Bachelor's Thesis

Identifying Future product needs by Clustering Companies

by

Markus Petrykowski

Potsdam, Juni 2015

Supervisor

Prof. Dr. Christoph Meinel

Internet-Technologies and Systems Group

Disclaimer

I certify that the material contained in this dissertation is my own work and does not contain significant portions of unreferenced or unacknowledged material. I also warrant that the above statement applies to the implementation of the project and all associated documentation.

Hiermit versichere ich, dass diese Arbeit selbständig verfasst wurde und dass keine anderen Quellen und Hilfsmittel als die angegebenen benutzt wurden. Diese Aussage trifft auch für alle Implementierungen und Dokumentationen im Rahmen dieses Projektes zu.

Potsdam, July 9, 2015

(Markus Petrykowski)

Kurzfassung

Da Unternehmen gerade in sozialen Netzwerken immer aktiver werden, ist es besonders interessant geworden diese Auftritte zu analysieren und Bedarfsanalysen der Unternehmen damit vorzunehmen. Diese Bachelorarbeit beschreibt zu Beginn eine Herangehensweise zur Gruppierung von Unternehmen. Die dabei entstandenen so genannten Cluster bestehen aus eng in Verbindung stehenden Unternehmen. Diese entstandenen Gruppen werden anschließend in Hinblick auf ihre Bedarfsentwicklung evaluiert. Daraufhin kann mit Hilfe der ausgesuchten Cluster der Bedarf von Unternehmen verhergesehen werden. Die Auswertung des Clusterings zeigt signifikante Korrelationen zwischen der Bedarfsentwicklung von Unternehmen und ihrer Merkmale auf.

Abstract

As companies are becoming more and more active in social networks, it has become interesting to evaluate these social media interactions and extracting demands from them. This thesis describes an approach that first groups companies. So called clusters consist of interconnected businesses. Second it evaluates the formed clusters regarding to their need development. Third it takes the chosen cluster and performs demand predictions of companies. The evaluation of the grouping shows significant correlations for the demand development of companies.

Contents

1 Introduction

Nowadays, as economy has passed boundaries and not only people but also companies are connected throughout the world, it has become impossible to keep track of everything. Companies interact with each other in lots of different ways like being competitors, exchanging employees, using the same infrastructure and more. Some of these influences may create similar struggles or needs for these businesses. Due to the growing presence of businesses online, especially on social media platforms can be used to analyze companies behaviour.

1.1 General Importance of Social Media

Social Media Networks have increased in importance for companies. They use it for creating a closer relation to their customers, for *hiring* new employees, to take care of their contacts, to *advertise* their offers or to *look for new products*. Online business networks like LinkedIn and Xing have a userbase of over 300 million [1] and 9.2 million [2] members.

So far new technologies, like the approach presented by Berger and Hennig[1], enable us to extract product relevant posts, which express a demand, from social media networks for certain products. Using this information a sales representative can actively engage with a new customer. This new form of selling products as a company also provides a lot more opportunities. One of this opportunities will be developed within this thesis.

To be able to sell products using social media, potential new customers have to claim a need in a social network. As this strategy is not widely spread yet, not all of the companies that have a demand do also claim it in a social network.

This thesis addresses this problem and presents an approach that is based on the main-thesis that *similar or strongly related businesses develop similar product needs*. The used way to prove this hypothesis is to cluster companies and evaluate how far companies within one cluster develop the same needs.

[1] https://www.linkedin.com/about-us
[2] https://corporate.xing.com/english/company/

By developing a tool for clustering and visualizing the formed groups it is possible to consider different clustering algorithms and companies' characteristics for the purpose of getting the most accurate result for performing predictions.

After having grouped the businesses successfully this thesis develops a strategy to identify future claims to solve the above problem.

2 Background

To identify companies with a similar demand, it is crucial to understand how companies develop product needs. This chapter will shortly describe Porters Theory [2] of economic clusters and some of his conclusions. Furthermore it is going to explain a subpart of Webster and Wind's model [3] of organizational buying behaviour. They describe environmental influences to which companies are exposed to.

Another important work that is necessary to prove the main thesis, that strongly related businesses develop similar product needs, is the lead extraction from social networks. This approach helps to create a dataset of raised company demands over a time-period. Having this information makes it possible to detect raised needs within a cluster over time.

2.1 Economic Clusters

An economic cluster is a group of companies that are strongly related to each other. This relations could exist through the same industry, a similar company size, the same products or other indicators.

According to Michael E. Porter [2] "Clusters are geographic concentrations of interconnected companies and institutions in a particular field"

These clusters comprise different companies of an industry, including suppliers of specialized inputs such as components, machinery and services, and providers of specialized infrastructure. A cluster contains linkages and complemetaries that are most important to competition.

A vital part of a cluster is an existing competitive attitude. It can survive only if belonging companies try to exceed each other. The quality with which companies compete in a perticular location is influenced by the quality of the local business environment. High quality goods can not be produced without good suppliers or an established transportation infrastructure.

This leads to the other important part of a cluster which is the cooperation. Com-

panies can learn from each other and build on an existing infrastructure of suppliers and providers for goods and services which belong to the cluster as well.

Porter emphasizes the importance of a companie's location for its success, even in times of global markets and faster transportation.

