandeil et al., 2014)
K-means, SOM & Particle Swarm Optimization (PSO)	Customer segmentation	Customer segmentation classification in mobile e-commerce	(Deng et al., 2011)
K-means & PSO	Market segmentation	Provide precise market segmentation for marketing strategy decision making	(Chiu, Chen, Kuo, & Ku, 2009)
SOM	Customer segmentation	Segmentation framework that considers LTV, current value and client loyalty to build client segments	(Cuadros & Domínguez, 2014)
SOM & k-means	Market segmentation	Cluster the customers based on questionnaire answers	(Kuo, An, Wang, & Chung, 2006)
SOM & k-means	Customer segmentation	Use demographic and transaction data to use suitable marketing strategies for the chain stores	(Koudehi et al., 2014)
SOM & k-means	Customer segmentation	Updating and improving an existing clustering model by adding relevant new variables	(Seret et al., 2015)

Table 4. Clustering techniques and their applications

When the summary of current literature on customer segmentation is reviewed, the most dominant clustering technique remains k-means. Although the Self Organizing Map (SOM) is also used regularly, it is usually applied in combination with k-means. Combining more than one clustering algorithm can overcome the individual deficiencies of the applied techniques. This multi-stage models are what Deng et al. (2011), Koudehi et al. (2014), Kuo, An, Wang, & Chung (2006) and Seret et al. (2015) proposed in their papers.

For example, some deficiencies for k-means are the dependency on the choice of the initial centroids and the possibility that convergence to a local minimum doesn't occur under certain conditions. The k-means clustering has already been discussed and presented in a formalized way in paragraph 4.5. As the SOM is also an important clustering technique within the field of customer segmentation, it will be explained in a more formalized way in this paragraph.

The SOM is an unsupervised learning neural network (Kuo et al., 2006) that classifies high-dimensional input data points into visually determinable clusters on a low-dimensional level, much like k-means (Tsai & Lu, 2009). The SOM was proposed and demonstrated by Kohonen (1990) as a new form of artificial neural networks algorithm. In the SOM, all the clusters in the low-dimensional level are interconnected. Note that the SOM maps new input data into existing clusters based on the similarity of data. In addition to this, it does not require the number of clusters to be specified beforehand. The aim is to have an output, as shown in Figure 18, which shows similar inputs as corresponding nearby outputs.

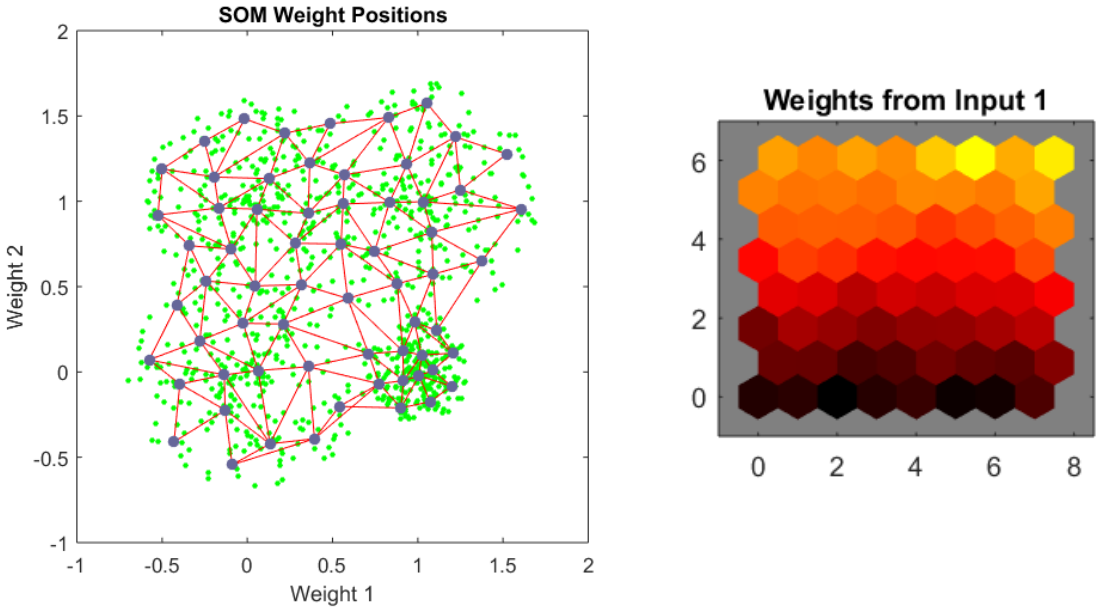

Figure 18. SOM weight positions and heat map in MATLAB R2015a

The SOM creates an output that is typically a two-dimensional lattice, as shown on the right in Figure 18, which consists of nodes that are all connected to the input space $I = \{x_1, \dots, x_n\}$ (Larose, 2005). The nodes can be considered the clusters in this clustering algorithm and will be indicated by K. Each of these nodes in the output space is represented by an n-dimensional weight vector w_j that corresponds to the dimensions of I (Deng et al., 2011). In Figure 19 a section of a two-dimensional lattice consisting of three nodes is shown. When a point from the input space x_i is handled by the SOM, it is mapped to a node in the output space when it's weight vector w_j is closest to that of the input vector of x_i.

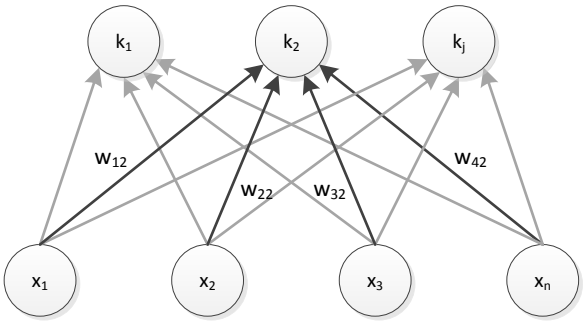

Figure 19. Three neurons k with corresponding weight vectors w and four input vectors x

The distance between each node and input space can be calculated by the squared Euclidean distance (Patel & Mehta, 2012). It calculates the difference between the vector from the input space x_i and the weight vector w_j for each node j. This can be done with the following formula,

$$d(x_i, w_j) = \sum_{i=1}^{n} (x_i - w_{ij})^2$$

When the minimum for $d(x_i, w_j)$ is found, the input point is mapped to the output of the winning node (Kohonen, 1990). In this way the high-dimensional input can be represented as two-dimensional output (Hotho et al., 2005). In addition to this the nodes j that are in the neighborhood of the winning node should adjust their weight vectors in the following way,

$$w_{ij} = w_{ij} + \eta(x_i - w_{ij})$$

The range of this neighborhood can decrease based on a time-variable (Kohonen, 1990), meaning that the SOM will adapt slower when the amount of steps increases. In the above formula, η is the learning rate of the SOM and can take a value between 0 and 1 e.g. $0 < \eta < 1$ (Larose, 2005). The initial weights corresponding to the nodes are random, but gradually the SOM will start "learning" and adapts to the weights of the input space.

5.3.2 Social media analysis

This paragraph will give an overview of existing literature on social media analysis, shown in Table 5. The papers presented in this table will be explained in more detail later in this chapter. The aim is to give a general idea about the focus of the research and the technique that is commonly used, given this focus.

Focus	Common technique	Literature
Topic modeling, summarization	LDA	(Naveed, Gottron, Kunegis, & Alhadi, 2011), (Weng, Lim, & Jiang, 2010), (Rosa, Shah, Lin, Gershman, & Frederking, 2011), (Gruber, Rosen-Zvi, & Weiss, 2007)
Sentiment assignment	Classification	(Schweidel & Moe, 2014), (Kouloumpis, Wilson, & Moore, 2011), (Pak & Paroubek, 2010)
Age determination	Support Vector Machine (SVM) Classification	(Santosh, Bansal, Shekhar, & Varma, 2013), (Nguyen, Gravel, Trieschnigg, & Meder, 2013), (Rao et al., 2010)
Gender determination	SVM Classification	(Santosh et al., 2013), (Rao et al., 2010)
Location dynamics	Extraction though application programming interface (API)	(Zhai et al., 2015), (Cao et al., 2015), (Cranshaw, Hong, & Sadeh, 2012),

Table 5 .Literature summary social media analysis

As seen in Table 5, there are five focuses within social media analysis that are relevant to this paper. Sentiment assignment, age determination and gender determination add additional attributes to the posts/documents that will be analyzed. This can be done through classification algorithms, which can be trained to categorize posts/documents into predefined categories e.g. age 20 – 25 or positive sentiment.

Aside from these classifiers, location dynamics includes research that aims to model the dynamics of cities though freely available location data that can be extracted from social channels. This paper's main interest is the raw location data, as this can be used to help create shared attributes between both archetype types. This data can be extracted through the various API's that social channels offer to developers or other interested parties. Lastly, it is important to be able to process all the posts and create a means of discovering trends or being able to summarize similar posts in order to get an overview of the content users are creating. This is called topic modeling or simply summarization.

5.4 Methods historic customer data

5.4.1 Data nature

In contrast to social activity data, the nature of the historic customer data is completely dependent on the data storing and customer monitoring approach of the enterprise that aims to implement hybrid clustering.

As discussed in chapter 2, there are four types of different scales per attribute, namely nominal, ordinal, interval and ratio scales (Stevens, 1946). These scales define the suitable properties of the attributes values and influence the type of clustering technique and distance measure that can be used. Interval and ratio scales, that are categorized as being continuous data, are the only data instances that are directly suitable for clustering with a k-means algorithm (Şchiopu, 2010). Meaning that categorical values should be either transformed to normalized numeric values or an alternative algorithm should be used. In addition to this, it is also possible to transform continuous values to discrete values, this is a process called discretization (Larose, 2005). Compared to continuous values, discrete values are more simple to use and comprehend, as they are closer to the knowledge-level representation (H. Liu, Hussain, Tan, & Dash, 2002).

