Reading Between the Lines of Qualitative Data – How to Detect Hidden Structure Based on Codes

Alexander Keller[1], Hans Achatz[2]

[1] University of Passau, Chair of Information Systems (Information and IT-Service-Management), Passau, Germany
alexander.keller@uni-passau.de
[2] University of Passau, Teaching Unit Information Systems, Passau, Germany
achatz@uni-passau.de

Abstract. While qualitative research is experiencing broad acceptance in the information systems discipline, growing volumes of heterogeneous data pose challenges to manual qualitative analysis. We introduce an unsupervised machine learning approach based on graph partitioning to detect hidden information and structure in qualitative data samples. With the clustering technique, we map coded data to a graph and formulate a partitioning problem which is solved by integer linear programming. As a result, clusters of information sources are identified based on similarities given in the coded data. We demonstrate the approaches' ability to detect hidden information in coded qualitative data by application on coded interview transcripts. With the approach, we draw on a technique from the operations research discipline and expand the repertoire of approaches being used to analyze qualitative data in the context of information systems.

Keywords: qualitative data mining, clustering, graph partitioning, unsupervised learning, integer linear programming.

1 Introduction

Since 2005 one can see a significant growth in qualitative research publications in information system (IS) journals which shows the growing relevance of qualitative research in the IS discipline [1]. Besides studies that are applying qualitative methods to investigate research topics (e.g. [2–4]), a lot of work is focusing on guidelines regarding how qualitative research should be conducted within the IS discipline (e.g. [5–7]). Additionally, qualitative research is often seen as a minor discipline because of findings which are said to be accompanied with biases (e.g. subjectivity) regarding the common quality standards [8].

Qualitative research in IS aims to gather a deep understanding of behavioral and technical issues regarding the role of information technology (IT) and often supplements quantitative studies in a mixed method approach [9]. Referring to Romano Jr. et al. (2003) with an increasing amount of available qualitative data it becomes

14th International Conference on Wirtschaftsinformatik,
February 24-27, 2019, Siegen, Germany

necessary to develop approaches to analyze growing volumes [10]. In the context of qualitative data, especially clustering mechanisms have been discussed in literature (e.g. [11]) but still remain underused [12]. In the case of clustering being used, researchers typically perform manual techniques to detect groups based on codes (i.e. marker for a relevant information given in qualitative data) in qualitative data which requires extensive resource commitments [13]. Due to this there is a growing field of scholars who are applying machine-learning techniques on qualitative data. Within this field, automatic coding techniques based on natural language processing and graph theory (e.g. [14, 15]) as well as clustering approaches (e.g. [16]) were applied on qualitative data. However, the clustering techniques are mostly used to cluster participants with similar profiles of codes instead of clustering codes based on their relationship to each other. Following this research stream, the paper investigates the research question how hidden structure in qualitative data sets can be detected automatically based on code similarities? Answering this question, we develop an unsupervised learning technique called *CodeClust* based on the concept of graph partitioning. We draw on a technique which we developed to solve clustering problems in the operations research (OR) domain and adopt it to mine and analyze coded qualitative data. With performing the approach, results are generated that grant additional and hidden insights regarding the structure and affiliation of qualitative data constructs. With this, we intend to expand the repertoire of approaches that are used in qualitative IS research and contribute to future studies by providing a new approach to analyze qualitative data samples.

The remainder of paper is structured as follows: First, we give a brief overview of data collection, coding techniques and existing quality criteria in qualitative research. This leads to the third chapter in which we outline fundamentals and related work in the domain and present *CodeClust*. Besides describing data preparation and essential features of the approach, we illustrate how to use the technique on coded data from expert interviews. The paper concludes in a discussion of the approach and gives an outlook towards future research directions.

2 Data Collection, Coding Techniques and Essential Quality Criteria in Qualitative Research

There exist different techniques of data collection to gather empirical material for the purpose of data analysis. Non-numeric information is usually collected directly in form of interviews or indirectly from secondary sources like text documents [17]. This results in textual information sources like interview transcripts and case descriptions which serve as a foundation for further analysis. For a detailed view of different methods, data collection techniques, modes of analysis and quality criteria used in qualitative IS research, see Keller (2017) [18].
After the data has been collected, the information sources have to be structured for further analysis in a next step. Grounded theory is a widely used approach and serves

as a foundation to structure textual information based on coding techniques [19]. The aim of the coding is to highlight segments of textual data that contain relevant information in regard to the underlying research question [20]. Therefore, codes represent information given in the data and serve as a generalization of information. In this field, Gläser and Laudel (2013) provide a detailed view on coding techniques and variations in objectives related to coding [19].