Companies within a cluster are closely related. They depend on each other and are highly influenced by the cluster. As the cluster changes, companies change too. If companies are influenced by the cluster, which is nothing else than companies that are related through their industry and location, than they will also develop together regarding their product needs. This supports our initial assumption that strongly related companies develop similar demands.

2.2 Organizational Buying Behavior

Webster and Wind [3] described a general model to explain organizational buying behavior.

The model addresses the influence factors that may raise new needs as well as the decision process within the company and the actual transaction. The influence factors are mostly relevant here. Following 6 types of environmental influences are mentioned by them:

- Economic (unemployment,economic growth)
- Political (public subsidies)
- Physical (goegraphic, climate, ecological)
- Technological (internet infrastructure)
- Legal (law restrictions)
- Cultural (Diverse working attitudes)

These influences are exerted through several institutions like suppliers, customers, competitors, governments, trade unions and political parties. They have their impact in four different ways.

First of all they define the availability of goods and services. Especially physi-

cal, technological and economic influences affect this impact. For example solar power plants are better situated in areas that provide a lot of sunlight like a desert.

Second they define general business conditions as the rate of economic growth, the level of national income, interest rates, and umemployment. Economic and political forces are the most dominant influences here. Businesses that need many employees are better situated in regions with higher unemployment and educated people.

Third, environmental factors define values and norms of interorganizational and interpersonal relationships between most of the market's participants like buyers, sellers, competitors and governments. Values and norms may be specified by law. But most important are cultural, social, legal and political forces.

Finally, information flow into buying organizations are influenced by environmental forces too. Most vitally to mention here is the flow of marketing communications from potential suppliers, through the mass media and through other personal and impersonal channel ÷ A variety of physical, technological, economic, and cultural factors are showing their effect here.

These influences are important to find measurements that group companies with similar circumstances. Ignoring them would lead to false results that do not represent companies that are exposed to the same influences. Only companies dealing with the same challenges would develop similar demands.

The challenge concerning the different influences is to find good measurements for each of them. Cultural, legal, physical, and political influences are especially tough to find. One attempt to cover those is to use a company's local information. A place can be defined through the country and therefore unites the political influence by the country and city, as well as the geographic conditions and the cultural attitudes of the people living there. The other two left influences can be described more easily by several publicly available indicators like the gross domestic product or the Human Development Index. The data used in this Thesis will mainly cover a company's location and its own economic values.

2.3 Generating Leads from social networks

Berger and Hennig's approach of converting social media posts to leads [1] helps to get a measurement of raised needs in companies.

They extract posts from social media, classify them with a two-stage classifier that sorts the posts by demand and tags certain products based on an already established knowledgebase created for the products.

Having the information of needs in companies makes it possible to address only companies that want to buy certain products.

Their two-stage classification not only makes it possible to analyse a general demand-evolvement for companies, but furthermore special products, which allows the evaluation of the thesis to be even more meaningful.

2.4 Clustering Algorithms

To accomplish the task of finding relationships between two or more companies, for example by grouping them, several algorithms are known. This part shortly describes and compares some of the major strategies to find the most convenient in order to cluster companies.

2.4.1 Clustering characteristics

Existing algorithms can be characterized by the following properties: [4]

- *Exclusive or nonexclusive*. An exclusive classification applies an entity to exactly one cluster, whereas a nonexclusive approach can assign multiple clusters for one entity.
- *Intrinsic and extrinsic clustering*. Intrinsic clustering only uses the calculated proximity matrix for asigning clusters. An extrinsic strategy would additionaly use previously taged values that may already provide some kind of clustering. This strategy is used to find different characteristics that are distinct for the different taged groups.

- *Hierarchical and paritional*. Only exclusive and intrinsic algorithms are subdivided in this two categories. A hierarchical algorithm is a sequence of partitions. It produces multiple clusterings, one per sequence, going from one cluster (contains all entities) to as many clusters as entities exist (one cluster per entity), which is the top-down approach called divisive. The bottom-up version works the opposite direction and is called agglomerative. The number of clusters does not have to be known for the algorithm but in return one has to select the most appropriate division produced by this algorithm. As against a partitional attempt consists of only one single partition. An partitional approach needs to know the number of clusters at the beginning. Then it chooses, more or less randomly, the cluster centres and applies the other entities. Thus a hierarchical classification is a special sequence of partitional classifications.

In lots of cases clustering algorithms are combined to get better results. The combination may allow to recognize outliers and reduce their impact on defining wrong clusters, or to determine a better approximation to the number of clusters. An example could be to first perform a hierarchical clustering to determine a good count of clusters, and afterwards to perform a partitional clustering in order to get improve the result.

2.4.2 Used Clustering Algorithm

Some clustering approaches need to know the number of clusters. Of course one could estimate a number of clusters by considering the number of industries as well as the number of different locations for each industry of a company, but this would still be an approximation to the number, which by the way would get invalid by adding more companies. Hierarchical algorithms have the advantage that they do not need to know the number of used clusters beforehand. This leads to the problem to figure out which of the multiple generated partitions should be used. So it is necessary to have a measurement for partitions to find out which one works best.

Furthermore the used clustering algorithms has to be *exclusive* and *intrinsic*. It would not be on purpose to find characteristics on predefined groups but rather

to define groups of companies. An exclusive approach would provide the information to which cluster a company belongs and that is what we are looking for.