5.4.1.1 Sample historic data

A sample of customer data from a telecom provider can be used to give an idea of the data nature of a customer, a single record is shown in Table 6. Sample of historic customer data. This sample consists of 3333 records and is used within the book “Discovering Knowledge in Data: An introduction to data mining” by Larose (2005). Originally this sample is used to predict the possibility of a customer churning, meaning that they will either switch to another company or stay. This is portrayed by the fact that a large part of the data is numerical and relatively little demographic data is present. The following table will show a part of the attributes and values for a customer.

State	Account Length	Area Code	Phone	Inter Plan	Day Mins	Day Calls	Day Charge	CustServ Calls	Churn
KS	128.00	415	382-4657	No	265.1	110	45.07	1	False

Table 6. Sample of historic customer data

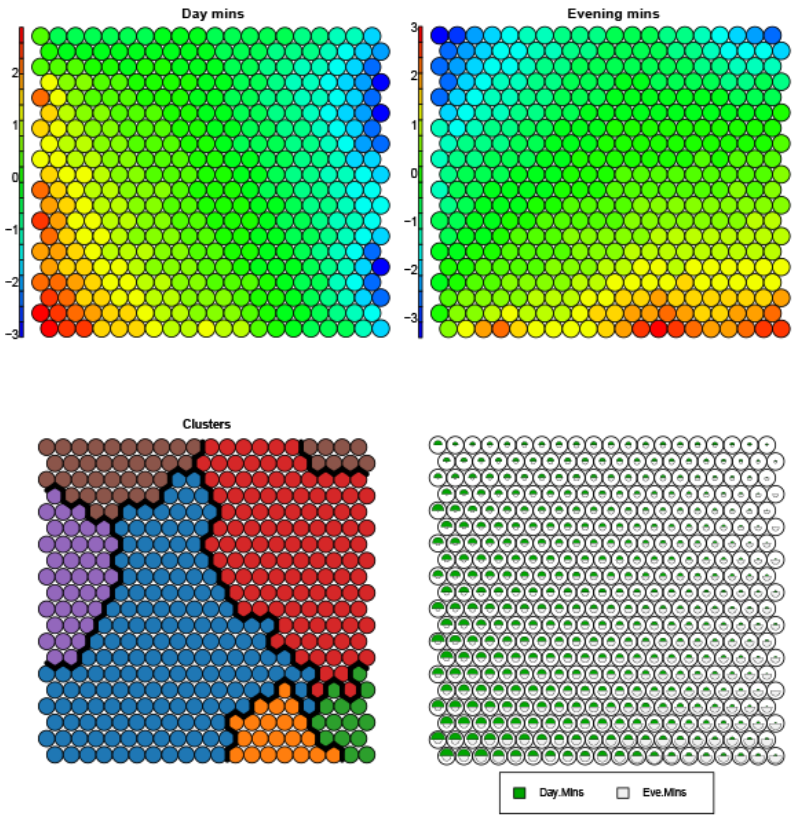

Figure 20. SOM and K-means in RStudio

For the visualizations in Figure 20 only the evening minutes and daytime minutes were taken into consideration. These visualizations were created using a package called “Kohonen” in RStudio (Cran R Project, 2015). This package contains functions to train supervised and self-organizing maps. The top two visualizations display the amount of minutes users called during the day and evening. Red shows heavy users and blue shows users that did not perform any calls during the evening. The bottom left visualization displays the clusters in which users are categorized, based on day and evening calls. The bottom right visualization shows the amount of day calls as a green portion of the circle and in the same holds for evening calls, which are displayed in white. Looking at the high concentration of green in the bottom left corner, this corresponds with the red that is displayed in the top left visualization. This method of visualizing should only be used for a limited amount of attributes, in order to avoid clutter.

5.4.2 Suggested algorithms

As seen in Table 4, K-means and SOM are used throughout more recent literature in relation to customer segmentation. When clustering algorithms are combined they can overcome the individual deficiencies of K-means and SOM (Deng et al., 2011). This is why it is highly beneficial to use more than just one algorithm.

The difficulty is the lack of a clear method that states how to combine different clustering or classification techniques in an appropriate manner (Hiziroglu, 2013). Meaning that this limits the implementation of such a combination to the literature that clearly describes the steps that were taken and the results it provided.

As shown in the paper of Deng et al. (2011) K-means can be applied to define the cluster centers for SOM. This is the same process as directly applying k-means to a data set. The difference is that the weight vectors from the SOM nodes are directly mapped onto the centroids of the k-means algorithm (Terlunen, Barreto, & Hellingrath, 2015).

An advantage of combining these algorithm it the fact that the resulting boundaries with SOM are hard to find, since it maps multi-dimensional data onto a two-dimensional space (Chiu et al., 2009). K-means will enable to make these boundaries more visible, as presented in the bottom left of Figure 20. In less recent literature by Vesanto & Alhoniemi (2000) the SOM algorithm is used prior to implementing the K-means algorithm which decreases the computational load. Making it possible to cluster larger data sets and different preprocessing strategies.

After applying K-means and SOM to the data set, the resulting segments could be further improved by applying the PSO algorithm. This leads to more efficient searching and classification. The complete process of combining K-means, SOM and PSO is clearly described by Deng et al. (2011), but since the main reason of applying PSO is to create a more efficient algorithm, this is left out of the suggested algorithms as this not a main aim of this paragraph but does increase the complexity.

5.5 Methods social activity data

5.5.1 Data nature

An important source of social activity data will be micro-blogs like Twitter, which allow users to post messages of up to 140 characters encoded in UTF-8. Aside from posts, it is also possible to identify shares and replies as a type of post (Nguyen et al., 2013). Posts from Twitter can be categorized as being short, since they contain less than 20 words (Timonen, 2013). Documents up to 100 words will be considered as being a short document in this paper.

The average tweet is 11 words long, with words only occurring once per tweet (O'Connor, Krieger, & Ahn, 2010). Meaning that the count for the term frequency (TF) is roughly the same as the count for all the documents it occurs in. This makes classical term weighting, like the inverse document frequency (TF-IDF), is not very suitable for short documents (Timonen, 2013).

This is because the TF for most short documents is 1. The problem of TF = 1 is called hapax legomena, meaning that a word only occurs once within the corpus of a document. This is why the papers in table 5 are explicitly aimed at short documents, as posts on social media are most of the times classified as being short. Each post, reply or share on social media can be treated as a separate document. This is done because the techniques that are used to analyze the posts are from the field of information retrieval, which normally cope with large corpora consisting of websites, papers or books that need to be retrieved by queries (Jackson & Moulinier, 2007). In addition to hapax legomena, posts are of a short and informal nature (X. Liu, Fu, Zhou, Wei, & Zhou, 2010) and many of them contain the same information in a slightly different format (Tao, Abel, & Hauff, 2013) which makes summarization of posts more difficult with common natural language processing techniques. Some posts can also add non-relevant information e.g. bots repeating advertisements, spam, weather reports, news feeds, copies of other people's messages and template messages (O'Connor et al., 2010).

5.5.1.1 Sample social data

To get an idea of the data structure of social activity data, Twitter was chosen to collect an initial data set from. Twitter is one of the most accessible sources of user-generated data (Miller, 2012). They make the majority of their user data freely available through its data access API. It is also the main source of data within Microsoft Social Engagement that would be implemented by Avanade at their clients, see Figure 1.

The data was collected using the open source integrated development environment (IDE) RStudio (RStudio, 2015). It uses R, a free open source statistical language. The package TwitteR was used to provide an interface to the Twitter web API (Gentry, 2015). Using this package it was possible to extract posts from Twitter based on the function searchTwitter(). This function allowed multi term queries and the specification for the number of tweets that needed to be returned, language of the tweets, date restrictions and location data in the form of a geocode. Since similar packages for other social channels exist, including public API's, for this paper it is assumed that the nature of the data is very similar. Although there are some channels specific characteristics e.g. hashtags, which are seen as out of scope for this paper.

The data that was retrieved was a list of posts. This was converted to a character vector using the built in function "sapply" within RStudio. This function extracts the text from the posts and put them into separate objects within the character vector. After the creation of the character vector a corpus can be created using the "Corpus" function from the text mining package "tm". This function converts a vector source into a corpus.

A corpus is a collection of individual documents, which in this case are the individual posts from twitter. Below a list of 3 tweets from the corpus of 300 tweets is shown after searching for the most recent Dutch tweets containing “Microsoft”.

1. "Sharicant: en toen weigerde mijnen outlook alle medewerking ... GRM BL ... oplossing van Microsoft - koop nen upgrade – geërgerd"
2. "MiedemaM: RT @Computeridee: #Windows10 is er! En het is nog niet ee ns 29 juli... http://t.co/ZmbQwRob06 #Windows #Microsoft"
3. "ACHTMEDIA: Software: Microsoft start uitrol definitieve Windows 10- versie naar Insiders | http://t.co/ZaqaXzNHEU | Officieel verschijnt Windo ws 10 p…"

Table 7. Three example tweets for search term “Microsoft”

Based on the tweets in Table 7, the following characteristics can be identified:

- Includes non-text characters, hyperlinks, numbers, upper and lower casing characters
- Little sentiment recognizable in tweet 1 and tweet 2, based on the term “geërgerd” in tweet 1 and based on the exclamation mark in tweet 2.

5.5.2 Preprocess data

The posts or documents need to be converted into a more manageable presentation called a feature vector, which is an entity without internal structure (Feldman & Sanger, 2007). Each document will be represented as a vector in this feature space. The process of preprocessing is shown in Figure 16.

