HPI Hasso Plattner Institut
IT Systems Engineering
Universität Potsdam

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 6, 2015

(Markus Petrykowski)

Kurzfassung

bla

Abstract

bla

Contents

1 Introduction

1.1 Outline

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 social media platforms can be used to analyze companies behaviour.

So far webtechnologies 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 thesis is going to present an approach that is based on the assumption that similar or strongly related businesses develop similar product needs. We are going to prove this assumption, explore existing correlations and develop a strategy to identify future claims.

2 Background

It is crucial to understand how companies develop product needs. This chapter will shortly describe Porters Theory of economic clusters and some of his conclusions. Furthermore it is going to explain a subpart of Webster and Wind's model 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

According to Michael E. Porter [1] "Clusters are geographic concentrations of interconnected companies and institutions in a particular field"

These clusters include 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. Companies 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.

2.2 Organizational Buying Behavior

Webster and Wind [2] 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 physical, technological and economic influences affect this impact.

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.

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 f̈low 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.

2.3 Generating Leads from social networks

Berger and Hennig's approach of converting social media posts to leads [3] 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 most known ones to find the most convenient in order to proof the main thesis.

Different Algorithms may belong to some of the following categories: [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.

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.

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, to mention only some of them.

They analyzed companies from 5 different industries and were able to assign most of the companies to the right 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.

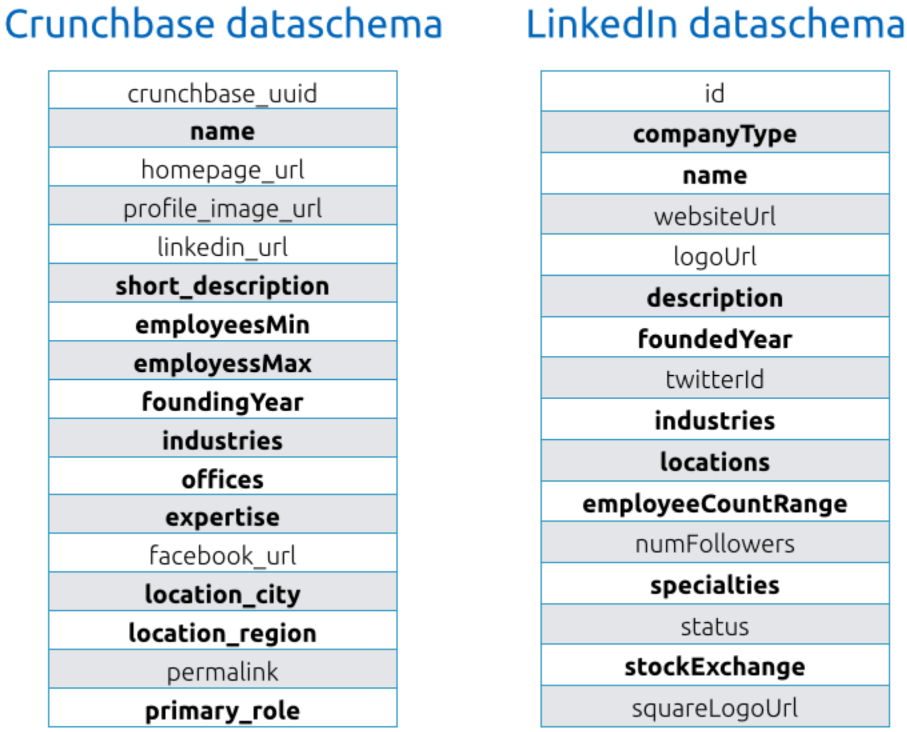
After explaining how to decompose variables into a new set of varibles without any interferences by the means of a principal components analysis they discussed criterias for grouping like group compactness.

The key aspect for the main 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 behaiviour its important to choose the criterias correctly in order to weight the values correctly.

4 Company Clustering Algorithm

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.