In most cases, the coding process itself is done manually. To ensure the quality of the results, the data sources are coded by different coders and their outputs are merged and evaluated for similarities. Within this coding process, software tools (e.g., *NVivo, Atlas.Ti, HyperResearch, MaxQDA*) are used by the coders and support in attaching codes to relevant text segments. Additionally, the tools provide functionalities in form of quantifications and visualizations to process the results.

In terms of quality criteria in qualitative research, validity and reliability are differently characterized than in the quantitative domain [21].
Although validity aims to ensure the quality and information value, Flick (2014) argues that flexibility is one major strength of qualitative approaches [22]. Therefore, communicative validation is a common approach to ensure credibility and accuracy in qualitative research [21, 23].
Reliability stands for the robustness of findings and the consistency of an approach. However, in qualitative research identical findings do not always represent reliable results (e.g. identical responses in interviews may point to prepared answers) [24]. In order to draw consistent conclusions in qualitative studies, the specific context and the data collection process itself must be described in very detail. With this, one can ensure traceability from an intersubjective point of view [25, 26].

In addition, credibility describes the internal validity of qualitative research which can be ensured by triangulation and negative case analysis [21]. As the researcher functions as a central part in the research process (i.e. data collection, coding, analysis, interpretation) some degree of bias is induced because of his personal perception which may lead to the problem of subjectivity [27, 28]. Therefore, qualitative research processes should be supplemented with standardized techniques to ensure a non-subjective evaluation and interpretation of qualitative data. Sarker et al. (2013) support this point of view by mentioning the strong "[…] need for clarity in the logic underlying data analysis […]" [29] in qualitative IS research.

3 Clustering Coded Qualitative Data

In qualitative research, scholars are interested in identifying behavioral or structural pattern to understand the studied phenomenon in the research context. For this purpose, the available heterogeneous qualitative data must be gathered and processed to structure the observed information. As mentioned in the last section, one central aspect of this systematic structuring process is the assignment of codes to specific information. While

most of this processing is done manually in qualitative research, the machine learning (also referred as statistical learning) domain provides some techniques to predict and structure categorical information in an automated manner on the one side and offers approaches to scale up the coding process on the other side [30].

3.1 Foundations and Related Work

When predicting categories (i.e. discrete variables with nominal or ordinal scale) two different approaches can be distinguished due to the underlying structure of available data: (i) supervised learning describes techniques for classification of information based on labeled data. (ii) unsupervised learning techniques process unlabeled data to detect structure within the given information. Labeled data refers to a sample that has been tagged with labels and hence includes information about categories. Therefore, in supervised learning labeled training samples are used to train different classifiers (e.g. naïve bayes, support vector machines or decision trees) aiming to categorize information according to the training data. In contrast, unsupervised learning aims to detect information in form of clusters (e.g. k-means) or topics (e.g. latent dirichlet allocation) in unlabeled samples without any prior knowledge about categories. Figure 1 shows the difference between the two approaches of machine learning and gives some examples.

Figure 1. Unsupervised vs. Supervised Prediction of Categories.

As we aim to detect hidden information in qualitative data samples, this paper focuses on unsupervised learning in form of clustering. Clusters are defined as sets of objects (e.g. text documents) which are grouped based on similarities. Most prior work in this field has its origin in text mining, where documents are clustered based on the similarity of words to identify certain semantic topics [31, 32]. Besides emphasizing content representation, clustering is also used to identify dominant information in given samples [33].

The foundation of previous research is in general based on partitional clustering methods like K-Means [34] or graph-partitioning approaches [35] where an object

belongs in exactly one cluster and the number of clusters is given in advance. Besides that, hierarchical clustering techniques are used where a set of nested clusters is built by successive merging or splitting [36, 37]. In general, hierarchical clustering strategies fall into two types: (i) the agglomerative or bottom up approach where clusters are merged on every hierarchy level and (ii) the divisive or top down approach in which splits are performed on every hierarchy level. Another approach is called topic modeling which decomposes information sources into topics and links the information sources to the identified topics based on certain probabilities. Latent dirichlet allocation introduced by Blei et al. (2003) is a common technique for topic modeling and is applicable in most textual based clustering scenarios [39, 40].

Most of the previous work has in common that the occurrence, frequency and/or combination of words are taken into consideration when building clusters. However, in qualitative research additional information in form of codes is available which can be considered in the clustering to detect hidden information. With this paper, we propose a clustering technique that grasps the relation of codes to each other.

3.2 Concept

Meeting the needs for an automated clustering technique based on the coded qualitative data, we develop a graph-partitioning approach called *CodeClust*. The presented technique is built upon existing clustering approaches and embraces the numerical structure of coded data which results from the assignment of text segments to codes (e.g. coding process in grounded theory and content analysis). The intention of the approach lies in clustering information sources and codes. As Marton (2013) points out, this is necessary because ”[…] collected data needs to be grouped […] in order to form relevant corpora for comparison” [41]. Such additional information about groups serves as a basis and guidance for detailed interpretation and supports the analysis of subgroups [16, 42].