5.5.2.1 Tokenization

Through the means of the tokenization process, the sentences of a post or document will be split into tokens (Russell, 2011). Tokens are separate words and are mostly the result of splitting the sentence on whitespaces. The tokenization process is fundamental to further analysis and extraction of higher-level information (Zhao, 2013). This means it is one of the first steps that should be conducted during the preprocessing of the data. This process also removes all non-text characters, but aims to keep meaningful punctuation marks, like abbreviations, intact (Hotho et al., 2005).

5.5.2.2 Filtering

The removal of stop words can be considered as a filtering method and is also called feature selection (Feldman & Sanger, 2007). Filtering removes irrelevant words from the dictionary of the document collection. Stop words are common words that add no information to the text, i.e. pronouns, prepositions and conjunctions (Gupta & Lehal, 2009). Apart from stop words, it is useful to remove a set of domain specific words which can be captured in a domain vocabulary (Sharma, 2012). This also could be terms which are used to find the social data, as it will occur in every individual document.

5.5.2.3 Stemming

After the tokenization and the filtering process, it is possible to convert each token to a standard form (Zhao, 2013). This process is called stemming or lemmatization. Each stem can capture many variations and synonyms of that token, but also words that are syntactically similar (Hotho et al., 2005)(Gupta & Lehal, 2009).

Stemming allows the reduction of the variation of tokens, because stemmed tokens are counted as occurrences of the stem. When the stemming process is only dictionary-based, it can occur that tokens get stemmed incorrectly, as some tokens can only be identified correctly by the means of rule-based stemming (Zhao, 2013). This type also takes grammar into account and will contain fewer errors than dictionary-based stemming.

5.5.3 Analysis of social activity posts

After the posts have been preprocessed, there is a collection of techniques that consist of the processing step within the complete process of hybrid clustering, as seen in Figure 15 and Figure 17. These techniques can be subdivided in the following parts: classification, extraction and summarization. Classification is used to create the additional attributes: age, gender and sentiment. Extraction is needed to gain geo-location data from posts. Lastly summarization is used to identify important terms or important posts. This will hold the added value of the social activity data. First the classification of age and gender will be discussed. These two classifications appear together in most papers, as their method depends on the same input.

5.5.3.1 Age and gender

Technique	Explanation	Literature
Multiclass classification through logistic and linear regression	Assign ages and life stages to Twitter users	(Nguyen et al., 2013)
Closed-vocabulary word-category analyses	Exploration of language that distinguishes people	(Schwartz et al., 2013)
Feature based machine learning using support vector machine algorithm and decision trees	Determining age, gender, native language or personality type of author by studying their sociolect aspect	(Santosh et al., 2013)
Stacked-SVM-based classification algorithms	Classify latent user attributes, including gender, age, regional origin, and political orientation solely from Twitter user language	(Rao et al., 2010)

Table 8. Predicting age through machine learning techniques, applied to posts.

Determining the age and gender through the means of a post/document requires a machine learning algorithm. This type of algorithm will need training data in order to able to assign the most likely age and gender to the owner of the post, based on the terms and emoticons that are used. This means that the first step of this analysis method is to retrieve tweets from users with prior knowledge about age and gender. The papers displayed in Table 8 and Table 9 used classifiers based on feature models which were processed by the aid of Support Vector Machines (SVM).

So-called features are used to describe an occurrence of a word and the weight that is associated with it. Some features might have a strong correlation with the female gender, meaning that if such a feature is present in the posts, it will receive a certain bonus for being a post from a female. SVM can be used to process the features in order to classify posts in a binary way. In the paper by Rao et al. (2010) an accuracy of 72.23% was achieved for classifying the gender and an accuracy of 74.11% for classifying the user above or below the age group of 30. Miller (2012) used the Naïve Bayes and Perceptron algorithms, which produced accuracy, balanced accuracy, and F-measure between 75% and 99% for gender classification.

Technique	Explanation	Literature
Feature based machine learning using SVM algorithm and decision trees	Determining age, gender, native language or personality type of author by studying their sociolect aspect	(Santosh et al., 2013)
Stacked-SVM-based classification algorithms	Classify latent user attributes, including gender, age, regional origin, and political orientation solely from Twitter user language	(Rao et al., 2010)
Perceptron and Naïve Bayes	Identify the gender of users on Twitter	(Miller, 2012)

Table 9. Predicting gender through machine learning techniques, applied to posts

5.5.3.2 Sentiment

As the papers in Table 10 show, the classification of sentiment requires training data. This enables the model to assign sentiment value to terms and emoticons that occur in a post. The total value for sentiment will then determine if a post is categorized as positive, neural or negative. There exist data sets that contain words labeled with their sentiment value. It is possible to create such a set, but this will require manual labeling which will be a labor intensive process. Every document d has a sentiment value s that is the result of the sentiment values s of all the terms t that are part of that document.

$$\text{Sentiment}(\text{t}, \text{d}) \in \{positive, neutral, negative\}$$

Meaning that, if the sentiment for a document is considered to be positive. The sentiment for the terms that are contained within the document should also be positive.

$$Sentiment(t, d) = positive \implies \forall_s [s \in d \implies Term(s, d) = positive]$$

Technique	Explanation	Literature
Classification model	Model the sentiment expressed in social media posts	(Schweidel & Moe, 2014)
Supervised classification using features	Detecting the sentiment of Twitter messages	(Kouloumpis et al., 2011)
Multinomial Naïve Bayes Classifier	Determine positive, negative and neutral sentiments for a document	(Pak & Paroubek, 2010)
Part of Speech tagging	Automatically detect sentiments on Twitter messages	(Barbosa & Feng, 2010)
LDA based model	Distill foreground topics and filter out longstanding background topics.	(S. Tan et al., 2014)

Table 10. Assign sentiment value to each post

5.5.3.3 Geographic data

The extraction of location or so-called geo-location data mostly involves a social channel specific API. Almost all papers in Table 11 stated having used the dedicated API to retrieve geo-location data from posts. These API's specifically contain a method that returns geo-location data that is linked to a post. This data will be returned as coordinates, which allows similarity measures to be applied more easily. Not all posts will contain geo-location data, which could be resolved in some cases where a user might explicitly refer to a location based check-in website in their post.

Technique	Explanation	Literature
Topic Modeling to Classify Patterns	Technique to analyze large-scale geo-location data from social media to infer individual activity patterns	(Hasan & Ukkusuri, 2014)
Spectral Clustering	Providing representations of the dynamic areas that comprise the city	(Cranshaw et al., 2012)
Spatiotemporal data cube model	Present a scalable computational framework to harness massive location-based social media data	(Cao et al., 2015)
Spatial and statistical analysis	Geographic location data was harvested through the corresponding API. Results were then mapped by using a geographic information system (GIS)	(Zhai et al., 2015)

Table 11. Location data that can be extracted from posts

5.5.3.4 Summarization

Technique	Explanation	Literature
LDA	Coping with sparsity and document quality in microblogs	(Naveed et al., 2011)
LDA	Identify interesting topics for people	(Weng et al., 2010)
LDA & k-means	Automatically clustering and classifying Twitter messages	(Rosa et al., 2011)
LDA & Hidden Topic Markov Models	Modeling the topics of words in a document as a Markov chain	(Gruber et al., 2007)
Near-duplicate clustering	Exploratory search application for Twitter	(O'Connor et al., 2010)
Mutual reinforcement model based sub-topic summarization	Time-line based framework for topic summarization in Twitter	(Hou & Yeung, 2012)
Document categorization, keyword extraction, keyword association	Propose multiple approaches for term weighting in short documents, of which the performance is compared against existing methods	(Timonen, 2013)
Concept-based optimization framework	Explore a variety of text sources for summarizing the Twitter topics	(F. Liu et al., 2011)
Phrase Reinforcement algorithm and Hybrid TF-IDF Summarization	Process collections of short posts on specific topics on the well-known site called Twitter and create short summaries	(Sharifi, Hutton, & Kalita, 2010a)
Phrase Reinforcement algorithm	Developed an algorithm that provides an automatically created summary of the posts related to a term	(Sharifi et al., 2010b)
Word-based and symbol-based features	Propose a novel speech act-guided summarization approach	(Zhang, Li, Gao, & Ouyang, 2013)
Heuristic (stream-based) approach	Propose a new summarization task, called sequential summarization, which aims to provide a serial of chronologically ordered short sub-summaries for a trending topic	(Gao, Li, Cai, Zhang, & Ouyang, 2014)

Table 12. Topic extraction

As seen in Table 12, there is no clear dominating technique that is used for the summarization or topic classification of posts. However, Gao et al. (2014) do state that the LDA method is widely used in topic modeling. This method is seen twice when coping with short documents e.g. posts on microblogs like Twitter and four times in total.

$$\begin{array}{cccc} & T_1 & \dots & T_m \\ D_1 & d_{11} & \dots & d_{1m} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ D_n & d_{n1} & \dots & d_{nm} \end{array}$$

It is possible to display all the terms T that occur in the collection of documents (posts) D in a document/term frequency matrix, shown above (Aggarwal & Zhai, 2012). Here each entry (n, m) can be seen as the frequency the n-th term occurs in the m-th document. A problem that occurs with such a matrix, is the fact that it is very sparse, meaning that for a lot of entries the term frequency will be 0. Instead of this approach, O'Connor et al. (2010) treated the corpora containing all documents as one giant document. This resulted in the TF for one giant document consisting of the concatenation of all query subcorpus messages.

Hou & Yeung (2012) introduce the concept of term frequency bursty in their paper. Term frequency bursty is most correlated to topic evolution and can be used to identify sub-topics along with their associated words. When TF is larger than two times the mean, the period is considered bursty.