Crunchbase dataschema
crunchbase_uuid
name
homepage_url
profile_image_url
linkedin_url
short_description
employeesMin
employessMax
foundingYear
industries
offices
expertise
facebook_url
location_city
location_region
permalink
primary_role

LinkedIn dataschema
id
companyType
name
websiteUrl
logoUrl
description
foundedYear
twitterId
industries
locations
employeeCountRange
numFollowers
specialties
status
stockExchange
squareLogoUrl

Figure 1: Comparison of dataschemas

4.1.1 Datasources

To ensure a good quality the data-sets were extracted from two different sources, LinkedIn and Crunchbase.

LinkedIn is a social business network with over 300 million user,[1] 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.[2] Company-datasets contain information like employees, competitors, industry and basic information as well. Like the wikipedia information can be maintained by everyone, which could lead to frequently updated information on the one hand, and to wrong information on the other hand.

Figure 1 shows a subset of attributes of companies that are provided by each source. [3] The characteristics that represent information to conclude a companies demand are printed bold. [4] Both datasets provide similar information but with a different structure. For example the number of employees. Crunchbase provides 2 attributes one for the mininum value and one for the maximum value as integers whereas LinkedIn delivers a string like “1001-5000” which requires further processing to extract the same information.

4.2 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 problem 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 a 100% match of both companynames. If companies have slightly different names in both sets, they will be matched if they have the same

[1] https://www.linkedin.com/about-us, 28th of June 2015

[2] https://info.crunchbase.com/about/ 28th of June 2015

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

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

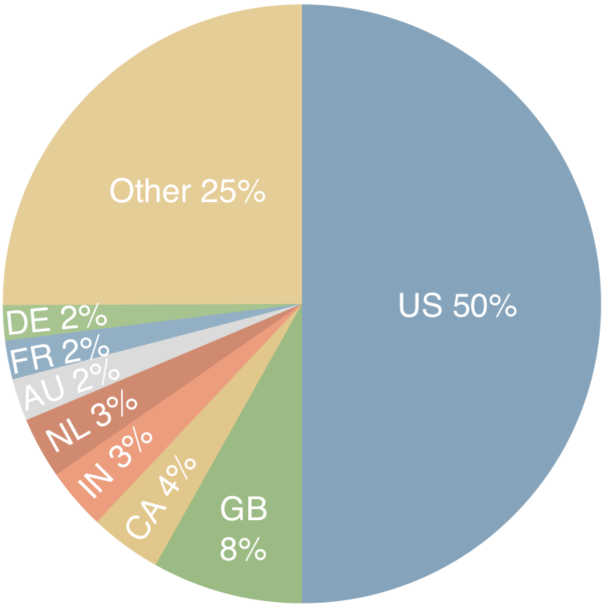


Figure 2: Distribution of companies per country (Total 236,235 companies)

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. The dataset contains companies from 220 countries.

[Maybe add industries as well.]

4.3 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 [1] 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 [1] 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 [2] 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 could also be used as features. But as the selected 3 features cover all the aspects discussed during the economic background, there is no need for more features at the moment. A comparison of results using different combinations of features could be part of future work. This thesis focuses on finding a correlation between the closeness of companies and their demand-evolvement.

4.4 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, 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. But this neither solves the problem of getting good cluster because one would still have to figure out which of the multiple generated clustering shemas should be used. So it is necessary to have a measurement of a cluster schema 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.

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. This requirement makes to possible to allow predictions on a cluster′s demand evolvement.

Therefore the approach will be to calculate the proximity between each of the companies. This makes it possible to look for existing correlations and form clusters. A agglomerative hierarchical algorithm will be used to perform this.

As the 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

4.5 Calculate Proximity

The calculation is performed on a randomly sampled subset of our total set of companies to save time and keep the amount of produced realations as small as possible. The subset contains 1192 documents. This would result in a maximum count of realtionships of 709.836, 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 4.3 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.

4.6 Cluster scoring

To be able to evaluate which feature wheight is the best and which cluster combination of the set that emerges from the hierarchical clustering 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 one demand.