In our scenario, we assume that these similarities are based on information sources (e.g. interview transcripts) expressing the same ideas and therefore standing in relation to each other. For example, groups of experts can be identified in interviews which are considered to be similar relating to their statements. As *CodeClust* is a partitional clustering each object is assigned to one cluster.

The provided graph-theoretic approach has originally been developed to simplify the complexity of staff scheduling problems in the OR discipline [43]. We adopt our method from this domain and adapt it to analyze qualitative data to cluster affiliated information sources in groups with regard to their coding. Before describing the approach table 1 defines the used mathematical notations.

Table 1. Mathematical Notations and Definition

Notation	Definition
n	Total number of information sources
m	Total number of codes
i	Index of information sources, $i \in \{1, \ldots, n\}$
V	Set of nodes within a graph, with $\lvert V \rvert$ being the total number of nodes
E	Set of edges between nodes of V within a graph
c	Index of codes, $c \in \{1, \ldots, m\}$
f_{ci}	Code-frequency of code c within information source i, $f_{ci} \in \mathbb{N}_0$
b_{ci}	Binary value for a code being coded above average within i

In a first step, the coded qualitative data must be represented as a graph to use the technique. Therefore, the codes c are represented as nodes V in a graph $G = (V, E)$. The total number of nodes $|V|$ is m, representing the total number of codes given in the data set. E represents a set of edges between nodes c and c' of V. An edge between two nodes exists, if both codes represented by the nodes are coded above average within at least one information source. This is modeled with the binary value b_{ci} that takes the following two states:

$$b_{ci} = \begin{cases} 1, if\ f_{ci} > \left\lceil \frac{1}{|V|} \sum_{c'=1}^{|V|} f_{c'i} \right\rceil \\ 0, otherwise \end{cases}$$

These binary values b_{ci} determine the graph, which serves as a basis for the partitioning technique. In addition, each edge of the graph contains a weight $g_{cc'}$ that stands for the number of information sources in which the two codes c and c' are coded above average:

$$g_{cc'} = \sum_{i=1}^{n} b_{ci}\, b_{c'i}$$

The graph partitioning problem is solved with integer linear programming (ILP) to identify two disjunctive sets of nodes $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$. For the ILP two binary auxiliary variables are introduced:

$$z_c = \begin{cases} 1, if\ the\ first\ set\ of\ nodes\ contains\ c \\ 0, otherwise \end{cases}$$

$$y_{cc'} = \begin{cases} 1, if\ an\ edge\ between\ c\ and\ c'\ leads\ from\ one\ subgraph\ to\ the\ other \\ 0, otherwise \end{cases}$$

The partitioning aims to minimize the weights of edges which lead from subgraph G_1 to G_2. To solve this minimization problem, the objective function is formulated under four constraints:

$$\min \sum_{(c,c')\in E} g_{cc'}\, y_{cc'}$$

$$(1) \quad z_c - z_{c'} - y_{cc'} \leq 0 \qquad \forall (c,c') \in E$$

$$(2) \quad z_{c'} - z_c - y_{cc'} \leq 0 \qquad \forall (c,c') \in E$$

$$(3) \quad \sum_{c \in V} z_c \geq \frac{|V|}{4}$$

$$(4) \quad \sum_{c \in V} z_c \leq \frac{3|V|}{4}$$

Constraints 1 and 2 ensure that the weights of edges are considered in the objective function if the particular edge leads from one subgraph into the other. E.g. the first and second constraint forces $y_{cc'}$ to be 1 when c and c' are nodes of different subsets (e.g. $z_c = 1;\ z_{c'} = 0$). Hence the weight of the related edge between c and c' $g_{cc'}$ is considered in the objective function. In the approach, the constraints 3 and 4 are used to prevent the clusters from becoming very small or vice versa very large. With the actual setting, it is ensured that a cluster contains at least the fourth part of the total amount of nodes. These boundaries can be set according to the individual preferences on the minimum cluster size.

In figure 2, an example is given which shows a graph G holding six nodes with different weights on edges between the nodes. After solving the ILP for the given example two subgraphs G_1 and G_2 could be identified where two edges with a weight of 1 are leading from subgraph G_1 to the subgraph G_2 and vice versa. In the example the minimal solution of the objective function results in $g_{2\,3} + g_{1\,5} = 2$.