To denote a form of distance between posts, like the calculation of the Euclidian distance between data points, the Jaccard distance can be used to show how similar documents are to each other. Meaning that using the Jaccard Distance like the Euclidian distance, allows clustering posts with the k-means algorithm (Shameem & Ferdous, 2009). The Jaccard distance allows to find the most important information as the number of similar tweets could be used to indicate the importance of the content (Tao et al., 2013). The Jaccard distance measures the dissimilarity between two sets and is defined as the difference of the size of the union and the intersection of two sets divided by the size of the union of the sets (Russell, 2011).

$$Dist(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Tweets are considered to be an unordered set of words, bag of words representation. Considering a document as a bag of words implies that the order in which the words occur can be ignored (Gruber et al., 2007). When the Jaccard Distance $Dist(A, B)$ is calculated, the outcome will have the following meaning:

- Value is small if tweet A and B are similar
- Value is large if tweet A and B are dissimilar
- Value is 0 if tweets are the same
- Value is 1 if the tweets have no overlapping words at all

Before clustering tweets, using the Jaccard distance, it is useful to group near-duplicate posts into one post, as near-duplicated or duplicates do not add any information to the total collection of posts (O'Connor et al., 2010). Tao et al. (2013) have clearly defined the levels of duplication and found that one fifth of the items in their test set were duplicates, meaning that it can be beneficial to apply this step before clustering posts. As stated above, the Jaccard distance is 0 for tweets that are the same, thus duplicates. Near-duplicates are really close to duplicates, only slight variations in character use, anything more than that is considered similar, but not a near-duplicate.

The above approach attempts to use each post as a bag of words or individual terms. To avoid spelling errors and enable better representation of acronyms and emoticons word level n-grams can be used (Miller, 2012). This requires each post to be stored as multi-word term pairs, in order to be able to utilize the n-grams in a classification or clustering algorithm. The "n" in n-grams represents the amount of words that are stored as a single feature. This approach takes the context of individual terms into account, in contrast to the bag of words representation.

It is common to use the words uni- bi- and trigrams for n = [1, 2, 3] of which unigrams can be seen as the regular TF (Zhang et al., 2013). The uni- bi- and trigrams for the sentence "*We then repeat this procedure*." will be given as an example of this technique in

Unigram	{"We", "then", "repeat", "this", "procedure"}
Bigram	{["We", "then"], ["then", "repeat"], ["repeat", "this"], ["this", "procedure"]}
Trigram	{["We", "then", "repeat"], ["then", "repeat", "this"], ["repeat", "this", "procedure"]}

Table 13.

Unigram	{"We", "then", "repeat", "this", "procedure"}
Bigram	{["We", "then"], ["then", "repeat"], ["repeat", "this"], ["this", "procedure"]}
Trigram	{["We", "then", "repeat"], ["then", "repeat", "this"], ["repeat", "this", "procedure"]}

Table 13. Uni- bi- and trigram word examples

A method to summarize the posts within a cluster is proposed by Rosa et al. (2011) which state that taking the representative posts or top few tweets in the collection of a cluster is an effective way of summarizing. This was achieved by implementing the novelty selection technique presented during the TREC 2010 Web Track. This algorithm is based on selecting the highest ranked posts based on their TF-IDF similarity with a centroid. Their TF-IDF similarity should be some specified threshold.

5.6 Archetypes

As mentioned in paragraph 4.4, archetypes can be seen as a kind of customer profile. They are abstractions of the information that is found in the social activity and historic customer data. More concretely, they are the result of data selection, preprocessing, classification and clustering algorithms and represent a type of customer. In this paragraph the details of the archetypes will be discussed, including the concrete method of combining the different archetypes in order to create the hybrid clusters.

5.6.1 Validity and number of clusters

After the process of assigning customers and posts to clusters, the validity of the clusters have to be evaluated. The evaluation should be conducted because the clustering algorithm may have created clusters that are meaningless. A badly chosen amount of clusters could be the cause of such a problem (Silva et al., 2013). This could lead to the creation of groups that do not represent the actual groups that are present in the data. It is recommended to use validity indexes to find the optimal amount of clusters, although the exact right amount of clusters does not exist in general (Peters et al., 2013). Apart from the number of clusters being badly chosen, the situation can arise in which there are no logical groups to be identified at all. In this case the clustering algorithm will also identify clusters that hold no value for segmentation.

5.6.1.1 Internal and external validity

There can be made a distinction between Internal and external validity (Yao et al., 2014). External validity indicates how well the results are generalizable (Hiziroglu, 2013). Whereas internal validity can be used to measure how effective a clustering or classification algorithm has been implemented. Internal validity can be subdivided into compactness and separation. Compactness measures the distance between members of a clusters and separation measures the distance between other clusters (Seret et al., 2015). Calinski-Harabasz internal cluster validity index and the Rand external validity index can be used to calculated the validity (Kandeil et al., 2014).

5.6.1.2 Population and sample

When the source of the data is considered and the generalizations that are made based on that data, it is important to note that only the customers who use social channels or are part of the historic data set can be analyzed. Meaning that only a subset of the total population of customers is considered. This subset is called a sample (Kline, 2013). Only a representative sample can accurately reflect the entire population. Each customer in the population should have an equal chance of being part of the sample. Normally this is done by the means of random samples, which implies that all observation should have an equal likelihood of appearing in the sample.

5.6.2 Attributes archetypes

Based on the papers and the techniques that were discussed in paragraph 5.5.3, the following attributes and values can be identified for the social activity data archetypes. Each attribute is explained and followed by a concrete example of the value it can take in Table 14.

Attribute	Explanation	Example value
Character count	Amount of characters occurring in the post	46
Terms	Terms occurring in the raw post	MiedemaM: RT @Computeridee: #Windows10 is er!
Tokenized post	Resulting terms after preprocessing	rt computeridee windows10 is er
Sentiment value	Value of sentiment ϑ, ranging from $-1 \leq \vartheta \leq 1$	0.62
Sentiment class	Assigned sentiment class, based on value of n e.g. sentiment is positive for $\vartheta \geq 0.5$	Positive
Gender value	Score m that determines if someone is more likely male or female, ranging from $0 \leq \rho \leq 1$	.43
Gender	Assigned gender class, based on value of m e.g. gender is female for $\rho \geq 0.5$	Male
Age value	Score u that determines the estimated age based on the terms t in a post D	32.54
Age group	Assigned gender class, based on value of m e.g. age group is 30 - 40 for $0.4 \leq u \leq 0.5$	30 - 40
Location	Extracted geo-location data in the format (latitude, longitude)	52.349418, 5.1903040
Cluster	Cluster to which the post got assigned e.g. {1,…,K}	4
Cluster topics / summary	Collection of terms that describe the most important terms within a cluster of posts	Windows 10 ready free full upgrade

Table 14. Attributes social archetype

For sentiment, gender and age there exist two attributes, one that contains the exact value that resulted from the classification algorithm and another that contains the class to which the post got assigned based on that value. Numeric values are more suitable for clustering and also give an indication of how posts with the same class differ from another. If it is required, multi-valued attributes should be split into two separate attributes e.g. location latitude and longitude.

5.6.3 Combining archetypes

To process of forming hybrid clusters is based on the fact that after the processing techniques, shown in Figure 17 and paragraph 5.5.3, shared attributes exist between the archetypes based on social activity data and historic customer data. These shared attributes could only be acquired through techniques that created additional data based on supervised classifying algorithms. Without these shared attributes, there is no method that can automatically combine both archetypes. This is only possible when this combination is based on intersected attributes from h and g. When such a combination is made, the resulting set of data is called a hybrid cluster. Hybrid since the resulting set consists data from 2 different sources and cluster since the objects that are part of the archetypes are grouped by clustering algorithms.

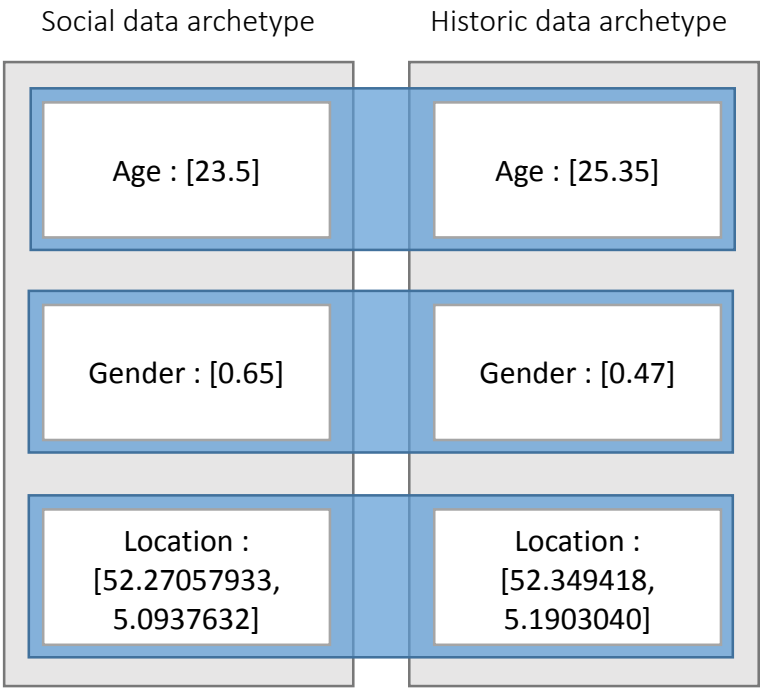

Figure 21. Comparison between attributes

The attributes shown in Figure 21, the attributes from social and historic can be displayed as the following row vectors, using the values used above.

$$h = \{23.5, 0.65, 52.27057933, 5.0937632\} \qquad g = \{25.25, 0.47, 52.349418, 5.190340\}$$

To compare the archetype attributes the Euclidean distance can be calculated and used to combine archetypes when this distance is below a certain threshold defined by θ. The Euclidean distance is calculated in the following way,

$$d(h_i, g_j) = \sqrt{(h_{age} - g_{age})^2 + (h_{gender} - g_{gender})^2 + (h_{location} - g_{location})^2}$$

In order to avoid attributes that have large values to influence those with small values, normalization should be applied (Larose, 2005). There a various normalization techniques in order to establish this. Since the minimum and maximum values of the attributes are unknown, Z-Score normalization is seen as a suitable technique (Patel & Mehta, 2012). The formula for this technique can be formalized in the following way,

$$y = \frac{x - mean(X)}{SD(X)}$$

In this formula, x is the value that needs to be normalized, $mean(X)$ is the mean of all attributes in the set X and $SD(X)$ is the standard deviation for the set X.