The aim to explore and furthermore predict the need evolvement could be achieved by grouping strongly connected companies. Companies that belong to one cluster should ideally have the same demands. To match the main thesis its important to find correlations between closeness of companies and their needs. Especially its important that a cluster evolves exactly one same need, according to the assumption we make that each company only raises one need. This requirement makes it possible to allow predictions on a cluster's demand evolvement.

Therefore the approach will be to group companies in that way that each cluster has on major product. In this thesis a bottom-up agglomerative hierarchical algorithm will be used as well as the partitional kMeans algorithm in comparison. Both algorithms are *exclusive* and *intrinsic*.

As the hierarchical algorithm produces different possible clusters a way to determine the best clusters is necessary. One clustercombination has to fullfill the following characteristics:

- All the clusters have a strong increase of exactly one demand each
- A cluster contains only companies that do not have the maximum possible proximity [3]

For the number of clusters to pass to the partitional clustering we use different values to test what works best.

[3] For detailed information on the proximity calculation see section 4.3

3 Related Work

This chapter introduces two papers that also described an approach to create clusters of companies and shortly explains their intention and strategy. Further more the key parts of each paper are going to be highlighted and connected to the main-thesis that strongly connected companies develop the the same demands.

3.1 Statistical Approach for grouping companies

Chen, Gnanadesikan and Kettenring [5] already described in 1974 an approach to group companies in their paper "Statistical methods for grouping corporations". Their general objecitve was to "detect, describe and distinguish relatively homogeneous groups of companies"

In their paper they compared a classification of companies by the use of a knowledgebase to a computed cluster analysis. As proximity measures they used fourteen self chosen normalized economic statistics like dividends per share, number of employees in proportion to net plant or the correlation of net sales to net plant as a mix of operational economic and financial variables, to mention only some of them.

They analyzed companies from 5 different industries and were able to insert most of the companies belonging to one industry in the same cluster, by only considering their economic measurements. As a consequence companies that belong to the same industry mostly act similar regarding to their economic statistics. This conclusion confirms the main-thesis insofar as businesses of the same industry may act in a similar way.

3.2 Economic Cluster Analysis

In their paper "Homogenous groups and the testing of economic hypothesis" Elton and Gruber [6] explore cluster analysis for the disaggregation of economic data into meaningful groups. Their main objective was to show the importance

of grouping companies and describe ways in order to test financial hypotheses. One key aspect was to get better results by decomposing measurements to avoid certain characteristics that may be represented by multiple variables. For example a company's stock price and its income per year. Both give information about a company's success and value. But above this they also provide slighty different information. So what they want is to break this variables down that the company's value does not count more than other characteristics.

After explaining how to decompose variables into a new set of varibles without any interferences by the means of a principal components analysis they discussed criterias for grouping like group compactness.

The most important part of this paper for this thesis is the prevention of possible interferences that can exist between some grouping criteria. Because analyzing financial values can give us information about a firm's possible buying behaviour its important to choose the criterias correctly in order to weight the values right.

4 Company Clustering Algorithm

As already discussed in section 2.4.2 this thesis will compare the results between a bottom-up agglomerative hierarchical clustering and the partitional kMeans clustering algorithm, which are both *exclusive* and *intrinsic*. The clustering and scoring is based on the assumption that *every company raises only one demand*. This is necessary to simplify the demand prediction and the cluster rating. It does not distort the results respectively to the correlation of companies' demands and their characteristics.

This section describes where the used data comes from and how it was processed and prepared for clustering. Further more this section will outline how the proximity between companies were calculated and how the resulting clusters were scored.

4.1 Data

To determine clusters of companies, its necessary to have a data-set that contains the relevant information for a company, and has to be big enough to get meaningful results.

4.1.1 Datasources

The entry point for the data was LinkedIn. To further enrich the data-set we also used Crunchbase.

LinkedIn is a social business network with over 300 million user,[4] with people from all over the world. Apart from user-profiles it also contains company-profiles with properties like year of foundation, industry or number of employees. The information are maintained by the companies itself.

Crunchbase is an open database containing startup-activity and company information.[5] Company-datasets contain information like employees, competitors,

[4] https://www.linkedin.com/about-us, 28th of June 2015
[5] https://info.crunchbase.com/about/ 28th of June 2015