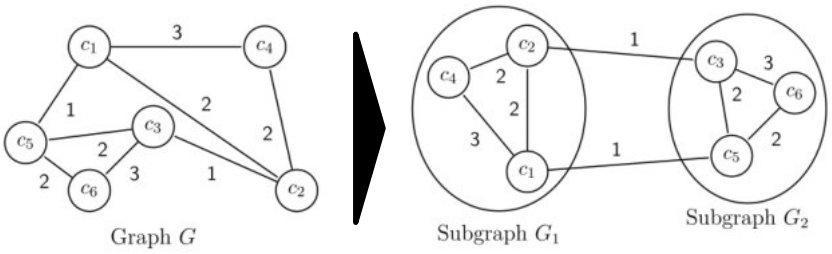

Figure 2. Example of Graph Partitioning

As shown in Achatz (1999) the technique can also be performed multiple times on large datasets to generate more than two subgraphs [43]. As multiple iterations can be performed by using the output from one iteration as input for the next one, large qualitative data samples can be analyzed with the approach in an automated manner. In settings where a code is not mentioned above average, one can either (i) exclude the code from the partitioning, (ii) assign code to smaller partition or (iii) assign code to larger partition. In scenarios where every node must be assigned to a cluster option (ii) or (iii) would be applied. However, in terms of clustering codes, we recommend using the first option. The specific code is excluded from the partitioning. This does not mean that the code should be excluded from further analysis. The information from the general coding can still be valuable for later interpretation.
As in each iteration the sample is only split into two clusters, the appropriate amount of clusters can be determined with existing measures like the elbow curve method [44] or a silhouette analysis [45]. In the first method the percentage of variance explained is used to find the appropriate number of clusters. The second method provides a graphical representation in form of silhouettes and measures how well an object fits to its cluster compared to other clusters. Identifying the correct number of clusters, however, is not a trivial task and there exists a multitude of cluster quality indicators (e.g. [46, 47]).

Regarding the behavior of the approach in higher orders of magnitude, there does not exist a high sensitivity to the number of observations. However, larger sample sizes will increase the runtime to solve the formulated partitioning problem with ILP.

3.3 Interpretation

The results from the clustering technique help to answer the question which set of information sources mention which specific codes. Hence, groups can be identified that include information sources based on the similarity of their coding. The clustering serves as an additional insight and makes it possible to identify structure in form of dependencies between information sources and codes. By just considering the codes, which are mentioned above average within the information source, the method uses the code-frequency for clustering. Therefore, it is assumed that information sources that mention the same codes above average are related to each other. This relation results from the fact that the information sources contain similar content regarding the research objective. E.g. if the information sources are interviews one can build groups of experts based on the similarity of their content-related statements. As a result, it can be identified which group of experts emphasizes which codes. Combining this with additional information about experts one can generate specific insights regarding the meaning and affiliation of identified codes.

4 Application to Field Data

The approach has been used on data from interviews with entrepreneurship experts to identify success indicators for IT-start-ups [48]. Besides entrepreneurs and investors,

business angels were chosen as experts to generate a holistic view on the topic. In total eleven interview transcripts were coded on the basis of the methodology proposed from Steigleder (2008) [49]. The theory orientated content coding technique results in 22 codes. Each code represents a separate success indicator in the domain. The data set in table 2 represents the results of the coding process of textual data from the interviews.

Table 2. Data Set of Coded Qualitative Data from Interviews

			Interview i										
	c	**Success Indicator**	1	2	3	4	5	6	7	8	9	10	11
Code c	1	Sales competence and marketing power	2	3	5	2	0	2	1	2	3	0	1
	2	Perseverance	0	0	0	1	2	2	1	1	0	0	0
	3	Value orientated thinking	4	1	0	0	2	2	0	0	0	0	0
	4	Entrepreneurship and professional experience	2	1	1	2	2	2	2	2	0	0	2
	5	Industry specific competence	2	0	3	3	1	0	1	3	1	0	0
	6	ICT competence	1	2	0	2	1	1	1	4	2	0	0
	7	Conversion capability and speed	3	0	2	1	0	0	0	0	1	0	2
	8	Staff management skills	2	1	0	1	1	0	0	0	1	0	3
	9	Scalability and market ability	2	2	0	1	0	2	3	1	3	0	0
	10	Proof of feasibility and verification	0	0	1	0	1	4	0	1	0	2	0
	11	Business model flexibility and independence	2	3	0	0	2	0	0	2	1	1	2
	12	Team composition	1	1	1	2	2	0	0	2	2	3	0
	13	Industry specific financing	1	0	1	7	0	1	1	6	0	2	0
	14	Accelerator or incubator program	0	0	7	0	0	0	0	0	1	2	0
	15	R&D cooperations	0	0	0	3	1	4	2	2	0	0	1
	16	Seed-customer and technology partner	1	5	0	0	5	1	1	2	1	0	1
	17	Handling of market conditions	0	7	0	1	4	1	4	5	1	0	0
	18	Political and regulatory business environment	0	1	0	3	5	0	0	2	3	0	1
	19	Industry specific norms and requirements	0	4	0	3	3	0	1	2	3	0	0
	20	Customer orientated problem solving	0	1	3	1	3	3	0	2	2	2	1
	21	Feedback driven product development	0	2	1	0	3	0	0	5	3	1	1
	22	Prototype orientated product development	1	0	0	2	7	0	1	1	1	2	1

