MASTER

Data mining journal entries

discovering unusual financial transactions

Arvaniti, V.

Award date:
2016

Link to publication
Data Mining Journal Entries: Discovering unusual financial transactions

Vasiliki Arvaniti

In partial fulfillment of the requirements for the degree of Master of Science in Business Information Systems.

Supervisors:

dr. ir. H. Eshuis (TU/e)
dr. R.J. (Rui) de Almeida e Santos Nogueira (TU/e)
dr. D. Fahland (TU/e)
Dennis van de Wiel (KPMG)

Eindhoven, August, 2016
ABSTRACT

Data mining has been described as "knowledge discovery in databases" or machine learning [12] and is used to discover patterns that are “hidden” in the data. Although data mining is a methodology that is widely applied in many scientific domains, data mining journal entries is a topic rarely researched. Journal entries are inspected by auditors and their aim is to create the company’s financial statement at the end of the year. However, these inspections are relying on predefined knowledge and do not reflect to the whole dataset but rather on specific subsets. In this research, we propose an Association rule mining approach to support auditors in finding unusual behavior out of the whole dataset. Together with a visualization tool and the extraction of unusual rules, this can increase the quality of review, provide insight into the whole dataset and consequently support domain experts in finding unusual financial transactions. The validation of our methodology took place in two parts. First, the effectiveness and usability of the visualization tool were evaluated by auditors. Auditors reacted positively to the visualization tool and they contributed to the exploration of further continuation. Second, the extraction of unusual financial transactions methodology was tested on a synthetic dataset in which unusual financial transactions were manually added. The evaluation indicated that the unusual financial transactions that were added, could be discovered by extracting the negative correlated rules.

Key words: Data Mining, Association Rules, Objective measures, Visualization, Journal Entries, Auditing
ACKNOWLEDGEMENTS

This thesis represents the completion of my Master’s program in Business Information Systems in Eindhoven University of Technology (TU/e). Several people supported me throughout the conduct of my master thesis. I would therefore firstly like to thank my head supervisor in the university, dr. ir. H. Eshuis for his valuable input, elaborate feedback and guidance from the very beginning to the end of the project. Next, I would also like to thank my second and third supervisors, dr. Rui Almeida and dr. Dirk Fahland, for taking the time and meeting with me when I mostly needed it. Their intermediate feedback and recommendations improved my thesis that I am delivering now.

Furthermore, I would like to thank all the colleagues at KPMG for helping me during my graduation project. It was a pleasure to meet real experts and learn from them. I am really grateful to Dennis van de Wiel, my company supervisor, who gave me the opportunity to conduct this research. He was always able to find some time to meeting with me, evaluate my work and provide me useful insights.

At this point, I would like to express my thanks to my dearest friends in Eindhoven, Elena, Sofia, Dimitris, Maria and Anthony for all the fun moments that we had during the last two years. I am also really grateful to them for all the support and love they gave me in the some difficult times. Moreover, special thanks to my friends in Greece, Vaso, Dimitra, Chrysa, Sofia, Marianna and Gerasimos, whether living in the same city or in a different country they were always caring and supportive.

This project is dedicated to my family for constantly believing in me and giving me unconditional love and encouragement throughout my life. None of my dreams would be possible without them. I conclude with a loving thank you to my boyfriend Angelos, after seven years, there still not a single day that he had not make me smile.

Vasiliki Arvaniti
August 2016
# Contents

Abstract ........................................................................................................................................ iii

Acknowledgements ......................................................................................................................... iv

List of Figures .................................................................................................................................. viii

List of Tables ................................................................................................................................... viii

1. Introduction ................................................................................................................................ 1
   1.1. Background .............................................................................................................................. 2
   1.2. Problem Statement .................................................................................................................... 5
   1.3. Research Questions ................................................................................................................... 6
   1.4. Scope ...................................................................................................................................... 6
   1.5. Methodology ............................................................................................................................. 7
   1.6. Main Contribution ..................................................................................................................... 8
   1.7. Outline ..................................................................................................................................... 9