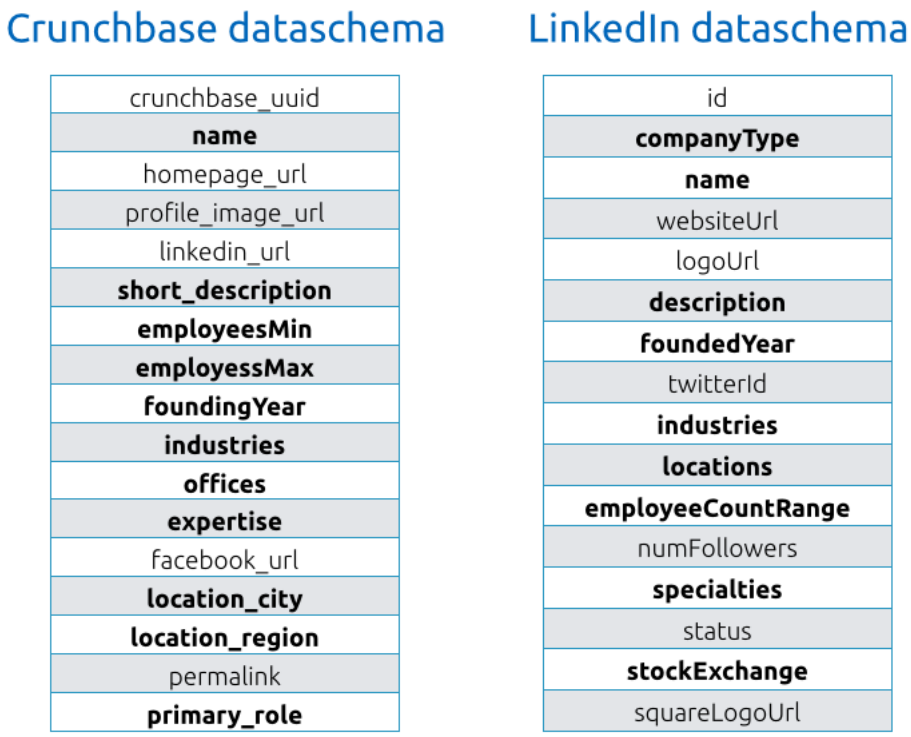

Figure 1: Comparison of dataschemas

industry and basic information as well. Like in wikipedia information can be maintained by everyone, which leads to frequently updated information on the one hand, and could lead to wrong information on the other hand.

Figure 1 shows a subset of attributes of companies that are provided by each source. [6] The characteristics that represent information to conclude a companies demand are printed bold. [7] Both datasets provide similar information but with a different structure. For example the number of employees. Crunchbase provides two attributes one for the mininum value and one for the maximum value as integers, whereas LinkedIn delivers a string like "1001-5000" which requires normalization to extract the same information.

[6] More detailed information can be found on http://data.crunchbase.com/v3/docs/organization and https://developer.linkedin.com/docs/fields/company-profile

[7] Regarding to influencing factors and a companies environment in chapter 2 and 3. See also Chapter 4.3

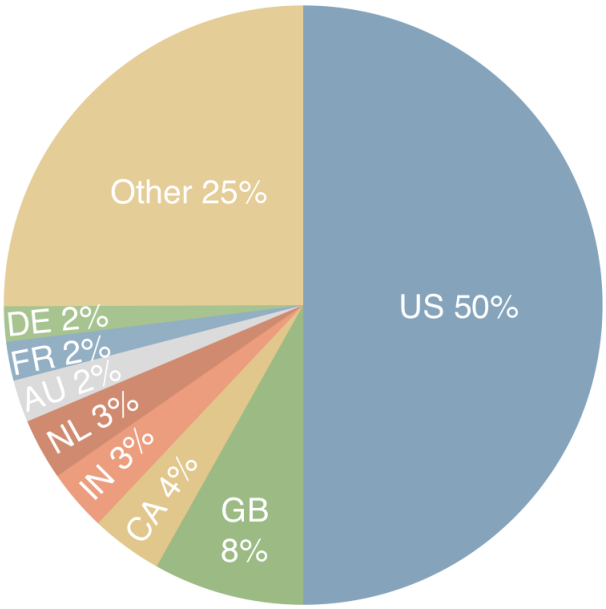

Figure 2: Distribution of companies per country (Total 236,235 companies)

4.1.2 Dataprocessing

Because both sources have different advantages and information and as mentioned in the last section a different structure, it makes sense to combine both datasets into one, that covers all the necessary information needed for clustering, and has one defined dataschema.

The biggest challenge in combining these two datasets is finding the right corresponding company in the respectively other dataset. The used approach was to join to datasets on an equal string match of both company names. If companies have slightly different names in both sets, they will be matched if they have the same website url given. Otherwise a new entry will be created in the resulting dataset. This resulted in a dataset of 236,235 companies. As you can see in Figure 2 the most companies are located in the United States. Since both datasources are US-companies and the main users are therefore from the united states that distribution is not surprising. The overall dataset contains companies from 220 countries.

4.1.3 Used data for clustering

Because this thesis looks for correlation between company characteristics and the company's demands, we can only use companies we also have demand information for. So we compared the existing posts from the NoiseToOppotunity[1] database with the crawled companies from LinkedIn and Crunchbase. This resulted in a test dataset of 1129 companies.

4.2 Company Features

Features are variables or a combination of variables that can describe certain characteristics of an entity. Using the right features is essential to prove that strongly related businesses develop similar product needs. In this case the features have to describe characteristics that influence a companies buying behaviour.

Regarding to Porter [2] a company's *location* has a high influence on how it acts. Companies will often rather know what happens next to them than at a totally different place. Steps taken by companies right next to each other will have a higher impact on how each of them reacts to particular circumstances, especially purchases made by one of the companies may lead to an economic advantage. Other companies are then forced to close this gap by doing similar purchases.

Of course the location is important but has less impact if the companies next to each other do not compete somehow. Referring to Porter [2] companies of the same *industry* are often shaped in clusters at one location. They are using the same infrastructure and increasing the clusters know how.

So the first two features that cause the highest influence from one company to another are a company's location and its industry.