The values given in the matrix stand for the code-frequency of a particular code in an information source (f_{ci}). E.g. code *c=1* is coded two times in interview *i=1* which results in f_{11} equals 2. The variable b_{ci} is a binary value which equals 1 if a code c is coded above average in an information source i. If this is not the case, b_{ci} is set to 0.

Before performing the approach, the graph must be generated based on the values of b_{ci}. An edge between two nodes (i.e. codes) c and c' exists, if for at least one information source i the multiplication of b_{ic} and $b_{ic'}$ equals 1. The number of information sources i for which this is true represents the weight $g_{cc'}$ of a particular edge between the nodes c and c'. As represented in figure 3 on the next page the three codes (c=10; c=15; c=20) are mentioned above average in the sixth interview (i=6) which results in $b_{10\,6} = b_{15\,6} = b_{20\,6} = 1$. Besides that, information sources 6 and 10 mention the codes 10 and 20 above average. Therefore, the weight of $g_{20\,10}$ equals 2.

In figure 3 the input and the output of the approach is visualized. The matrix on the left side shows an unsorted data set of binary values b_{ci} based on the qualitative data given in table 2. This represents the graph which serves as an input for the presented approach. The matrix on the right side in figure 3 represents the data set sorted into two groups based on the graph partitioning. The first group contains the sources 1, 3, 6 and 10 as well as the codes 20, 12, 10, 14 and 3. The second group is determined by the rest of sources and codes, while code 2 is an artefact that cannot be assigned to any group

because it is not mentioned above average in any information source. One can see that no source of the second group mentions a code of the first group above average, which means that each code of the first group is uniquely assigned to it.

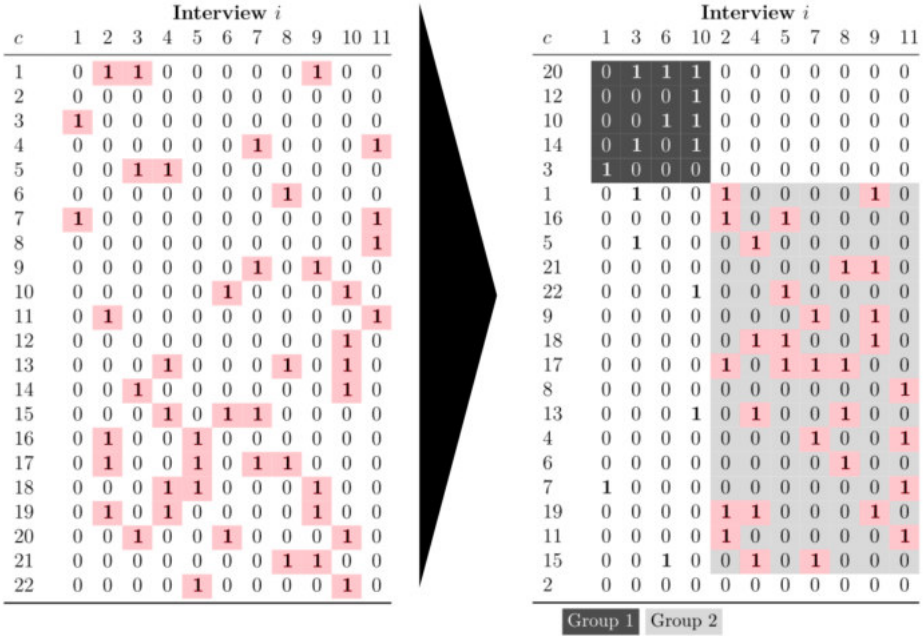

Figure 3. Results Generated by the Clustering Technique

Using the approach, we identified that in the first group experts are clustered which have a strong relation to phases at the beginning of the start-up life cycle. In this group is one business angel supporting start-ups in the seed and start-up phase and the rest of experts are entrepreneurs standing at the beginning of their ventures. This observation is accompanied with the fact that the codes of the first group represent success indicators which are important when founding and developing the firm. E.g. *team composition* (c=12) and *accelerator or incubator program* (c=14) are aspects entrepreneurs must deal with at the very start of a venture.