2. Preliminaries ................................................................................................................................. 10
   2.1. Fraud and Anomaly Detection in Accounting .......................................................................... 10
       2.1.1 Fraud categories .................................................................................................................. 10
       2.1.2. Objectives of Auditors .................................................................................................... 12
       2.1.3. Traditional Methods for Fraud and Anomaly Identification ............................................ 13
       2.1.4. Conclusion ....................................................................................................................... 14
   2.2. Association Rules Data mining ............................................................................................... 15
       2.2.1. Basic Concept .................................................................................................................... 15
       2.2.2. Association Rules Algorithms ............................................................................................ 17
       2.2.3. Drawbacks and solutions of applying Association rules mining ...................................... 19
   2.3. Post-Processing methods in association rules mining ............................................................... 19
       2.3.1. Redundancy ...................................................................................................................... 20
       2.3.2. Interestingness measures .................................................................................................. 21
       2.3.3. Summarization and Grouping ............................................................................................. 25
       2.3.4. Visualization ..................................................................................................................... 25
   2.4. Conclusion ............................................................................................................................... 26

3. Data Understanding, SELECTION and Transformation .............................................................. 28
   3.1. Data Understanding .................................................................................................................. 28
   3.2. Data selection .......................................................................................................................... 29
LIST OF FIGURES

Figure 1: structure model for accounting data [11]
Figure 2: An overview of KDD process [42]
Figure 3: Report Outline
Figure 4: Example of fraud outlier [54]
Figure 5: Apriori Algorithm [14]
Figure 6: Classification of objective measures[36]
Figure 7: Detailed design of each layer [14]
Figure 8: Transaction database information
Figure 9: Visualization graph
Figure 10: Rules Extraction from Phi-coefficient
Figure 11: Confusion Matrix
Figure 12: Accounting tree [11]
Figure 13: Level of Description of Accounts IDs
Figure 14: Classification of data mining algorithms [12]
Figure 15: Current auditing procedure for Manual Journal Entries
Figure 16: Paper classified per purpose, the final scenario which were assigned [76]
Figure C.1: Graph “0-100”
Figure C.2: Graph for “1.000.001-10.000.000”
Figure C.3: Graph for “50.001-100.000”
Figure C.4: Graph for “1.001-10.000”
Figure C.5: Graph for “10.001-50.000”
Figure C.6: Graph for “1.001-10.000”
Figure C.7: Lift extraction plot
Figure C.8: Cosine extraction plot
Figure C.9: Certainty extraction plot
Figure C.10: Cosine extracted rules
Figure C.11: Lift extracted rules
Figure C.12: Phi extracted rules
Figure C.14: Certainty extracted rules
Figure D.1: Certainty extraction plot for the synthetic dataset

LIST OF TABLES

Table 1: 2x2 contingency table [75]
Table 2: Manual journal entries in the ledger
Table 3: MJE database
Table 4: Transaction Database of MJE example of 15 Journal Entries
Table 5: Transaction Database
Table 6: Example of generated association rules
Table 7: Rules per Value Category
Table 8: Extracted rules
Table 9: Visualization Evaluation methodology
Table 10: attributes created
Table 11: Synthetic transaction database
Table 12: Rules with the highest support
Table 13: Added “unusual” transactions
Table 14: Certainty measure for the unusual transactions
Table 15: Extracted unusual rules
Table 16: Inspected rules subset
Table 17: All rules from the synthetic dataset
Table A.1: Attributes description
Table A.2: Chart of Accounts
1. INTRODUCTION

In recent years, volume and complexity of financial transactions stored in information systems has increased significantly. Myriads of such transactions are recorded as journal entries in a general ledger, a database at the core of an information system providing the data required to build financial reports. To record transactions, accounting systems use double-entry accounting, implying that transactions are always recorded using two sides, debit and credit, and that the sum of debit side amounts should equal to the sum of credit side amounts [24]. A sufficient amount of transactions recorded in the database is audited to make sure that no risk is left unaddressed. Transactions auditing is still conducted manually, using procedural methods such as interviews and inspections of selected documents [11], or other conventional methods (substantial audit procedures performed for single business transactions, based on “Analytics” and “Test of Detail”) [8]. Nevertheless, this approach is considered time consuming and error-prone [11].

Despite the fact that such complicated financial transactions can now be presented in a more comprehensive format by running varied query commands, auditors are still required to make assumptions and interpret results. These methods, however, do not allow them to extract information from the available data when it is unknown or hidden [24]. As a result, threats like fraudulent activities and human errors, which pose a great risk to the area of journal entries, might not be traced [18]. In addition, considering the vast amount of data requiring inspection, such methods could lead to a process that does not even correspond to the real data. Therefore, current auditing procedures present deteriorating sufficiency [11].