An increasing number of employees within a company leads to a higher complexity. Also bigger companies have other needs and higher expenses than smaller ones have. Therefore companies of similar size are more related to each other than to smaller sized companies. This leads us to the third feature, a company's size measured by its *number of employees*.

According to Webster and Wind [3] companies are exposed to 6 different influences. These influences are already covered by the selected features. For example by selecting a company's location the legal, economic and political influences which are the strongest ones are considered.

Other characteristics mentioned in chapter 4.1 and displayed in figure 1 could also be used as features. But as the selected 3 features cover all the aspects discussed during the economic background, we stick to the three features *location*, *industry*, *number of employees*. A comparison of results using different combinations of features could be part of future work. This thesis focuses on finding a correlation between this features and the company's demand demand-evolvement.

4.3 Calculate Proximity

The calculation is performed on the set described in section 4.1.3, that contains 1129 documents. This would result in a maximum count of realtionships of 636.756, where each company has a relationship to each other, but not itself. Thats the result of following function where n equals the number of companies.

$$\text{relations} = \frac{n*(n-1)}{2}$$

For the proximity we take all of the features and weight them according to their influence. Regarding to the conclusions in chapter 4.2 we assume that location and industry have a high weight whereas the company size does not have that much impact on a company's buying behaviour.

The calculation takes two companies and compares the defined features. We check for an industry match I_m, whether both companies have at least one common industry, for a location match L_m, whether both companies have at least one office in the same city or country, and finally for and employees count match E_m, whether both companies have the same employees range. If a match exists the corresponding value is one otherwise it is zero.

$$\text{proximity} = 1 - (I_w * \frac{I_m}{I_w+L_w+E_w} + L_w * \frac{L_m}{I_w+L_w+E_w} + E_w * \frac{E_m}{I_w+L_w+E_w})$$

The proximity formula, with E_w for the company's size weight, L_w for the company's location weight and I_m for the company's industry weight return a value between zero and one. The smaller the value the closer 2 companies are.

4.4 Cluster scoring

Annahme: Jede Company entwickelt nur ein demand / What clustering algorithm used? What exactly? cluster company level -> Partition

To be able to evaluate which feature weight works best and which partitions of the ones formed by the clusterings it is necessary to have some characteristics to compare.

The first and most important one is the function score(X) that calculates how good a cluster is according to the fact whether it strongly develops only a demand for one specific product.

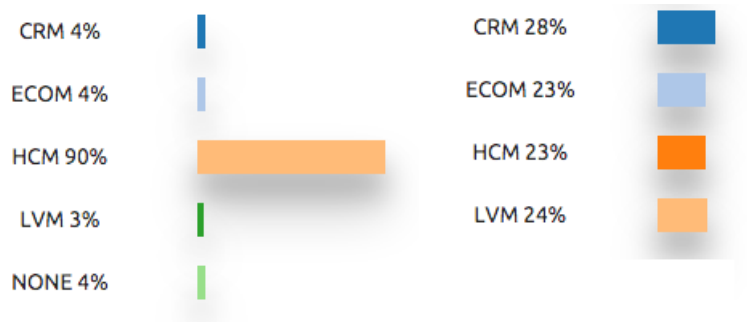

Figure 3: A good and a bad example of a cluster's demand development

Figure 3 shows a good demand development on the left and a bad one on the rigth. The used dataset of demand-posts covers 5 different products: CRM, ECOM, HCM and LVM. The percentage value says how many of the companies within the according cluster raised a demand of the corresponding product. The left distribution reveals that the remaining companies in the cluster may also be interested in a HCM product.

A good distribution can be recognized easily as human beings. But we do not want to have a closer look at each product distribution to decide wether a par-

tition is good or bad. In order to process each cluster combination using we developed a function to distinguish between bad and good product distributions.

$$\text{Let } X \subseteq \mathbb{Q}, score(X) = \frac{maxVal((max2(X)-avg2(X)),1)}{(max(X)-avg(X))}$$

$$max(X) = \text{maximum value of set X}$$
$$avg(X) = \text{average value of set X}$$
$$max2(X) = max(X \backslash \{y|y = max(X)\})$$
$$avg2(X) = avg(X \backslash \{y|y = max(X)\})$$
$$\text{Let } k, l \in \mathbb{Q}, maxVal(k,l) = \text{returns the bigger value}$$

Figure 4: The scoring function

The function score(X) returns a value that describes the ratio between the difference of the highest value and the average to the difference of the second highest value and the average without the highest value. The closer the value is to 0 the better the cluster.

If the highest value strongly differs from all the others than it has a high difference to the average value. If the highest value is by far the highest than the difference between the second highest value and the average without the higest value will small. To prevent a wrong result the denominator has to be at least 1 because otherwise the whole value could be 0 even if the highest value does not have a high difference to the average.

To clarify the formula we are going to calculate for the two examples in figure 3 The Left distribution:

$$\text{score([4,4,90,3,4])} = \frac{maxVal((4-3{,}75),1)}{(90-21)} = \frac{1}{69} = 0.0144$$

The right distribution:

$$\text{score([28,23,23,24,6])} = \frac{maxVal((24-19),1)}{(28-20{,}8)} = \frac{5}{7{,}2} = 0.6944$$

The left distribution has as expected a better value than the right distribution. To evaluate a whole cluster combination the rating for each cluster gets calculated. All the ratings will then be averaged according to the clusters size. So a good rating within a small cluster will not have as much impact as a good rating in a bigger sized cluster.