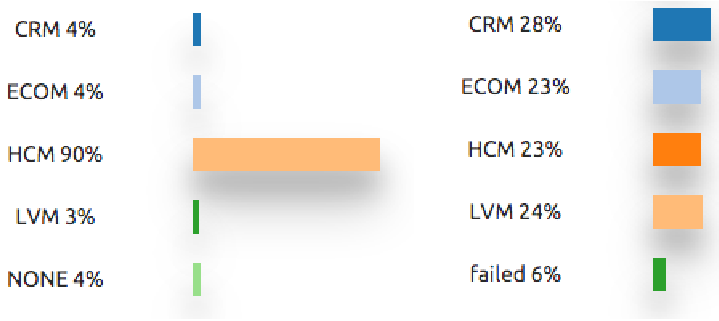


Figure 3: A good and a bad example of a cluster's demand development

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 allows to conclude that the remaining companies in the cluster may also be interested in a HCM product.

$$\text{Let } X \subseteq \mathbb{Q}$$

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

$$\text{max(X) = maximum value of set X}$$

$$\text{avg(X) = average value of set X}$$

$$\text{max2(X)} = max(X\backslash\{y|y = max(X)\})$$

$$\text{avg2(X)} = avg(X\backslash\{y|y = max(X)\})$$

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.

5 Evaluation

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

Table 1: Different feature weights and their result

Nr.	Clusters	Avg Rating	Level	High Avg	Big Cluster	Weight Size Industry Location	Tree depth
1	8	1.0500	45	34%	732 911	0,0,1	777
2	4	1.0560	119	22%	233 357	0,1,0	352
3	3	1.0356	148	27%	136 211	1,0,0	284
4	4	0.7597	61	27%	46 64	1,1,1	107
5	7	0.8480	51	7%	43 94	2,2,1	94
6	6	0.8556	60	7%	43 99	2,8,1	103

5.1 Correlation of company closeness and need development

Table 1 shows the resulting measurements regarding to the different feature weights. To see what impact each feature has without the influence of the other ones, an own cluster combination for each feature was created [5].

The results for this combination 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 [6] it contains. 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. According to the results for clusterings with single weights only,

[5] See tablerows 1-3 in **??**

[6] A one company cluster is a cluster with only one element

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 [1] and Webster and Wind [2] 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.

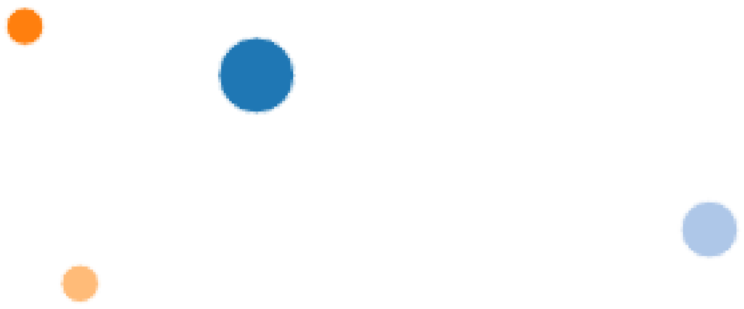


Figure 5: Visualized cluster for the cluster combination with the best rating

Figure 5 shows the 4 clusters belonging to table row number 4 and the distance between each of the cluster. The one company clusters are ignored in this visualization because they do not provide any further value. Even if this cluster combination provides the best average rating it is still not good engough to perform any useful predictions. On the one hand it covers only around a twentieth of the overall companie set and on the other hand the cluster score of 0.75 is still to high for reliable demand forecasts.

5.2 How could the result could be improved?

To test what influence each feature has we will have a look at the

Therefore the first approach weights the location and industry 0.45 each and the size 0.1 times.

Show coverage of companies within cluster that match a product related demand. At the beginning , midtime and at the end. Evaluate the goodness of the feature distribution by this development.

Showing that the clustering makes sense. Explain why its ok to ignore clusters that are only a company by itself. -> To show the spreading over time within a cluster to show the well choseness of a cluster cant be shown with one company cluster

6 Conclusion

7 Future Work

References

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

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

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

[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.