Taking up the aforementioned challenges, the present thesis proposes an association data mining model to support auditors in identifying unusual financial transactions. Financial transactions are represented by the accounts posted inside a journal entry. For this purpose, a case study was conducted within the KPMG Risk Consulting IT Advisory department of Eindhoven. Access to a general ledger was given containing manual journal entries. In this chapter, first a general ledger will be described, then a brief introduction about data mining algorithms will be given and last, former research findings about data mining in journal entries will be analyzed. Furthermore, an effort will be made to appoint an accurate definition to the research problem, derived both from literature and KPMG needs. Finally, research questions and scope of thesis will be provided, followed by the research methodology and contribution.
1.1. Background

This section introduces initially a general ledger, then data mining algorithms are briefly described and finally it provides some background knowledge based on literature about data mining journal entries.

**General Ledger**

The general ledger is a repository of financial transactions. It works as a database containing all recorded transactions that impacted a company’s assets, liabilities, equity, revenue and expenses. At the end of the year, the companies’ financial statement is prepared based on the general ledger, which is the primary and authorized source of data for financial accounting reports, including the income statement and balance sheet. The recorded transactions in a general ledger are following the Double-entry bookkeeping system. The golden rule of this system is that the amount debited equals the amount credited. Transactions are represented by journal entries, therefore each journal entry can consist of at least two lines (a simple journal entry). Each line of the journal entry lists the accounts to be credited and the accounts to be debited with their corresponded amounts. A journal entry should also include other information about the posted accounts, such as date of the entry, a unique number to identify the entry, name of the person recording the entry etc. Example of a journal entry structure is shown in Figure 1.

<table>
<thead>
<tr>
<th>Account Number</th>
<th>Description</th>
<th>Date</th>
<th>Debit Amount</th>
<th>Credit Amount</th>
<th>Details</th>
<th>Account Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>60000</td>
<td>Salaries</td>
<td>27/Aug/2007</td>
<td>Debit 56,070</td>
<td>Credit</td>
<td>Payment of salaries Expenses</td>
<td></td>
</tr>
<tr>
<td>30000</td>
<td>Bank</td>
<td>27/Aug/2007</td>
<td>Debit</td>
<td>Credit 56,070</td>
<td>Payment of salaries Cash &amp; Cash Equivalents</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: structure model for accounting data [11]

Many types of journal entries are automated and posted in the general ledger through specific periodic scheduled processing. However, journal entry corrections or adjustments might be needed outside of the automated processing. This kind of journal entries are called Manual journal entries(also called top-side journal entries) and have been posted to the general ledger by a user and not by a system. A general ledger at the end, contains both automatically generated journal entries and manual journal entries.

**Data Mining Algorithms**

Data mining has been described as "knowledge discovery in databases" or machine learning [12]. Data mining algorithms can be used to discover patterns, relationships, factors, clusters, associations, profiles and predictions previously “hidden”. Data mining can be classified into two major models: predictive and descriptive. Further classification of data mining models provided by paper [12] can be found in Appendix A1.

According to [29], predictive models’ conceptual logic is that given a sequence of desired outputs, the machine has to learn to produce the correct output given the new input, in which case input data is called “training data” and has a known label or result such as spam/non-spam or a stock price at a time. A model is developed through a training process in which predictions are made and corrected if wrong. The training process continues until the model achieves a desired level of accuracy on the training data. Example algorithms are classification and regression. Although both predictive and descriptive models include functions capable of finding different hidden patterns in large data sets, in predictive models (supervised
learning) there is a clear measure of success or failure that can be used to judge adequacy in particular situations and compare the effectiveness of different methods over various situations.

Descriptive models consist of unsupervised learning that describes the historical events and the presumed or real relationships between elements creating them. The input data is not labelled and does not have a known result. A model is prepared by deducing structures present in the input data. This may be to extract general rules through a mathematical process to systematically reduce redundancy, or to organize data by similarity. There are several unsupervised learning techniques applicable for data preprocessing in a complementary way. Among the most commonly used in practice are Association Rules, Clustering Analysis, Principal Components and Self Organization Maps [29]

Association rules mining aims to find interesting associations (relationships, dependences) in large sets of data items and obtain the hidden associations between two or more items in databases. These items usually appear forming sets called transactions. When items tend to co-exist in most of the transactions, they are possibly related in some way, and this relationship can be described by an association rule. Thus, association rules can be viewed as general relations between two or more attributes described by means of a convenient quantifier.