Other measurments to value a cluster combination are the total number of companies within the clusters or the highest average of a products demand. The more companies covered, the more efficient the demand predictions are. Also the higher the average covering is for the products, the more actively are the companies spreading demands within a cluster. An average covering takes only the highest coverage from each of the existing clusters and averages them. According to our exmaple in figure 3 we would calculate the average of 90 and 28.

4.5 Downsides

As we simplified our approach by assuming that every company raises one product demand only, we have also have to consider what happens if companies also raise more than one demand, which is the more realistic case.

The assumption allowed us to use an exclusive clustering algorithm. So if companies raise more demands they have to be in more than one cluster according to their demand. This allows us to still be able to use the same scoring function but it would support multiple demands per company.

This would require a fuzzy algorithm. C-Means could be used for this task. It is an extension of the standard kMeans algorithm and was first introduced by Bezdek [7]. This implementation could be part of the future work.

5 Prototype

For this thesis a tool was developed that easily allows to cluster companies by different feature weights, select the most appropriate cluster from all the cluster combinations emerged by a hierarchical clustering and visualize this cluster and show the important information and statistics for each group.

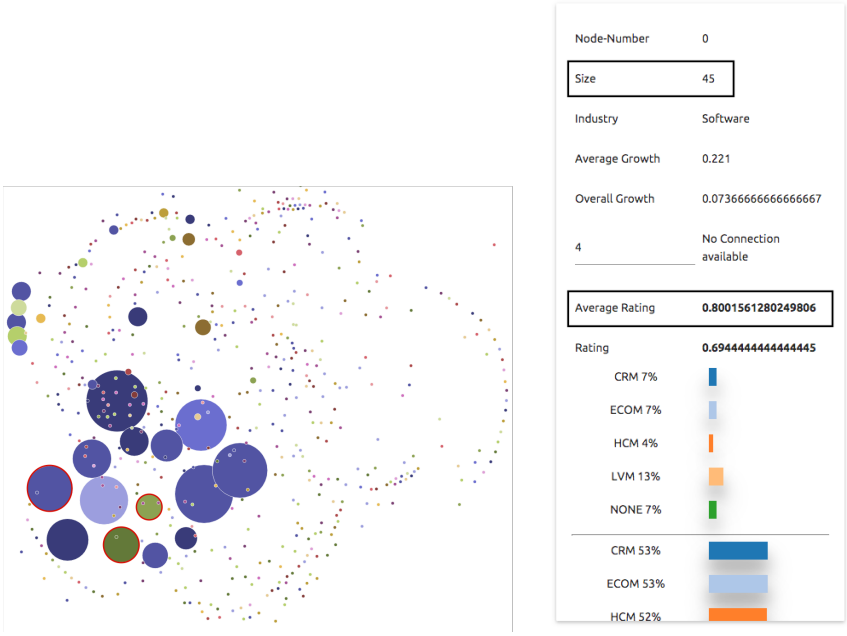

Figure 5: A cluster visualization

Figure 5 shows an example visualization of clusters produced by the tool. Each circle represents an own cluster. The bigger each circle's radius, the more companies are contained within this cluster. The distance between each cluster shows their distance towards each other. Circles with a red border include companies that raised a demand in a timeframe that can be adjusted. The different coloring of the clusters illustrates the dominant industry. A click on a cluster gives an overview of the most important values, as what needs were raised, what is the clusters or the overall demand distribution for all products or what is the accordingly cluster score.

6 Evaluation

This section evaluates the result of different clusterings with different weights of the used features. Multiple clusterings with different feature weights were processed to get a deeper understanding of the correlation between the features and the demand development.

We differentiate between test that were made on all clusters from one result and from clusters without their outliers. The resulting partitions contained mostly more than one clusters, where only clusters with more than one company so calles one company cluster were considered. One company clusters are bad for deducing a correlation of company closeness and need development because a company within a one company cluster can not have influence on any other companies within this cluster.

Companies covered by one partition are not evenly distributed among the clusters, so we have clusters that contain more companies than others. This could lead to a heavily irregular distribution. In the following we will first have a look at the whole partition and second at the partition without its outliers.

The strategy for the tests were quite similar. The features that were used were once weighted alone to see their impact. Then different weighting combinations were performed in order to get better results.

For the partitional clustering we used partitions with the cluster size 100, 600 and 800 as a partitional clustering needs to know the number of clusters beforehand. As the hierarchical clustering does not have to know the number of clusters to generate we took the best performing partition for each iteration.

The used quality measures are:

- *Avg. Rating*: The average rating which is the result of the weighted average score [8] of one partition
- *Clusters*: The number of clusters that contain more than one company and occure in the partition

[8] Using the scoring function described in section 4.4

- *Weight*: The used weight for the iteration with I=Industry, S=Size, L=Location
- *Biggest Cluster*: Number of companies in the biggest cluster
- *Covered Companies*: Number of companies covered by the clusters of this partition
- *Level*: The selected tree-depth which performs best (hierarchical clustering only)
- *Tree depth*: Number of possible Partitions the clustering build (hierarchical clustering only)
- *k-value*: The number of clusters that should be created by the partitional clustering (partitional clustering only)

6.1 Using total set of cluster

This subsection evaluates the findigs where the outliers within each cluster were still considered.