In contrast, experts in the second cluster handle start-up firms in the growth and expansion phase (e.g. investors, advanced founders). Looking at the codes it becomes evident that the success indicators *handling of market conditions* (c=17) and *political and regulatory business environment* (c=18) are related to advanced phases in the start-up lifecycle as well. Additionally, no expert of group two is mentioning success indicators of group one above average. In contrast, some experts of group one do mention codes of group two. This shows that no expert in group two (i.e. expert related to advanced start-up life cycle phases) emphasizes the success indicators of early start-up phases because they seem to be less relevant. Experts in group one (i.e. experts

related to early start-up life cycle phases) consider the future by focusing not only on success indicators for the beginning of a venture but also on success indicators of advanced phases of the start-up lifecycle.
While choosing start-up stakeholders as experts we did not focus on their affiliation regarding different start-up lifecycle phases. With performing the clustering, we are able to detect hidden structure in form of subgroups which generates valuable findings concerning the affiliation of success indicators to lifecycle phases.

5 Limitations and Future Research

As shown in the example the approach provides valuable results and assists in the detection of hidden structure in qualitative data samples. However, the technique is subject to some limitations.
First, it is depending on the underlying data. This means, that the results generated can only be as good as the quality of the gathered qualitative data. Therefore, the clustering should only be performed on top of a robust data collection and coding process.
Second, although it is transparent how clusters are generated, the findings might still be biased by a subjectivity in data pre-processing steps and by a wrong interpretation of clustering results. Therefore, it is important to mention that the clustering results offer guidance for interpretation and subgroup analysis but should not be applied as strict decision guidelines.
Third, like other techniques analyzing qualitative data, the approach provides an indication for the affiliation of codes with regard to the underlying research subject. In contrast to "[…] statistical methods which require representative data, cluster analysis does not find generalizable characteristics" [12]. Hence, it is suitable for qualitative research which aims to understand complex phenomena regarding a specific topic of interest.
Fourth, although the used code-frequency is a suitable measure to create the graph of codes it can make sense to include other aspects like the context a code is related to or actual speech such as laughter. Hence, the frequency measure could be enhanced or replaced with other subject related aspects.
Fifth, like in other clustering techniques, it has to be considered that there may be settings where the clustering does not result in interpretable results. This is indicated by a high value for the objective function related to the sum of all weights of edges. In this case, there exist outliers which indicate an unstable clustering. Therefore, we recommend combining multiple manual and automated clustering techniques to confirm accuracy of generated cluster solutions.

Regarding the limitations, future research should investigate how the approach can be combined with already existing qualitative data analysis techniques. Especially the possibilities for complementation should be considered to validate findings on the one side and supplement them on the other. In addition, the behavior of the presented approach could be studied in more detail. E.g. the clustering technique can be tested with a variation of the underlying data. With this the behavior of the partitioning

procedure can be evaluated with focus on the robustness of the clustering. Furthermore, we intentionally applied a small data example to introduce the approach in a simple and understanding manner. Although the clustering delivers stable results for larger data samples in its original discipline [43], further research should investigate the behavior with different dimensions of volume. A benchmark of the approach against other techniques can be based on agreement measures like the Cohen's kappa [50, 51]. In a next step, the technique could also be compared to traditional connectivity-based clustering techniques like hierarchical clustering. As distance metric the proposed $g_{cc'}$ will be considered to identify similarity of codes.
In addition, the approach is not only suitable in qualitative research scenarios but should also be considered in the domain of business analytics. Although in these scenarios the data modeling might be different, the problem formulation remains the same.

6 Conclusion

We introduced a machine learning approach to detect hidden information and structure in qualitative data samples. Therefore, we used a clustering technique based on a graph theory to group information sources based on the similarity of codes. With the approach, we map the coded data to a graph and formulate a graph partitioning problem which is solved with ILP. As a result, the technique separates the graph into different clusters based on the similarities given in the coded data. The technique is designed to be used on textual qualitative data which results from any kind of qualitative coding process. Hence, the approach can be performed independently of the underlying methodology and does not replace any existing procedures but complements them.

Until now interpretation of qualitative data mostly relies on the system of codes, their assigned textual passages and different frequency measures. Most clustering in this context is performed manually or semi-automated which is contradictory to the essential quality criteria of intersubjective traceability [22, 29]. Hence, with the presented automated approach we present a new data modeling and clustering technique which adds to the existing repertoire of qualitative data analysis methods in IS. Although the data pre-processing (i.e. data collection, transcription and coding process) might still be biased by subjectivity, the approach could increase the reliability of qualitative research approaches regarding interpretation. However, the clustering should guide and support the researcher in context informed interpretation and code relationship analysis but should not be used in form of decision rules to split qualitative data sets.
Referring to Sarker et al. (2012) about 60% of qualitative studies in the IS discipline use coding procedures to analyze empirical data [1]. Because of this, our approach could address and complement many existing and future qualitative studies in the IS domain in terms of applying a mixed methods approach to answer research questions. Especially when dealing with large volumes of empirical data, the potential of the clustering could be exploited.