Clustering techniques are widely used for summarizing data objects and capturing key data characteristics. The goal of cluster analysis is to partition the observations into groups (“clusters”) so that the pairwise dissimilarities between those assigned to the same cluster will be smaller than the ones assigned to different clusters. Principal components constitute a sequence of data projections, mutually uncorrelated and ordered in variance. Finally, Self-Organizing Maps (SOM) is a method that can be viewed as a constrained version of K-means clustering, in which prototypes are encouraged to lie in a one- or two-dimensional manifold in the space provided.

Data mining journal entries

Although data mining is a methodology that has been applied in almost every scientific domain, gathering scientific literature related to data mining in general ledger and real-world applications of the method was challenging. Existing studies [19-28] focus on applications of data mining technology in audit datasets, such as outlier detection, statistical analysis, association analysis, and decision trees. Despite the fact that scientific papers available emphasize the importance of data mining for fraud detection, only a few include analyses of real case studies of data mining techniques on journal entries. Lack of published research can be attributed to the difficulties researchers encounter in gaining access to journal entries, as databases are proprietary [24].

Data mining techniques on real datasets have been applied in the following studies:

- In paper [21], cluster analysis is applied in the accounting domain, for anomaly detection in audit for a life insurance company. With cluster analysis, data is grouped so that claims with similar characteristics are put in the same cluster. The study intends to examine the use of clustering technology in automating fraud filtering during audit. Having no fraud samples, this study demonstrates that clustering analysis for auditing can be effective in fraud and anomaly detection, assuming that clusters with small populations and single claims differing from other claims in the same cluster are to be flagged for further investigation.

- A noteworthy case is analyzed in [24], where digital analysis (also first-digit law) is initially used as a statistical technique to predict if the first digit of journal dollar amount differs from that expected
from Benford’s Law by calculating the Chi-square distribution. Journal entries presenting variations in distribution are flagged as suspicious and require further investigation by the auditors. Additionally, a last digit analysis is conducted. Particular journal entries with abnormal distribution of last digits are also considered to be fraudulent. Finally, unusual patterns in means of journal entry activities are investigated. This study recommends that more data mining techniques should be applied in journal entries in order to start seeding the dataset with fraud indicators (e.g. pairs of accounts that would not be expected in a journal entry).

- The main purpose of [19] is to detect journal entries that can materially misstate financial statements. This paper proposes a bipartite model for detecting “suspicious” journal entries in order to assist auditors in identification and risk assessment of material misstatement to financial statements. The paper defines “suspicious” journal entries as journal entries having both a large monetary amount and a low probability of occurring; in other words, “suspicious” journal entries are rare and include a monetary amount that is large enough to materially misstate financial statements.

- Also, paper [20] suggests a model based on SOM, an unsupervised learning algorithm ensuring that journal entries are free of material misstatements. As before, the paper intends to assist auditors in detecting suspicious journal entries. More specifically, SOM is employed to derive a reference model capable of describing the behavior of legitimate journal entries. Novel journal entries are compared to the reference model to determine whether they are “suspicious” or legitimate; the comparison is made by calculating the quantization error, a distance metric between individual journal entries and the SOM-based reference model. A journal entry having a quantization error that exceeds a specified threshold is deemed to be “suspicious”.

- A previous thesis conducted in KPMG uses general ledger data as input to extract instances of business processes in a format suitable for interpretation by financial professionals for audit as well as fraud investigations [2]. In the research, the link analysis data mining technique applied is categorized as an unsupervised modeling method, as it can detect patterns in a dataset based on graphs rather than statistical features. It is assumed that transactions are more likely to be connected when they are closer to each other in the journal entry. This approach can produce useful results when applied to journal entries that are relatively small in lines, while for more complex journal entries it quickly becomes computationally infeasible.

To conclude, review of research findings aforementioned shows that detecting unusual behavior in journal entries is a challenging task. Data mining techniques could leverage its ability to inspect the whole spectrum of the data and detect unusual behavior, but, as stated in paper [24], it is still a topic that is not in depth researched. All above literature studies conducted explorative research in order to assist auditors in more effective and efficient ways to detect unusual behavior, using unsupervised data mining techniques in order to identify unusual patterns from the data due to the lack of labelled journal entries as fraudulent.

However, none of the above papers took the account pairs of the journal entries into consideration but rather inspected the account amounts individually, to determine if the transaction is unusual or not. Yet, WorldCom Inc., in one of the most egregious financial statement frauds, used inappropriate account relocations in the journal entries to hide their fraudulent activities [24]. For instance, significant transfers were made from what was effectively a suspense expenditure account, “Prepaid Capacity Costs,” to the “Construction in Progress” account, which was treated as capital expenditure [24]. From the previous real
example and from papers’ [24] [2] suggestions, detecting pairs of accounts that would not be expected in a journal entry might be a fraud indicator.