Table 1: Different feature weights and their result for hierarchical clustering with variance

Nr.	Weight[9]	Clusters	Level	Highest Avg	Biggest Cluster	Covered companies	Tree depth	Avg. Rating
1	0_L,0_I,1_S	8	45	34%	732	911	777	1.0500
2	0_L,1_I,0_S	4	119	22%	233	357	352	1.0560
3	1_L,0_I,0_S	3	148	27%	136	211	284	1.0356
4	1_L,1_I,1_S	4	61	27%	46	64	107	0.7597
5	2_L,2_I,1_S	7	51	7%	43	94	94	0.8480
6	2_L,8_I,1_S	6	60	7%	43	99	103	0.8556

Table 1 shows the results for the test with the bottom-up agglomorative hierarchical clustering.

[9]Weight matches Size , Industry , Location

The results for rows 1-3, where each feature was weighted alone, were quite similar. Their average rating is between 1.03 and 1.05 whereas the tree depth differs a lot. A hierarchical clustering always produces a tree as its outcome structure. This tree represents the clustering. Each node within this tree represents an own cluster that contains every child element of this node. So the deeper the tree the more one company clusters it contains. According to the results for clusterings with single weights only, the clustering resulted from the location weighted tree has a much higher depth than any other result set. Therefore the location itself as a result is useless. The other two single weighted trees are useless for predictions as well. Their clusters show a balanced demand development for each of the products. So they are not suited for demand predictions by themselves.

This result is not astonishing as economic processes are very complex and can not be described by a single measurement. Thats why we had a look at the influence of all three of the features. According to Porter [2] and Webster and Wind [3] industry and location do both have a high influence. So we had a look at combination where this features where wheighted more than the size feature.

The unexpected result was that the evenly ditributed weight to all of the three features leads the best outcome. It has a similar depth like table rows number 5 and 6 but much better average rating which is the most interesting measurement to look for. Even if it only covers two thirds of the other two result sets it has a higher prediction potential. But although it provides the best average rating it is still not good enough to perform any useful predictions. On the one hand it covers only around a twentieth of the overall companies set and on the other hand the cluster score of 0.75 is still to high for reliable demand forecasts.

Table 2 shows the results for the test with the partitional kMeans clustering.

The average rating for each of the partitions is higher than 1. So none of iterations are really well for predictions. This algorithm gives nicely distributed clusters, as you can see in row 1. With 460 companies it covers nearly the half of all possible companies. The biggest cluster contains only 33 companies, so the other 86 cluster contain around 5 businesses each. Another interesting ob-

[11] Weight matches Industry , Nr. of employees
[11] k-value describes the number of cluster that should be generated

Table 2: Different feature weights and their result for kMeans clustering with variance

Nr.	Weight[10]	Clusters	k-value[11]	Highest Avg	Biggest Cluster	Covered companies	Avg. rating
1	1_I,1_S	87	800	33%	33	460	1.0026
2	1_I,1_S	35	100	24%	162	873	1.0972
3	0_I,1_S	7	800	23%	275	933	1.1303
4	0_I,1_S	7	100	23%	275	933	1.1303
5	1_I,0_S	1	800	18%	1129	1129	1.1076
6	1_I,0_S	46	600	41%	99	550	1.0565
7	1_I,0_S	30	100	32%	191	777	1.1036

servation are rows 3 and 4. They are exactly the same and weighted with the size only. As the number of different sizes is smaller than the clusters to form, no more than 7 clusters exist. As in both cases the k-value is higher than the possible number of clusters both iterations evaluate to the same result.

Overall the results of the partitional kMeans considering the outliers does not provide a well base for identifying future product needs.

6.2 Using set of cluster without outlier

This subsection evaluates the findigs where the outliers within each cluster were not considered.

Table **??** shows the results for the test with the bottom-up agglomerative hierarchical clustering without the outliers. This means the biggest cluster was not considered in the calculations.

In comparison to Table **??** the average rating became better for nearly all partitions. By ignoring the biggest cluster the coverage for all partitions sank. But nevertheless the last row looks promising. Indeed it only covers 36 companies

[12]Weight matches Size , Industry , Location

Table 3: Different feature weights and their result for *hierarchical clustering* without variance

Nr.	Weight[12]	Clusters	Level	Highest Avg	Biggest Cluster	Covered companies	Tree depth	Avg. rating
1	0_L,0_I,1_S	4	28	27%	10	26	777	0.8461
2	0_L,1_I,0_S	12	44	23%	68	157	352	0.9618
3	1_L,0_I,0_S	4	62	22%	137	360	284	1.0070
4	1_L,1_I,1_S	7	48	24%	28	82	107	0.9291
5	2_L,2_I,1_S	11	6	29%	10	41	94	0.7804
6	2_L,8_I,1_S	4	13	42%	11	36	103	0.6296

but it has the best average rating. This result matches the theories of Porter[2] and Webster and Wind [3] that the industry and location of a company have a high impact on its demand development.