References

1. Sarker, S., Xiao, X., Tanya, B.: Towards an Anatomy of "Successful" Qualitiative Research Manuscripts in IS: A Critical Review and Some Recommendations. In: Thirty Third International Conference on Information Systems (ICIS 2012). pp. 1–21 (2012).
2. Chua, C.E.H., Yeow, A.Y.K.: Artifacts, Actors, and Interactions in the Cross-project Coordination Practices of Open-source Communities. J. Assoc. Inf. Syst. 11, 838–867 (2010).
3. Davidson, E.J., Chismar, W.G.: The Interaction of Institutionally Triggered and Technology - Triggered Social Structure Change: An Investigation of Computerized Physician Order Entry. MIS Q. 31, 739–758 (2007).
4. Vogelsang, K., Steinhüser, M., Hoppe, U.: A Qualitative Approach to Examine Technology Acceptance. In: International Conference on Information Systems (ICIS 2013). pp. 234–245 (2013).
5. Urquhart, C., Lehmann, H., Myers, M.D.: Putting the "Theory" Back into Grounded Theory: Guidelines for Grounded Theory Studies in Information Systems. Inf. Syst. J. 20, 357–381 (2010).
6. Klein, H.K., Myers, M.D.: A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems. MIS Q. 23, 67–93 (1999).
7. Schultze, U.: A Confessional Account of an Ethnography about Knowledge Work. MIS Q. 24, 3–41 (2000).
8. Conboy, K., Fitzgerald, G., Mathiassen, L.: Qualitative Methods Research in Information Systems: Motivations, Themes, and Contributions. Eur. J. Inf. Syst. 21, 113–118 (2012).
9. Venkatesh, V., Brown, S.A., Bala, H.: Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems. Manag. Inf. Syst. Q. 37, 21–54 (2013).
10. Romano Jr., N.C., Donovan, C., Chen, H., Nunamaker Jr., J.F.: A Methodology for Analyzing Web-Based Qualitative Data. J. Manag. Inf. Syst. 19, 213–246 (2003).
11. Guest, G., McLellan, E.: Distinguishing the Trees from the Forest: Applying Cluster Analysis to Thematic Qualitative Data. Field methods. 15, 186–201 (2003).
12. Macia, L.: Using Clustering as a Tool: Mixed Methods in Qualitative Data Analysis. Qual. Rep. 20, 1083–1094 (2015).
13. Brickey, J., Walczak, S., Burgess, T.: A Comparative Analysis of Persona Clustering Methods A Comparative Analysis of Persona Clustering Methods. (2010).
14. Crowston, K., Liu, X., Allen, E.E.: Machine learning and rule-based automated coding of qualitative data. In: Proceedings of the ASIST Annual Meeting. pp. 1–2 (2010).
15. Tierney, P.: A Qualitative Analysis Framework Using Natural Language Processing and Graph Theory. Int. Rev. Res. Open Distrib. Learn. 13, 173 (2012).
16. Henry, D., Dymnicki, A.B., Mohatt, N., Allen, J., Kelly, J.G.: Clustering Methods with Qualitative Data: a Mixed-Methods Approach for Prevention Research with Small Samples. Prev. Sci. 1007–1016 (2015).
17. Myers, M.D.: Qualitative Research in Information Systems. MIS Q. 21, 241–242 (1997).
18. Keller, A.: How to Gauge the Relevance of Codes in Qualitative Data Analysis ? – A Technique Based on Information Retrieval. In: 13th International Conference on Wirtschaftsinformatik. pp. 1096–1110 (2017).
19. Gläser, J., Laudel, G.: Life With and Without Coding: Two Methods for Early-Stage Data Analysis in Qualitative Research Aiming at Causal Explanations. Forum Qual. Sozialforsch. 14, (2013).