1.2. Problem Statement

Auditors and fraud examiners could use manual means to review the general ledger. However, this generally proves ineffective given the breadth of the ledger and the limitations of the human eye. Although a person’s judgment remains of high value when reviewing entries, relying exclusively on manual means might not be the most effective approach. It is clear that current auditing procedures, as described by experienced auditors of KPMG (and shown in Appendix B2), may involve unusual financial transactions (journal entries containing accounts) stored in the ledger, unidentifiable using data filtering (running query commands) and sample reviewing. Concerning Manual Journal Entries (MJEs), which are executed by humans and not by system-generated tasks, the risks are even higher [2]. Among associated risks in review of entries deriving from literature [2] [11] [18] [24] are the Quality of Review leading to Opportunity for Fraud given the company’s constraints in comprehension, review and control of journal entries.

The present thesis’ purpose is to support auditors in a more comprehensive way, by using data mining methods to inspect the whole dataset, in an effort to trace unusual financial transactions. As it was mentioned in the previous section, existing literature (1.1) has presented methods for the detection of unusual journal entries, but none of the scientific papers reviewed within the framework of this research focuses on the pairs of accounts in a journal entry.

Given that journal entries use the double entry booking system and contain two accounts at minimum, debit and credit, it is important for auditors to have an overview of the whole spectrum of the dataset. In other words, auditors need to know how accounts (financial transactions) are related with each other and how much money is involved in them. This could improve the quality of review of journal entries and expose unusual financial transactions that might need further investigation.

Unusual financial transactions in this thesis are defined as account pairs inside journal entries containing an unusual amount of money compared to their frequent behavior.

In this research there is no specific dataset labelled as unusual, thus predictive data mining models cannot be applied. Yet, unsupervised data mining algorithms identifying structures and relations among data can be used.

In addition, since this research focuses on account pairs (sets) included in journal entries and aims to find ways of comparing their co-occurrences with their frequent behavior and provide an overview of the accounts connection in the whole dataset, association rules mining could be an appropriate method to be applied. In that sense, the transactions in a transactional database containing item sets will represent journal entries in a general ledger containing accounts information. Thus, accounts related in some way can be described and viewed by an Association Rules.

Recent research approaches in Association Rules are based on obtaining different kinds of knowledge, referred to as peculiarities, infrequent rules, exceptions or anomalous rules. Information obtained by these new types of rules can in many cases be more useful than that obtained by simple association rules. Several authors over the years have proposed additional measures in association rules mining, in order to
detect the anomalous or unexpected/unusual behavior, and defined this step as a post-processing step.[14-18][32-35][37][45][49].

1.3. Research Questions

Based on the previous section, we aim to address our research problem by applying association rules mining in the general ledger. Association Rules mining are applied in a transactional database in which each transaction contain a set of items. Financial transactions are represented by journal entries which contain account pairs (or a set of accounts). The goal is to support auditors in finding unusual financial transactions. To achieve that, we aim to increase the quality of review of the journal entries and, in top of that, extract unusual financial transactions (as they have been defined in the previous section) that might need further inspection.

Therefore, the main research question addressed in this thesis is: “How can association rule techniques support domain experts in finding unusual financial transactions from the whole dataset of a general ledger?”

To comprehend this problem and reach an answer to the main research question, the following questions will be investigated in more detail:

A1. What kind of data transformation in the general ledger dataset is required to apply association rules?

A2. What are the necessary parameters to be set in the Association Rules Algorithm (ARA)?

A3. How can we obtain a holistic view of the financial transactions, represented by rules, in order to improve the quality of review?

A4. How can we extract the unusual financial transactions, represented by rules, among the results?

A5. How can we evaluate our results?

1.4. Scope

The dataset analyzed contains all manual journal entries that were recorded in one fiscal year in the Netherlands by a customer of KPMG. Thus, in the next chapter, the term “journal entries” will be exclusively used to describe manual journal entries. Constraints of time availability, experience and specification of goal setting limit the study to this specific ledger. As a result, assurance that applied method will be similarly effective on other ledgers has to be as limited.