Table 4: Different feature weights and their result for *kMeans clustering* without variance

Nr.	Weight[13]	Clusters	k-value[14]	Highest Avg	Biggest Cluster	Covered companies	Avg. rating
1	1_I,1_S	86	800	34%	26	427	0.9852
2	1_I,1_S	34	100	25%	125	711	1.1161
3	0_I,1_S	6	800	23%	191	658	1.0803
4	0_I,1_S	6	100	23%	191	658	1.0803
5	1_I,0_S	-	800	-%	-	-	-
6	1_I,0_S	45	600	42%	55	451	1.0201
7	1_I,0_S	29	100	31%	116	586	1.0995

Also table 4 shows a better result than table 2. It also shows that the best result was generated when considering not only one feature but both that were used.

[14]Weight matches Industry , Nr. of employees
[14]k-value describes the number of cluster that should be generated

6.3 Comparing hierarchical clustering and kMeans

The two different clustering algorithms have both in common that their result improved when ignoring the outliers. For the used features the hierarchical cluster performed better. But its difficult to compare both algorithms based on the fact that both used another featureset.

But it shows that 2 features are not enough and that the result becomes even better with more features. But what also results from the tests, that the more features used the less companies were covered by by the clusters. The challenge here is to find as less features as possible with as much companies covered as possible.

6.4 Correlation of company closeness and need development

The tests prove an existing correlation between the companies' characteristics and their need development. Especially table 3 in row 6 confirms the main-thesis. The table also shows a development of the average rating. In row 3 were only the industries were considered the average rating was above 1. But with increasing the number of features and weighting them accordingly to Porter[2] and Webster and Wind[3] the average rating becomes better.

6.5 Improving the result

The result proves an existing correlation between the used features and the demand development. But still the outcome is not sufficient enough. Obviously the used characteristics do not cover up the whole complex structure that describes the buying behaviour of companies.

One approach to solve this problem is to use more characteristics. Appropriate ones could be the monthly income, a more detailed company description or other metrics for a company's economic situation.

Another approach to improve the result is to use a fuzzy clustering. As companies are not restricted to develop one demand only it has to be possible to assign

one company to different clusters. So it would be interesting to see how the average rating would develop when using a fuzzy clustering. We assume that the result will also cover up more companies.

7 Future Work

As this thesis is more a proof of concept than a final solution the main aspect of the future work will be to improve the clustering by refering to more metrics as already mentioned in section 6.5. Therefore it could be useful to also mention additional datasources like the Compustat database which provides financial, statistical and market information on companies throughout the world.

Since the datasets used for this thesis were entirely small in comparison to the information thats available on LinkedIn or Crunchbase, the clustering tool should be able to handle millions of datasets. Useful predictions can be generated from a lot of data only.

Indeed this thesis discussed and developed a model for identifying future product needs, but this approach still needs to be implemented. So a tool for predicting future demands from companies will be necessary.

Current tools enable sales people to have an overview over their social networks and keep track of companies that show interest in certain products. The tool developed with this thesis would provide a way to multiply this demand. This would strongly increase the salesmen's efficiency.

Apart from predicting future demands this tool could also be used for marketing evaluations. As it provides an overview of companies and visualizes the raised demands per cluster, evaluations like a spread rate or number of companies reached can be generated. One could see how successful certain marketing campaigns were and even better who exactly they may have reached. This would also provide information about a products user group. All based on real demand tracking in social networks.

8 Conclusion

This thesis presented an approach to group companies, evaluate different clusters regarding to their need development and predict demands of companies.

In addition to this thesis we developed a tool that allowed us to perform the tests for the evaluation. We compared the two clustering algorithms bottom-up agglomorative clustering and the partitional kMeans clustering. Both perfomed well and the results developed as expected. However, as we used only two features for the partitional clustering the hierarchical algortithm produced the better result.

The data we used was sufficient enough to fullfill the needs for the used features. For further tests it would be helpful to enrich the companies with even more data. Especially economic ones like the income per year. This would allow to get even better clusters.

This thesis could not give the final answer to the problem, it is further more a proof of concept that the primary objective to identify and predict future product needs is possible. It reveals existing correlations between a company's demand development and its different characteristics.

The opportunities are promising. Knowing companie's demands beforehand would create a big advantage to the ones using it.

References

[1] Philipp Berger Patrick Hennig: *Noise-to-opportunity conversion for social media posts*.

[2] Michael E. Porter: *Clusters and the new economics of competition*. Harvard Business Review, pages 77–90, November - December 1998.

[3] Frederick E. Webster JR. Yoram Wind: *A general model for understanding organizational buying behavior*. Journal of Marketing, 36:12–19, April 1972.

[4] Anil K. Jain and Richard C. Dubes: *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, New Yersey, 1st edition, 1988.

[5] Hwei Ju Chen R. Gnanadesikan J. R. Kettenring: *Statistical methods for grouping corporations*. The Indian Journal of Statistics, Series B, pages 1–28, February 1974.

[6] Edwin J. Elton Martin J. Gruber: *Homogeneous groups and the testing of economic hypotheses*. Journal of Financial and Quantitative Analysis, 4:581–602, January 1970.

[7] James C. Bezdek: *Fcm: The fuzzy c-means clustering algorithm*. Computer Geosciences, 10:191–203.