5.7 Output hybrid clustering

After executing all the processes described in Figure 15, the result will be a collection of combined sets of attributes from two different data sources. These sets did not have any prior shared attributes that would allow a meaningful connection between them. Based on the attributes: age, gender and location, this connection was made possible. There are, however, three scenarios in which these shared attributes might not result in a meaningful connection.

- Age and gender do not correlate in a positive or negative way in relation to the service or product that is provided by the enterprise. Meaning that the segments based on historic customer data will not contain significant differences when age and gender are taken into account. The same problem holds for location.
- Posts that contain limited user content are difficult to classify (Schwartz et al., 2013)
- Not all posts will contain location data, which results in a weaker link (Hasan & Ukkusuri, 2014)

6 Validation of framework

6.1 Framework outline

This paragraph serves as a quick summary of the most important concepts that are part of the framework presented in chapter 4. Table 15 lists these concepts together with a brief explanation.

Attribute	Explanation
Segmentation	The process of dividing a set of objects into smaller sets of objects with similar characteristics. When these objects are customers, the process can be described as dividing the customer base into smaller homogenous groups.
Customer	An entity within the database that describes a real life person through attributes and corresponding values. There can be made a distinction between existing and potential customers.
Historic customer data	Values assigned to attributes that exist for each customer.
Social activity data	Values assigned to attributes that exist for each post.
Archetype	The result of data selection, preprocessing, classification and clustering algorithms and represent a type of customer. It exist for both historic customer data and social activity data, meaning that one collection is based on existing customers and the other one is based on posts. Archetypes from both sources have similar attributes, making comparison and combing possible.
Clustering	An unsupervised data mining technique that groups objects in homogenous groups according to logical relationships. This is done in such a way that objects from the same cluster are similar and objects from different clusters are dissimilar. Objects have no class label before the clustering process starts.

Table 15. Framework concepts

As seen in Figure 22, the concepts explained above are placed into context. Both data sources start separate from each other. After the first step, clusters are created through the means of clustering algorithms. These algorithms enable similar objects to be grouped together. These groups or clusters are the segments that will be used to target marketing campaigns on. The second step involves selection of relevant data for the historic customer data and attribute creation though classification and extraction for the social activity data. When both archetypes contain all relevant data e.g. cluster and similar attributes, the comparison can be made, which allows to combine both archetypes based on similarity. When archetypes are similar enough they can be combined, creating so called hybrid clusters.

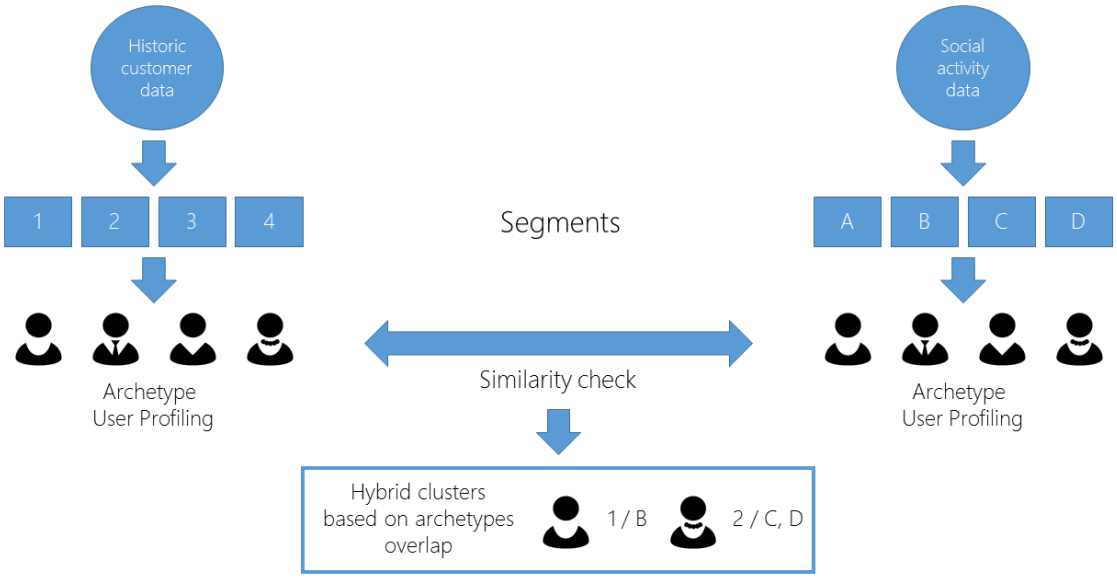

Figure 22. Framework visualization

6.2 Comparison requirements

In chapter 3 the requirements of hybrid clustering were discussed. This chapter serves as a validation of the presented framework in chapter 4 and the more detailed explanation of the techniques that could be used to reach the required situation described in the framework.

The requirements were aimed to overcome the weakness of the current implementation that aims to give insights in social channels. This implementation, called social engagement, had three main weakness.

1. It does not provide detailed information on potential customers expressing themselves in social channels
2. Lack of integration with the information currently available about customers, historic customer data
3. Only monitors trends that are related to manually set terms

6.2.1 Detailed information on potential customers

Social engagement provided the user with general information that was based on manually set terms e.g. brand name or products. This information can be seen as general since it consisted of a word cloud, meaning that frequent occurring terms were shown. The most important providers of this information were also shown e.g. users with the highest amount of posts related to the set terms.

With hybrid clustering it has become possible to add summarizations of social activity data to customer segments. These summarizations provide more information than the previous word cloud and in addition to this the summarizations are grouped together based on similarity, meaning that each social segment will have a unique set of terms associated with it.

6.2.2 Lack of integration with historic customer data

The previous implementation was purely aimed at social activity data. Meaning that this source of information had no formal implementation within the existing customer segmentation strategy. By adding attributes to the social activity data, it was possible to create a means of combining historic customer data and social activity data. This creates the opportunity to combine the segments based on both data sources.

6.2.3 Manually set terms

It is still required to manually set terms to retrieve social activity data. For each give term, posts will be returned that contain that term. Since these terms are bound to the enterprise implementing the solution this step requires a manual action. Aside from this, it would be strange to use information besides your own products and brand. General trending topics can be monitored without directly using this information. This is also possible with the twitter package “TwitteR” that was used in RStudio. It can even return the tending topics within a certain location.

7 Conclusion

The goal of this paper was to find a service which would be able to combine social activity data and historic customer data for companies focusing on B2C. This needed to be done in such a way that marketers could focus on one type of segmentation, instead of both customer segments based on historic customer data and social trends that arise in the social engagement service. It also needed to overcome the existing shortcomings of the social engagement service.

7.1 Main and sub questions

This paper aimed to answer the following main question: *How should enterprises implement customer analytics to be able to segment customers in a more detailed way based on both historical customer data and social activity data in order to deploy more focused B2C campaigns?*

In chapter 2 the main concepts and relevant background literature was introduced. The main topics that were discussed consisted of customer analytics, data sources, data mining, text mining, information retrieval and natural language processing. This provided the main background literature for this paper. Based on the requirements in chapter 3 and additional literature the conceptual framework in chapter 4 was formed. This framework organizes and defines the concepts that are involved with hybrid clustering. An extensive literature research on existing customer segmentation solutions and classification techniques for social activity data resulted in chapter 5, where the concepts are explained in more detail and are supported by mostly recent literature. Each of the processes within the preprocessing and processing steps of the complete model presented in paragraph 5.2 is discussed. Lastly the presented model and framework are validated in chapter 6 based on the formulated requirements in chapter 3.

To give an answer to the main question of this paper. The following sub questions were identified and answered throughout the paper.

1. How can different segmentations be combined?
2. Does the combination of segmentations based on historical customer data and social activity data provide more detailed customer segmentations?
3. How will the combination of historical customer data and social activity data enable more focused B2C campaigns?

Sub question 1 will be answered in paragraph 7.1.1, sub question 2 will be answered in paragraph 7.1.2 and sub question 3 will be answered in paragraph 7.1.3.

7.1.1 Different segmentations combined

As described in paragraph 4.3, each set of customers has a set of attributes with corresponding values, this set of customers is called the historic customer data. There also exists a set of objects containing posts, which is described by a set of terms. The set of customers and the set of objects containing posts can only be combined when they share one or more of the same attributes. This would allow a union of the attributes for both objects when the intersection is non-empty. Meaning that an object in the historic customer data, a customer, and an object in the social activity data, a post, will be combined.

To achieve this non-empty meaningful intersection, additional attributes had to be added to the objects in the social activity data. For this reason intermediary customer profiles called archetypes were created, as discussed in paragraph 4.4. These archetypes were formed for both data sources. The social archetypes contain the social activity data objects and the added attributes. Aside from the posts consisting of terms, the attributes age, gender and location were found through classification and extraction techniques. A social archetype exists for each cluster found in the social activity data. This was discussed in paragraph 5.6.

Likewise, the historic archetypes contain a selection of domain relevant data, used for clustering, in addition to the demographic attributes age, gender and location. Each cluster found in the historic customer data will also result in an archetype. The historic and social archetypes create a means for intersecting and combining both sets. This can be done based on the similarity that is found between the shared attributes in the archetypes, as discussed in paragraph 5.6.3.

Instead of combining separate objects within the archetypes, an entire group of objects, the complete archetype, will be combined. This would mean that the average of the archetype's attributes is taken instead of individual values of objects. The resulting combinations are the so called hybrid clusters which are the products of combining archetypes from different sources of data and are achieved through clustering algorithms.