The scope of this thesis includes tasks as:
1. Applying ARA in journal entries to identify unusual financial transactions (accounts combinations in journal entry) for further inspection, in terms described in the previous chapter.
2. Providing a holistic view of financial transaction (accounts combinations in journal entry) stored in the general ledger.
3. Evaluation of the quality of the review tool by domain experts.
4. Since we did not have any suspicious or unusual dataset, the evaluation of the method applied for the extraction of the unusual rules is based on synthetic dataset.
Not included in the scope:
1. Determine which objective measure performed better in the identification of unusual financial transactions.

1.5. Methodology

The methodology for this research derives from the KDD (Knowledge Discovery in Databases) process [42]. Data mining is the core of KDD process involving the inferring of algorithms that explore, develop model and discover unknown patterns. It is used for large datasets that cannot be manually analyzed and the final goal is to extract useful/actionable knowledge. This methodology is extremely popular in association rules mining due to its significant component, the post processing step [67]. Therefore, this thesis also follows this KDD methodology for association rules mining. The methodology consists of six phases: Selecting the problem area, Selecting and creating a dataset on which discovery will be performed, Preprocessing of the data, Data mining algorithm, Post processing of the knowledge derived, evaluation and knowledge integration [67]. In Figure 2, the phases of the general KDD process are shown. The phases are in sequence, but loops may also occur during the project. The following picture illustrates all possible combinations within the process.

![Figure 2: An overview of KDD process [42]](image)

Different steps of the proposed methodology are briefly analyzed below:

**Selecting the Problem:** Prior to any processing, this phase aims to identify the goal of the knowledge discovery. Therefore, the first phase aims to understand the company’s environment and its auditing process. Meeting with experts both from a data and an auditing perspective will take place in order to define the problem and understand the project objectives and requirements in terms of business. During this phase, a literature study is conducted, to indicate most suitable models for appliance. Knowledge acquired in this phase will be converted into a data mining problem definition and a preliminary plan will be designed to achieve the list of objectives.
Selecting and Collecting the Dataset: This phase begins when the company providing the manual journal entry ledger for analysis purposes. A close review of the data acquired will be conducted for further data familiarization and quality problem identification, leading to export of first insights. Since data provided by the case study is stored and accessed in a SQL database, Microsoft SQL Server will be the tool in use for our purpose. After finding out what data is available and what it represents, the selection of attributes needed for our research problem is performed.

Data preprocessing: The data preparation phase will be implemented in both the SQL server and the R environment. R is a free tool, available in any platform, and produces excellent graphics. It was selected among other data mining tools (Weka, RapidMiner, Matlab) because of its wider optionality of association rules packages to automate particular tasks. This phase covers all activities needed to construct the final dataset, which will then be fed into the models. Data should be in a format capable to apply ARA for this specific business problem (Research Question A1).

Modeling: In this phase, the dataset prepared in the previous step will be applied in a chosen ARA. In a preliminary stage, a decision is made on which algorithm of ARA to use. Since algorithms of ARA include parameters, a decision is made to define the thresholds of the returned rules (Research Question A2).

Post-processing: In this phase, the results of knowledge extracted in the previous step are further analyzed. The R environment provides a vast amount of post-processing methods for the resulted rules. Based on the research goal, providing a holistic overview of the financial transaction(account combinations in journal entries) and the discovery of unusual account combination that need further inspection, the post-processing methods are selected. After this task, we will be able to collect the rules detected as unusual amongst all the resulted rules and have an overview of the financial transactions stored in the dataset (Research Question A3, Research Question A4).

Evaluation: In the final phase, the models (tools) developed in the previous step are evaluated. Quality of review findings will be evaluated by domain experts by conducting interviews. Additionally, for the unusual rules that need further inspection, and since we do not have labelled data to test whether measures provide accurate results, a synthetic dataset is constructed which will contain “unusual” transactions to measure the performance of our approach (Research Question A5).

Deployment: In this phase, depending on the requirements of the project, it can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. The project team decides whether to finish this project and move on to deployment, initiate further iterations, or set up new data mining projects. Given the time limitations of this project, deployment will include knowledge gained from evaluation as conclusion, as well as limitations and further research direction.

1.6. Main Contribution

In this research, we aim to apply association rules data mining with additional, post-processing methods, in order to gain an overview of the frequent and infrequent pairs(or set) of accounts and extract the unusual/unexpected ones that need further inspections from the auditors.

The practical contribution of this thesis is providing KPMG with a tool which provides a visualization overview of financial transactions (account pairs inside journal entries) and, in addition, the extraction of
the unusual ones for further investigation. Considering current techniques auditors use, the results reflect only to a selected portion of the dataset, which derives from data filtering, manual means and auditors’ judgment.