20. Kelle, U.: Theory Building in Qualitative Research and Computer Programs for the Management of Textual Data. Sociol. Res. 2, (1997).
21. Lincoln, Y.Y., Guba, E.G.: Naturalistic Inquiry. Sage, Beverly Hills (1985).
22. Flick, U.: Gütekriterien qualitativer Sozialforschung. In: Baur, N. and Blasius, J. (eds.) Handbuch Methoden der empirischen Sozialforschung. pp. 411–423. Springer, Wiesbaden (2014).
23. Bryman, A.: Research Methods and Organization Studies. Unwin Hyman, London (1989).
24. Flick, U.: An Introduction to Qualitative Research. Sage, Los Angeles (2014).
25. Kruse, J.: Qualitative Interviewforschung: Ein integrativer Ansatz. Beltz Juventa, Weinheim (2014).
26. Steinke, I.: Kriterien qualitativer Forschung. Beltz Juventa, Weinheim (1999).
27. Zahedi, F.M., Van Pelt, W.: Web Documents' Cultural and Femininity Masculinity. J. Manag. Inf. Syst. 23, 87–128 (2006).
28. Wrona, T., Gunnesch, M.: The One Who Sees More is More Right: How Theory Enhances the 'Repertoire to Interpret' in Qualitative Case Study Research. J. Bus. Econ. 1–27 (2015).
29. Sarker, S., Xiao, X., Beaulieu, T.: Guest Editorial: Qualitative Studies in Information Systems: A Critical Review and Some Guiding Principles. MIS Q. 37, iii–xviii (2013).
30. Chen, N., Kocielnik, R., Drouhard, M., Peña-Araya, V.: Challenges of Applying Machine Learning to Qualitative Coding. In: ACM SIGCHI Workshop on Human-Centered Machine Learning (2016).
31. Pons-Porrata, A., Berlanga-Llavori, R., Ruiz-Shulcloper, J.: Topic Discovery Based on Text Mining Techniques. Inf. Process. Manag. 43, 752–768 (2007).
32. Gupta, V., Lehal, G.S.: A Survey of Text Summarization Extractive Techniques. J. Emerg. Technol. Web Intell. 2, 258–268 (2010).
33. Guo, X., Wei, Q., Chen, G., Zhang, J., Qiao, D.: Extracting Representative Information on Intra-Organizational Blogging Platforms. MIS Q. 41, 1105–1127 (2017).
34. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Comput. Surv. 31, 264–323 (1999).
35. Aggarwal, C.C., Wang, H.: A Survey of Clustering Algorithms for Graph Data. In: Managing and Mining Graph Data. Advances in Database Systems. pp. 275–301. Springer, Boston (2010).
36. Berkhin, P.: A Survey of Clustering Data Mining Techniques. Group. Multidimens. Data. 25–71 (2006).
37. Lee, C.-J., Hsu, C.-C., Chen, D.-R.: A Hierarchical Document Clustering Approach with Frequent Itemsets. Int. J. Eng. Technol. 9, 174–178 (2017).
38. Blei, D.M., Edu, B.B., Ng, A.Y., Edu, A.S., Jordan, M.I., Edu, J.B.: Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022 (2003).
39. Gefen, D., Endicott, J.E., Fresneda, J.E., Miller, J., Larsen, K.R.: A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code: Studying Online Reviews and the Stack Exchange Community. Commun. Assoc. Inf. Syst. 41, 450–496 (2017).
40. Debortoli, S., Müller, O., Junglas, I., vom Brocke, J.: Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. Commun. Assoc. Inf. Syst. 39, 110–135 (2016).
41. Marton, A.: Purposive Selection and the Quality of Qualitative IS Research. In: Thirty Fourth International Conference on Information Systems (ICIS 2013) (2013).
42. Renaud, A., Walsh, I., Kalika, M.: Is SAM Still Alive? A Bibliometric and Interpretive Mapping of the Strategic Alignment Research Field. J. Strateg. Inf. Syst. 25, 75–103 (2016).

43. Achatz, H.: Partitionierung von Werkern aufgrund von Qualifikationsprofilen. In: Kossbiel, H. (ed.) Modellgestützte Personalentscheidungen. pp. 143–158. Rainer Hampp, München (1999).
44. Thorndike, R.L.: Who Belongs in the Family? Psychometrika. 18, 267–276 (1953).
45. Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. J. Comput. Appl. Math. 20, 53–65 (1987).
46. Ackerman, M., Ben-David, S.: Measures of Clustering Quality: A Working Set of Axioms for Clustering. Adv. Neural Inf. Process. Syst. 121–128 (2009).
47. Nguyen, Q.H., Rayward, Smith, V.J.: Internal Quality Measures for Clustering in Metric Spaces. Int. J. Bus. Intell. Data Min. 3, 4–29 (2008).
48. Keller, A.: Zum Erfolg von IKT-Start-ups in der deutschen Elektrizitätswirtschaft - Eine explorative Studie auf Basis von Experteninterviews. Dr. Kovač, Hamburg (2016).
49. Steigleder, S.: Die strukturierende qualitative Inhaltsanalyse im Praxistest: eine konstruktiv kritische Studie zur Auswertungsmethodik von Philipp Mayring. Tectum-Verlag, Marburg (2008).
50. Reilly, C., Wang, C., Rutherford, M.: A Rapid Method for the Comparison of Cluster Analyses. Stat. Sin. 15, 19–33 (2005).
51. Fraley, C., Raftery, A.E.: How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. Comput. J. 41, 578–588 (1998).