7.1.2 More detailed customer segments

The segmentations are essentially the formation of clusters through clustering techniques like SOM and k-means, see paragraph 5.4.2 and 5.5.3. These are applied to each individual source of data e.g. the historic customer data and social activity data. Meaning that the segments are essentially the same during the first step of segmentation. Which would imply that the customer segments are not more detailed.

There is however a second step that is focused on combining archetypes, which are essentially segments. This step adds the archetype from the historic customer data and social activity data together, the process is described in paragraph above. Since this step does add data to the regular customer segments, one can argue that the customer segments have become more detailed since it contains more information that can be utilized. The segments have not become more detailed in the sense that they became smaller. This is completely dependent on the number of clusters k.

7.1.3 More focused B2C campaigns

In the past, campaigns haven't been tailored to specific customer groups because of a lack of understanding of their behavior. Nowadays this has completely changed due to an increase in research, availability of customer data, computing power and competition. The direct goals for campaigns differ widely, but altogether aim to maintain good customer relationships and enhance customer value, as discussed in paragraph 2.1.1. It is important for campaigns to have a clear, preferably measurable goal and customer target audience in order to be effective.

The social information that will be available together with the historic customer data in the archetype will allow marketers to get a better understanding of their target audience, since it provides additional information that would first have to be interpreted from the social channels directly instead of through hybrid clusters.

The fact if these hybrid clusters actually cause campaigns to be more effective has to be tested through an A/B test. Meaning that one campaign will be set up using the old method (A) and another campaign will be set up using the new method (B) they should have the same goal, only the information on the target audience should be different. There should be a measure that defines the campaigns effectiveness e.g. amount of clicks for an e-mail campaign or conversion rate in a webshop. This measure can then be compared for campaign A and B in order to decide which one has been the most effective.

8 Discussion and future work

In this paper the focus was put solely on Twitter messages, this was because a lot of research had already been done on Twitter and the information that it could provide. It was also stated that Twitter is one of the most accessible sources of user-generated data. Aside from this, the API proved to be very user-friendly and the existing packages for Twitter that were available for RStudio were well documented. In principle the same should at least hold for Facebook as this has a similar API and also has packages for RStudio.

During the extraction process with the TwitteR package for RStudio, it was possible to set the desired language for the tweets. In this paper the focus was on Dutch and English tweets, which resulted in sufficient tweets. Using the TwitteR package, it required no effort to retrieve 5000 posts given a term. When using the ISO 639-1 codes for languages, tweets from almost all languages can be retrieved. Experiments with Russian and Korean tweets proved no problem. Since the Jaccard distance cannot be used between different language, each language will have to be separated. As users from other counties are likely to have different topics and trends that need to combined with other historic customer data, this seems like a logical step.

The literature that was used in paragraph 5.5.3.1 which discussed the determination of age and gender through the classification of tweets focused on general tweets. It did not focus on a specific kind of tweets like the type that is used in this paper. Research or experiments would have to be conducted in order to verify that tweets within the brand/product context hold the same type of information as the tweets used in the papers.

There has been made the assumption that these is some kind of correlation between the products and services an enterprise is providing and the age, gender and location of their existing and potential customers. This would mean that if there does not exist such a correlation, the connection made through the archetype could be meaningless after all. In such a case the only means of making a meaningful connection is through manual labeling. For this approach the archetypes can still be utilized, but lose the value of being able to make a non-empty intersection.

In order to create a simple means of combining different segments or clusters, crisp clustering algorithms were used. The advantage this brings is the very strictly defined clusters either containing an object or not. It might however occur that this will result in objects that do not necessarily belong in the cluster they are assigned to. Having stated this, it will be interesting to conduct further research to see how fuzzy clustering could improve this problem.

In the paper θ is used to define the level of similarity that is need between the shared attributes of the archetypes in order to combine them through a union. Further research should aim to conduct experiments on what value delivers useable results, meaning that the amount of combination is such that at least half of the historic customer data archetypes have a social activity archetype.

In paragraph 4.6, a brief explanation is given of how the visualization of hybrid clusters should work ideally. User test should be conducted with marketers in order to determine if this proposed visualization is really something they would work with. It provides a really high level representation that also needs to be worked out for lower levels e.g. levels that show more detailed information.

9 References

9.1 Academic references

Aggarwal, C., & Zhai, C. (2012). A survey of text clustering algorithms. *Mining Text Data*. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4614-3223-4_4

Ahola, J., & Rinta-runsala, E. (2001). *Data Mining Case Studies in Customer Profiling*. *Research report TTE1- 2001-29*.

AlFalahi, K., Atif, Y., & Abraham, A. (2014). Models of Influence in Online Social Networks. *International Journal of Intelligent Systems*, *29*(2), 1–23. doi:10.1002/int

Baker, P. (2010). Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender and Language*, *4*(1), 125–149. doi:10.1558/genl.v4i1.125

Barbosa, L., & Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. *Proceedings of the 23rd International Conference on …*, (August), 36–44. Retrieved from http://dl.acm.org/citation.cfm?id=1944571

Bellot, P., Bonnefoy, L., Bouvier, V., Duvert, F., & Kim, Y. (2014). *Innovations in Intelligent Machines-4*. (C. Faucher & L. C. Jain, Eds.) (Vol. 514). Cham: Springer International Publishing. doi:10.1007/978-3-319-01866-9

Berry, M. J. a., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. *Portal.Acm.Org*. Retrieved from http://portal.acm.org/citation.cfm?id=983642

Bijmolt, T. H. A., Leeflang, P. S. H., Block, F., Eisenbeiss, M., Hardie, B. G. S., Lemmens, A., & Saffert, P. (2010). Analytics for Customer Engagement. *Journal of Service Research*, *13*, 341–356. doi:10.1177/1094670510375603

Blythe, J. I. M. (2008). *Essentials of marketing* (3rd ed.). Pearson Education.

Brito, P. Q., Almeida, C. S. S., Monte, A., & Byvoet, M. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing*, 1–8. doi:10.1016/j.rcim.2014.12.014

Canhoto, A. I., Clark, M., & Fennemore, P. (2013). Emerging segmentation practices in the age of the social customer. *Journal of Strategic Marketing*, *21*(5), 413–428. doi:10.1080/0965254X.2013.801609

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2015). A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems*, *51*, 70–82. doi:10.1016/j.compenvurbsys.2015.01.002

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications*, *34*(4), 2754–2762. doi:10.1016/j.eswa.2007.05.043

Chiu, C.-Y., Chen, Y.-F., Kuo, I.-T., & Ku, H. C. (2009). An intelligent market segmentation system using k-means and particle swarm optimization. *Expert Systems with Applications*, *36*(3), 4558–4565. doi:10.1016/j.eswa.2008.05.029

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, *12*(Aug), 2493–2537. Retrieved from http://arxiv.org/abs/1103.0398

Corrigan, H. B., Craciun, G., & Powell, A. M. (2014). How Does Target Know So Much About Its Customers? Utilizing Customer Analytics to Make Marketing Decisions. *Marketing Education Review*, *24*(2), 159–166. doi:10.2753/MER1052-8008240206

Cranshaw, J., Hong, J. I., & Sadeh, N. (2012). The Livehoods Project : Utilizing Social Media to Understand the Dynamics of a City. *Icwsm*, 58–65.

Cuadros, A. J., & Domínguez, V. E. (2014). Customer segmentation model based on value generation for marketing strategies formulation. *Estudios Gerenciales*, *30*(130), 25–30. doi:10.1016/j.estger.2014.02.005

Davenport, T. H. (2006). Competing on Analytics Competing on Analytics. *Harvard Business Review*, *84*(1), 98–107.

Deng, X., Jin, C., Higuchi, Y., & Han, J. C. (2011). An efficient hybrid clustering algorithm for customer segmentation in mobile e-commerce. *ICIC Express Letters*, *5*(4 B), 1411–1416. doi:10.4018/jeco.2013040105

Dhandayudam, P. (2012). An improved clustering algorithm for customer segmentation. *International Jouranl of Engineering Science and Technology*, *4*(02), 695–702.

Dunk, A. S. (2004). Product life cycle cost analysis: The impact of customer profiling, competitive advantage, and quality of IS information. *Management Accounting Research*, *15*(4), 401–414. doi:10.1016/j.mar.2004.04.001

Engels, G., Förster, a, Heckel, R., & Thöne, S. (2005). Process Modeling using UML. *Process-Aware Information Systems: Bridging People and Software Through Process Technology*, 85–117. doi:10.1002/0471741442.ch5

Fan, W., & Bifet, A. (2013). Mining Big Data : Current Status , and Forecast to the Future. *ACM SIGKDD Explorations Newsletter*, *14*(2), 1–5. doi:10.1145/2481244.2481246

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook - Advanced Approaches in Analysing Unstructured Data*. Cambridge University Press.

Gao, D., Li, W., Cai, X., Zhang, R., & Ouyang, Y. (2014). Sequential summarization: A full view of twitter trending topics. *IEEE Transactions on Audio, Speech and Language Processing*, *22*(2), 293–302. doi:10.1109/TASL.2013.2282191

Geffet, M., & Dagan, I. (2009). Bootstrapping Distributional Feature Vector Quality, (November 2008). doi:10.1162/coli.08-032-R1-06-96

Germann, F., Lilien, G. L., Fiedler, L., & Krausd, M. (2014). Do Retailers Benefit from Deploying Customer Analytics ? *Journal of Retailing*, *90*(4), 587–593. doi:10.1016/j.jretai.2014.08.002

Grishman, R. (1997). Information Extraction : Techniques and Challenges Why the Interest in Information Extraction ? *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, 10–27.