Furthermore, this thesis aims to contribute to literature by applying association rules mining in journal entries for supporting auditors to identify unusual journal entries. Existing work [19] [20] [21] [24] performs similar experiments. However, it does not take the account pairs into consideration and therefore uses different data mining models.

1.7. Outline

The Report outline follows the research process conducted during project execution. Figure 3 shows the activities carried out during research, including corresponding research questions and subjects of each report chapter.

First, the problem area is introduced and defined (Chapter 1), in addition to findings of previous studies already included in the Introduction section. Chapter 2 includes preliminaries about what is perceived as a suspicious journal entry in accounting, as well as description and findings from literature on Association rules and Post-processing activities. In this chapter, decisions about what is going to be applied in the next chapters are taken. Chapter 3 describes the case study conducted in this research, concerning the data, selection of attributes and transformation activities. Chapter 4 describes the model output after setting the association rules parameters. In Chapter 5, post-processing activities are applied: first, redundant rules are removed, then results are visualized and finally, interestingness measures are applied and the unusual rules for further inspection are extracted. Chapter 6 describes the evaluation of results. Finally, in Chapter 7, the contribution of this work, limitations and possible directions for future work are discussed.
2. PRELIMINARIES

The preliminaries chapter provides insights related to the research topic. Since unusual financial transactions are the motivation of the thesis, the chapter initially presents a summary of literature analysis conducted to investigate the definition of fraudulent journal entries in accounting and how they are currently found. Then, detailed literature analysis about Association Rules Algorithms is conducted, in order to obtain a depth knowledge about the algorithms characteristics since in this thesis we intent to discover the co-occurrences of the account pairs in the whole dataset. The previous sections are followed by one on Association Rules Algorithms’ post-processing techniques. The analysis of the post-processing technique aims to summarize the research that has conducted to analyze further the association rules that are derived from an Association rule mining.

2.1. Fraud and Anomaly Detection in Accounting

The goal of this thesis is to support auditors in finding unusual financial transactions. Among the unusual financial transactions there might occur transactions that are uncommon but do not indicate a suspicious or fraudulent activity, but on the other side, the unusual financial transactions might also include fraudulent activities.

Fraud, according to [51], is an intentional act by one or more individuals among management, employees or third parties, resulting in a misrepresentation of financial statements. Fraud can also be seen as the intentional misrepresentation, concealment or omission of the truth for the purpose of manipulation to the financial detriment of an individual or an organization. Therefore you are not supposed to find it fraud [54]. Even if you are looking for it, it can be hard to find. Auditors have traditionally been trained to find evidence to support a premise that assertions embodied in financial statements are right. However, audit processes designed to detect such fraud remain problematic as there is no deployed body of knowledge among auditors about what fraud looks like.

2.1.1 Fraud categories

The Auditing Standard Board defines two categories of occupational fraud: misappropriation of assets and fraudulent financial report [5]:

Misappropriation of assets

Asset misappropriation fraud happens when people entrusted to manage the assets of an organization steal from it. Asset misappropriation fraud involves third parties or employees in an organization, who abuse their position to steal from it through fraudulent activity. It is also known as insider fraud. This type of fraud can be committed by company directors, or its employees, or anyone else entrusted to hold and manage the assets and interests of an organization. Typically, the assets stolen are cash or cash equivalents, such as credit notes or vouchers. However, the fraud can extend to include company data or intellectual property [6].

With misappropriation of assets an organization’s assets are taken away through trickery or deception, rather than by force (e.g., theft of company cash, false billing schemes or inflated expense reports). Next to the "act" of theft, concealment and conversion should also preside. In a recent study by the ACFE (Association of Certified Fraud Examiners), approximately 85% of all asset misappropriation cases involved
the misuse or theft of cash. Symptoms of fraud are often referred to as “red flags”. A research conducted by [52], indicated the red flags for misappropriation of assets:

1. Accounting anomalies, such as faulty journal entries, inaccuracies in ledgers, or fictitious documents.
2. Internal control overrides and breakdowns.
3. Analytical fraud symptoms, including procedures or relations that are unusual or too unrealistic to be plausible. For example, transactions or events that happen at odd times or places; that are performed by or involve people who would not normally participate; or that include odd procedures, policies or practices. They may also include transaction amounts that are too large or too small.
4. Lifestyle symptoms (people who commit fraud usually meet their immediate need and then gradually start to increase their lifestyles).
5. Unusual behaviors of people (people who are involved in fraud often feel stress and, as a result, change their behaviors to cope with this stress).
6. Tips and complaints that something is suspicious.

