Bachelor's Thesis

Future product needs of Companies by intelligent clustering

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, June 24, 2015

(Markus Petrykowski)

Kurzfassung

bla

Abstract

bla

Contents

1 Introduction

1.1 Outline

- Thesis: Strongly connected companies develop similar needs
- Clustering with emphasis on features that describe companies with similar needs
- Number of clusters are unknown
- clustering could be either done by classification using some kind of knowledgebase or clustering using clustering algorithms

  as numbers of clusters are unknown classification does not make sense because a knwoledgebase can not be build

Nowadays, as economy has passed borders and not only people but also companies are connected throughout the world, it has become impossible to keep watch 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 at these businesses. So it may be possible that companies being affected by those forces may also make related decisions.

This thesis explores the need development of strongly connected firms and the existing correlations. It further more tries to use this correlation to perform predictions of needs in the near future.

2 Background

To make assertions on when companies raise needs it is crucial to understand what drives companies to want special products and what influences them. Therefore this chapter will shortly describe Porters Theory of economic clusters and some of his conclusions he made. Further more it is going to explain a subpart of Webster and Wind's model of organizational buying behaviour which describes environmental influences which companies are exposed to that may force them to take action.

Another important work that is necessary to prove the main thesis is the lead extraction from social networks. This approach helps to create a dataset of raised company needs 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 "Clusters are geographic concentrations of interconnected companies and institutions in a particular field" [1]

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 can 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.

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:

- Physical (goegraphic, climate, ecological)
- Technological
- Economic
- Political
- Legal
- Cultural

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 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.

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 need-evolvement for companies, but further more special products, which allows the proving test of the thesis to be even more meaningful.]

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

Hwei-ju Chen, Gnanadesikan and Kettenring [4] 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 that 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 [5] 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 to do that in order to test mainly 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 assign the values the correct weight.

4 Company Clustering Algorithm

4.1 Data

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

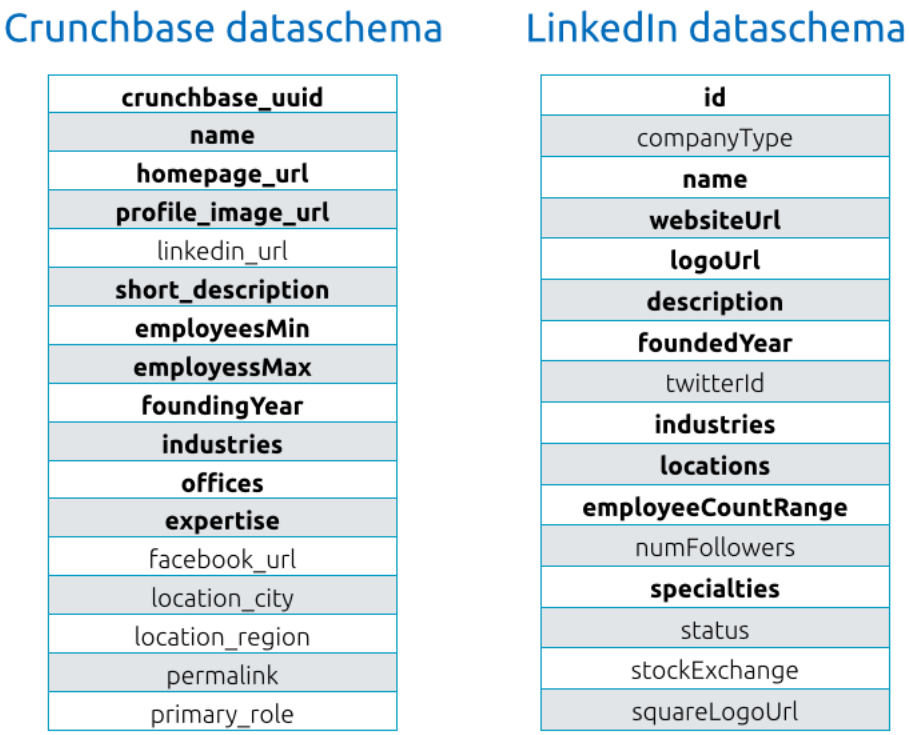

Figure 1: Comparison of dataschemas

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] including people from all over the world. Apart from user-profiles it also contains company-profiles with data-properties like year of foundation, industry or number of employees. The informations are mostly 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. The 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 relevant characteristics are printed bold. 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 for extracting 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 website url given. Otherwise a new entry will be created in the resulting dataset. This resulted in a dataset of 236235 companies. As you can see in figure 2 the

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

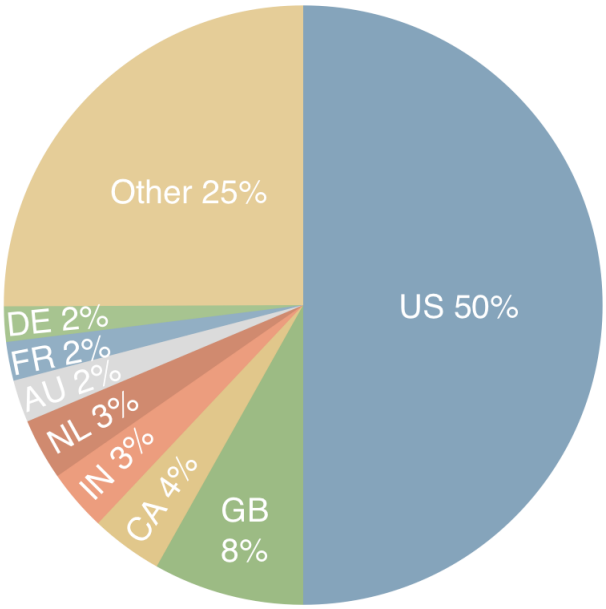

Figure 2: Distribution of companies to countries

most companies are located in the United States. The dataset contains companies from 220 countries.

[Maybe add industries as well.]

4.3 Clustering Features

Features are variables or a combination of variables that can describe certain characteristics of an entity. Using the right features is essential for the mainthesis to get proved. 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. Therefore 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 may be forced to close this gap by doing similar

purchases.

Of course the location itself is important but has a less impact itself if the companies next to each other do not compete somehow. According again to Porter [1] companies of the same *industry* are often shaped in clusters at one location. There they are using the same infrastructure and increasing to 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*.

These features will be evaluated later on after building the first clustering.

4.4 Calculate Proximity

4.5 Used Clustering Algorithm

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: [6]

- *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 algorith 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 a better result. 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.

Hierarchical algorithms for Clustering 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 could get invalid by adding more companies. Again it is important to mention here that these algorithms are exclusive and the only information that exists for one company would be to which cluster it belongs, not how close they belong to each other.

So our approach has to be, to get an evaluation of all companies according to their closeness to each other. This would also create some kind of clusters, but we would still have the information of closeness between each company. An appropriate datastructure to store this information could be a graph.

The aim to explore and further more predict the need evolvement could be achieved better by having the additional information of closeness between each

company. To match the main thesis its important to find correlations between closeness of companies and their needs.

- Hierarchical or non hierarchical alg? - Shortly describe possible clustering algs
  - Number of resulting clusters not known
  - Hierarchical cluster doees not need to know the cluster count bottom-up or top-down no exact clustering chosen
- Combining existing Cluster-definitions which both have a huge impact onto similar need evolvement. (Location / Industry)

5 Evaluation

5.1 Correlation of company closeness and need development

5.2 Treats for raising certain needs

6 Conclusion

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] 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.

[5] 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.

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