Gruber, a, Rosen-Zvi, M., & Weiss, Y. (2007). Hidden topic Markov models. *Proceeding of the International Conference on Artificial Intelligence and Statistics*, 163–170. doi:10.1145/1143844.1143967

Grunbaum, B. (1984). The Construction of Venn Diagrams. *The College Mathematics Journal*, *15*(3), 238–247.

Guerra, L., McGarry, L. M., Robles, V., Bielza, C., Larrañaga, P., & Yuste, R. (2011). Comparison between supervised and unsupervised classifications of neuronal cell types: A case study. *Developmental Neurobiology*, *71*(1), 71–82. doi:10.1002/dneu.20809

Gullo, F., Domeniconi, C., & Tagarelli, A. (2013). Projective clustering ensembles. *Data Mining and Knowledge Discovery*, *26*(3), 452–511. doi:10.1007/s10618-012-0266-x

Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76. doi:10.4304/jetwi.1.1.60-76

Hand, D. J., Blunt, G., Kelly, M. G., & Adams, N. M. (2000). Data Mining for Fun and Profit. *Statistical Science*, *15*(2), 111–131. Retrieved from http://projecteuclid.org/euclid.ss/1009212753

Hasan, S., & Ukkusuri, S. V. (2014). Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, *44*, 363–381. doi:10.1016/j.trc.2014.04.003

Hiziroglu, A. (2013). Soft computing applications in customer segmentation: State-of-art review and critique. *Expert Systems with Applications*, *40*(16), 6491–6507. doi:10.1016/j.eswa.2013.05.052

Hotho, A., Andreas, N., Paaß, G., & Augustin, S. (2005). A Brief Survey of Text Mining. *Ldv Forum*, *20*(1), 19–62.

Hou, Z. M., & Yeung, H. (2012). Twitter Topic Summarization by Ranking Tweets Using Social Influence and Content Quality, *4*(96), 763–780.

Jackson, P., & Moulinier, I. (2007). *Natural Language Processing for Online Applications* (Vol. 5). Amsterdam: John Benjamins Publishing Company. doi:10.1075/nlp.5

Josiah, A., Ikenna, O., Jennifer, A., Chinaedum, I., Justina, R., & Nnamonso, A. (2015). The Relevance of Analytical CRM and Knowledge Management in an Organisation : A Data Mining Structure, *4*(2), 208–215.

Kandeil, D. A., Saad, A. A., & Youssef, S. M. (2014). A Two-Phase Clustering Analysis for B2B Customer Segmentation. *2014 International Conference on Intelligent Networking and Collaborative Systems*, 221–228. doi:10.1109/INCoS.2014.49

Kao, A., & Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. Springer.

Karna, N., Supriana, I., & Maulidevi, U. (2014). Social CRM using Web Mining, (November), 24–27.

Keupp, M. M., & Gassmann, O. (2009). Determinants and archetype users of open innovation. *R and D Management*, *39*(4), 331–341. doi:10.1111/j.1467-9310.2009.00563.x

Kline, R. B. (2013). Sampling and estimation. In *Beyond significance testing: Statistics reform in the behavioral sciences (2nd ed.).* (pp. 29–65). Washington: American Psychological Association. doi:10.1037/14136-002

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480. doi:10.1109/5.58325

Koudehi, F. A., Rajeh, S. M., Farazmand, R., & Seyedhosseini, S. M. (2014). A Hybrid Segmentation Approach for Customer Value. *Journal of Contemporary Research in Business*, *6*(6), 142–152.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis : The Good the Bad and the OMG ! *Artificial Intelligence*, 538–541. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2857/3251

Kroeze, J. H., Matthee, M. C., & Bothma, T. J. D. (2003). Differentiating data- and text-mining terminology, 93–101. Retrieved from http://dl.acm.org/citation.cfm?id=954014.954024

Kuo, R. J., An, Y. L., Wang, H. S., & Chung, W. J. (2006). Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation. *Expert Systems with Applications*, *30*(2), 313–324. doi:10.1016/j.eswa.2005.07.036

Larose, D. T. (2005). *Discovering Knowledge in Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471687545

Lewis, D. D. (1994). A Comparison of Two Learning Algorithms for Text Categorization 1 Introduction 2 Text Categorization : Nature and Approaches. *In Proceeding of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR`94)*, 1–14.

Li, C. (2015). *Data Mining, Data, Data Preprocessing*. *Data Mining*. University of Texas.

Liu, F., Liu, Y., & Weng, F. (2011). Why is "SXSW " trending ? Exploring Multiple Text Sources for Twitter Topic Summarization. *Proceedings of the Workshop on Language in Social Media*, (June), 66–75. Retrieved from http://www.aclweb.org/anthology/W11-0709\nhttp://www.hlt.utdallas.edu/~feiliu/papers/LSM_2011.pdf

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, *6*(4), 393–423. doi:10.1023/A:1016304305535

Liu, X., Fu, Z., Zhou, X., Wei, F., & Zhou, M. (2010). Collective Nominal Semantic Role Labeling for Tweets, 1685–1691.

Luo, Y., Cai, Q., Xi, H., Liu, Y., & Zhu, G. (2013). Customer segmentation for telecom with the k-means clustering method. *Information Technology Journal*, *12*(3), 409–413. doi:10.3923/itj.2013.409.413

Mannila, H. (1996). Data mining: machine learning, statistics, and databases. In *Proceedings of 8th International Conference on Scientific and Statistical Data Base Management* (pp. 2–9). IEEE Comput. Soc. Press. doi:10.1109/SSDM.1996.505910

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval. Information Retrieval*. Cambridge: Cambridge University Press. doi:10.1109/LPT.2009.2020494

McCallum, A. (2005). Information extraction. *Queue*, *3*(9), 48. doi:10.1145/1105664.1105679

Miller, Z. (2012). Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. *International Journal of Intelligence Science*, *02*(24), 143–148. doi:10.4236/ijis.2012.224019

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. Cambridge, MA: MIT Press.

Nauck, D. D., Ruta, D., Spott, M., & Azvine, B. (2006). Being proactive - Analytics for predicting customer actions. *BT Technology Journal*, *24*(1), 17–26. doi:10.1007/s10550-006-0017-x

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Searching microblogs: coping with sparsity and document quality. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 183–188. doi:10.1145/2063576.2063607

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2 PART 2), 2592–2602. doi:10.1016/j.eswa.2008.02.021

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). "How old do you think I am ?": A study of language and age in Twitter. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM '13)*, 439–448.

O'Connor, B., Krieger, M., & Ahn, D. (2010). TweetMotif : Exploratory Search and Topic Summarization for Twitter. *ICWSM*, 384–385.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Lrec*, 1320–1326. doi:10.1371/journal.pone.0026624

Patel, V. R., & Mehta, R. G. (2012). Data clustering: Integrating different distance measures with modified k-means algorithm. *Advances in Intelligent and Soft Computing*, *131 AISC*(VOL. 2), 691–700. doi:10.1007/978-81-322-0491-6_63

Patterson, M. (2005). Business requirements for campaign management — A sample framework. *Journal of Database Marketing & Customer Strategy Management*, *12*(2), 177–192. doi:10.1057/palgrave.dbm.3240254

Peters, G., Crespo, F., Lingras, P., & Weber, R. (2013). Soft clustering - Fuzzy and rough approaches and their extensions and derivatives. *International Journal of Approximate Reasoning*, *54*(2), 307–322. doi:10.1016/j.ijar.2012.10.003

Rajagopal, S. (2011). Customer Data Clustering Using Data Mining Technique. *International Journal of Database Management Systems (IJDMS)*, *3*(4), 1–11. doi:10.5121/ijdms.2011.3401

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents - SMUC '10*, 37. doi:10.1145/1871985.1871993

Raykov, T., & Marcoulides, G. A. (2012). *Basic statistics: an Introduction with R*. Rowman & Littlefield Publishers.

Rehman, A., & Ali, A. R. (2014). Customer Churn Prediction , Segmentation and Fraud Detection in Telecommunication Industry, 1–9.

Rosa, K. Dela, Shah, R., Lin, B., Gershman, A., & Frederking, R. (2011). Topical Clustering of Tweets. *SIGIR 3rd Workshop on Social Web Search and Mining*, cited 2.

Russell, M. A. (2011). *Mining the Social Web* (Second Edi.). O'Reilly Media.

Rybalko, S., & Seltzer, T. (2010). Dialogic communication in 140 characters or less: How Fortune 500 companies engage stakeholders using Twitter. *Public Relations Review*, *36*(4), 336–341. doi:10.1016/j.pubrev.2010.08.004

Saarijärvi, H., Karjaluoto, H., & Kuusela, H. (2013). Customer relationship management: the evolving role of customer data. *Marketing Intelligence & Planning*, *31*, 584–600. doi:10.1108/MIP-05-2012-0055

Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). Author Profiling : Predicting Age and Gender from Blogs Notebook for PAN at CLEF 2013. *PAN - Uncovering Plagiarism, Authorship, and Social Software Misuse a Benchmarking Activity on Uncovering Plagiarism, Authorship and Social Software Misuse*, 23–27. Retrieved from http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-papers-final/pan13-author-profiling/santosh13-notebook.pdf

Şchiopu, D. (2010). Applying TwoStep cluster analysis for identifying bank customers' profile. *Buletinul*, *LXII*(3), 66–75.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, *8*(9). doi:10.1371/journal.pone.0073791

Schweidel, D. a., & Moe, W. W. (2014). Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice. *Journal of Marketing Research*, *LI*(August), 140319055406004. doi:10.1509/jmr.12.0424

Seret, A., Maldonado, S., & Baesens, B. (2015). Identifying next relevant variables for segmentation by using feature selection approaches. *Expert Systems with Applications*, (March). doi:10.1016/j.eswa.2015.01.070

Seret, A., vanden Broucke, S. K. L. M., Baesens, B., & Vanthienen, J. (2014). A dynamic understanding of customer behavior processes based on clustering and sequence mining. *Lecture Notes in Business Information Processing*, *171 LNBIP*(10), 237–248. doi:10.1007/978-3-319-06257-0

Shameem, M.-U.-S., & Ferdous, R. (2009). An efficient k-means algorithm integrated with Jaccard distance measure for document clustering. In *2009 First Asian Himalayas International Conference on Internet* (pp. 1–6). IEEE. doi:10.1109/AHICI.2009.5340335

Sharifi, B., Hutton, M. A., & Kalita, J. K. (2010a). Experiments in microblog summarization. *Proceedings - SocialCom 2010: 2nd IEEE International Conference on Social Computing, PASSAT 2010: 2nd IEEE International Conference on Privacy, Security, Risk and Trust*, 49–56. doi:10.1109/SocialCom.2010.17

Sharifi, B., Hutton, M., & Kalita, J. (2010b). Summarizing Microblogs Automatically. *Computational Linguistics*, *15*(June), 685–688. doi:10.1007/s10531-004-1065-5

Sharma, N. K. (2012). Discovering Topical Experts in Twitter Social Network Master of Technology.

Shum, S. B., & Ferguson, R. (2011). Social Learning Analytics. *Knowledge Media Institute & Institute of Educational Technology*, (June), 1–26.

Siegel, E. (2013). *Predictive analytics: The power to predicy who will click, buy, lie, or die*. John Wiley & Sons.

Silva, J. a, Faria, E. R., Barros, R. C., Hruschka, E. R., Carvalho, A. C. P. L. F. De, & Gama, J. (2013). Data stream clustering. *ACM Computing Surveys*, *46*(1), 1–31. doi:10.1145/2522968.2522981

Sirsat, S. R., & Chavan, V. (2014). Mining knowledge from text repositories using information extraction : A review, *39*(February), 53–62.

Smola, A., & Vishwanathan, S. V. N. (2008). *Introduction to machine learning*. Cambridge University Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science (New York, N.Y.)*, *103*(2684), 677–680. doi:10.1126/science.103.2684.677

Stoll, R. R. (1979). *Set theory and logic*. Courier Corporation.

Sun, S. (2009). An Analysis on the Conditions and Methods of Market Segmentation. *International Journal of Business and Management*, *4*(2), 63–70.

Susena, E. (2014). Using Data Mining Techniques In Higher Education., *4*(9), 68–72.

Tan, A. (1999). Text Mining : The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, *8*, 65–70. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.7672&rep=rep1&type=pdf

Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., … He, X. (2014). Interpreting the public sentiment variations on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, *26*(5), 1158–1170. doi:10.1109/TKDE.2013.116

Tao, K., Abel, F., & Hauff, C. (2013). Groundhog day: near-duplicate detection on twitter. *Proceedings of the 22nd International Conference on World Wide Web*, 1273–1283. Retrieved from http://dl.acm.org/citation.cfm?id=2488499

Terlunen, S., Barreto, G., & Hellingrath, B. (2015). Application and Evaluation of Multi-criteria Clustering Algorithms for Customer-Oriented Supply Chain Segmentation. In *Logistics Management* (pp. 121–133). doi:10.1007/978-3-319-13177-1_10

Timonen, M. (2013). *Term Weighting in Short Documents for Document Categorization , Keyword Extraction and Query Expansion Mika Timonen*.

Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547–12553. doi:10.1016/j.eswa.2009.05.032

Tsiptsis, K., & Chorianopoulos, A. (2010). *Data Mining Techniques in CRM: Inside Customer Segmentation*. John Wiley & Sons. doi:10.1002/9780470685815

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*(3), 586–600. doi:10.1109/72.846731

Waller, M. a, & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, *34*(2), 77–84. doi:10.1111/jbl.12010

Wang, S., Xu, C., & Wu, R. (2009). Clustering method based on fuzzy multisets for web pages and customer segments. *2008 International Seminar on Business and Information Management, ISBIM 2008*, *2*, 125–128. doi:10.1109/ISBIM.2008.171

Weng, J., Lim, E., & Jiang, J. (2010). TwitterRank : Finding Topic-sensitive Influential Twitterers. *New York*, *Paper 504*, 261–270. doi:10.1145/1718487.1718520

Wu, R. S., & Chou, P. H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, *10*(3), 331–341. doi:10.1016/j.elerap.2010.11.002

Yao, Z., Sarlin, P., Eklund, T., & Back, B. (2014). Combining visual customer segmentation and response modeling. *Neural Computing and Applications*, *25*(1), 123–134. doi:10.1007/s00521-013-1454-3

Zhai, S., Xu, X., Yang, L., Zhou, M., Zhang, L., & Qiu, B. (2015). Mapping the popularity of urban restaurants using social media data. *Applied Geography*, *63*, 113–120. doi:10.1016/j.apgeog.2015.06.006

Zhang, R., Li, W., Gao, D., & Ouyang, Y. (2013). Automatic twitter topic summarization with speech acts. *IEEE Transactions on Audio, Speech and Language Processing*, *21*(3), 649–658. doi:10.1109/TASL.2012.2229984

Zhao, Y. (2013). *Text Mining*. doi:10.1016/B978-0-12-396963-7.00010-6

9.2 Non-academic references

Gartner. (2014). *Moving On Up: Microsoft, Oracle, Salesforce.com and SAP's Social Applications for CRM.* Gartner.

Stodder, D. (2012) *Customer analytics in the age of social media.* TDWI.

Meloen, S. (2015, 04 15). Requirements hybrid clustering. (L. à. Campo, Interviewer)

9.3 Consulted websites

Analytics as a Service. (2015, 04 23). Retrieved from http://www.sas.com/en_ie/software/analytics-as-a-service.html

Avanade history. (2015, 04 23). Retrieved from http://www.avanade.com/en-us/about-avanade/about-us/avanade-history

Big Data Analytics as a Service. (2015, 04 23). Retrieved from http://www.datameer.com/blog/announcements/welcome-to-big-data-analytics-as-a-service-bdaaas.html

Cran R Project. (2015, 07 19). *kohonen: Supervised and Unsupervised Self-Organising Maps*. Retrieved from Cran R Project: https://cran.r-project.org/web/packages/kohonen/

Customer Analytics Engine. (2015, 04 23). Retrieved from http://accuracyinfolabs.com/?page_id=53

Customer churn for SMBs using Predictive Analytics As a Service. (2015, 04 23). Retrieved from http://www.simafore.com/blog/bid/128738/Customer-churn-for-SMBs-using-Predictive-Analytics-As-a-Service

Gentry, J. (2015, 07 16). *twitteR: R Based Twitter Client*. Retrieved from Cran R Project: http://cran.r-project.org/web/packages/twitteR/index.html

IBM Unites Marketing. (2015, 04 23). Retrieved from http://www.informationweek.com/cloud/software-as-a-service/ibm-unites-marketing-e-commerce-in-experienceone/d/d-id/1252682

Microsoft alliance. (2015, 04 23). Retrieved from http://www.avanade.com/en-us/about-avanade/working-with-us/accenture-microsoft-alliance

Microsoft Social Listening for CRM. (2015, 04 23). Retrieved from http://download.microsoft.com/download/5/6/6/5668A146-286A-46D9-AA77-E88A8C20E278/eBook%20_Microsoft_Social_Listening_for_CRM.pdf

RStudio. (2015, 07 16). *Take control of your R code*. Retrieved from RStudio: https://www.rstudio.com/products/rstudio/

Social Media Analytics Software as a Service. (2015, 04 23). Retrieved from http://www-03.ibm.com/software/products/en/social-media-analytics-saas

Appendixes

Appendix I: Planning

	W#	Dates			Subject	Deliverable
Part-time 4 days	10	2-Mar	-	8-Mar	Getting started / training	
	11	9-Mar	-	15-Mar	Orientation and reading, proposal corrections	
	12	16-Mar	-	22-Mar	Orientation and reading, proposal corrections	
	13	23-Mar	-	29-Mar	Research proposal finalization	Research proposal
	14	30-Mar	-	5-Apr	Start literature study	
	15	6-Apr	-	12-Apr	Literature study, finish thesis outline	Thesis outline (chapters)
	16	13-Apr	-	19-Apr	Write introduction	
	17	20-Apr	-	26-Apr	Write literature review & list related work	
	18	27-Apr	-	3-May	Write literature review & list related work	Draft version #1
	19	4-May	-	10-May	Indonesia research trip from 1 to 16 May	
	20	11-May	-	17-May		
	21	18-May	-	24-May	Process feedback, chapter 2	
	22	25-May	-	31-May	Finish background information	Chapter 2
	23	1-Jun	-	7-Jun	Chapter 3 & 4	Chapter 3
	24	8-Jun	-	14-Jun	Process feedback, chapter 3 & 4	
	25	15-Jun	-	21-Jun	Chapter 4, 5	

Fulltime 5 days	26	22-Jun	-	28-Jun	Rstudio, process, data structure	
	27	29-Jun	-	5-Jul	Process feedback, chapter 5 & 6	Draft #2
	28	6-Jul	-	12-Jul	Chapter 5 & 6	Framework draft complete → Theo
	29	13-Jul	-	19-Jul	Chapter 4, 5	Chapter 4
	30	20-Jul	-	26-Jul	Feedback, Chapter 5	
	31	27-Jul	-	2-Aug	Chapter 5 & 6	Pre final draft incl. chapter 5 & 6
	32	3-Aug	-	9-Aug	Discussion and conclusion, process feedback, presentation university planning	Final draft version + abstract
	33	10-Aug	-	16-Aug	Presentation Avanade, final corrections, working on final	Final version thesis
	34	17-Aug	-	23-Aug		
	35	24-Aug	-	30-Aug	Presentation university, chapter 7	Avanade recommendations