Findings from ACFE 2004 Report to the Nation on Occupational Fraud and Abuse into the audit showed that frauds involving the misappropriation of assets identified that cash is the targeted asset 93.4 per cent of the time. This increase from the ACFE’s 2002 Report showed that cash was the targeted asset in 90.1 per cent of the cases studied [7].

Financial statement fraud
While misappropriation of assets is important, detection of financial statement fraud is of greater concern. Financial statement fraud is the least frequent form of occupational fraud, but by far the most costly. This type of fraud usually has greater probability of giving rise to a material misstatement and of being committed by upper management. The exploratory research of paper [24] concluded and enumerated fraudulent behavior of the following types:

1. Inappropriate accounting reallocations, including transfers from flows to stocks. For example, significant transfers were made from what was effectively a suspense expenditure account, “Prepaid Capacity Costs” to a “Construction in Progress” account, which was treated as capital expenditure.
2. Accounting treatments designed to influence disclosure rather than recognition. For example, line costs were transferred to accounts that rolled up into “Selling, General and Administrative Expenses (SG&A).”
3. Adjustments that may have not changed the reported profits, but did change the allocation between gross and net profit disclosures.
4. Ill-concealed journal entries, with large adjustments in rounded amounts that would be obvious to the most casual of inspections.
5. Inappropriate journal entries often accompanied by failures in documentation and breaches in normal internal controls.
6. Adjustments almost universally being carried out at the corporate level. In many cases, however, these “top side” adjustments made at the corporate level required adjustments in operating divisions and international operations.
A review of the circumstances leading to frauds perpetrated in the past few years by companies such as Enron, Cendant, WorldCom and HealthSouth, to name but a few, shows that they were involved in deliberate manipulation of financial reports. Perpetrators used false entries including fraudulent journal entries in a variety of schemes to manipulate revenue and earnings, falsely capitalize expense items as assets, conceal liabilities and tamper with reserves [54].

In an effort to provide guidance to auditors in meeting the challenge of exposing financial statement fraud, the AICPA issued SAS No. 99 in 2002 [5]. Among other things, such standards provide auditors with a checklist of risk factors to consider when making a fraud risk assessment. The highest risk “red flag” in financial statement fraud is found in false accounts allocations [24] [60]. Committee of Sponsoring Organization Report [62] revealed that about 50 per cent of frauds involve overstated revenues, either by reporting revenues prematurely or by creating fictitious revenue transactions. Other fraudulent account relocations included significant or unusual accrual transactions [60].

Furthermore, based on study [60], the frequency of financial statement fraud has not shown signs of decline. On the contrary, a 2005 biennial survey of more than 3,000 corporate officers in 34 countries conducted by PricewaterhouseCoopers revealed a 140 per cent increase in discovered number of financial statement fraud from 10 per cent of companies reporting financial statement fraud in a 2003 survey to 24 per cent in a 2005 survey.

2.1.2. Objectives of Auditors

Based on the Audit Guideline of Fraud [53] defined by the European Court of Auditors, fraud is not the main focus of an audit. Auditing standards require auditors to carry out specific fraud-related procedures. Indeed, an auditor’s job is to make sure that financial statements as a whole are free from material misstatement, whether caused by fraud or error. Thus, auditors must have a reasonable basis for expecting that their work will detect material fraud. The objectives of auditors are:

a. To obtain sufficient and appropriate audit evidence regarding the assessed risk, through designing and implementing appropriate responses.
b. To respond accordingly to fraud or suspected fraud identified during audit.
c. To maintain professional skepticism throughout the audit.
d. To discuss the susceptibility of the entity to fraud, including how fraud might occur, with the audit team.
e. To document the auditor’s consideration of fraud.

The auditor, however, needs to be aware that:

- The risk of not detecting a material misstatement resulting from fraud is higher that the risk of not detecting a material misstatement resulting from error, as fraud may involve sophisticated and carefully organized schemes designed to conceal it. Such could be forgery, deliberate failure to record transactions, or intentional misrepresentations given to auditors. Attempts of concealment may be even more difficult to detect when accompanied by collusion, which may trick auditors to believe in false audit evidence.
- The risk of auditors’ not detecting a material misstatement resulting from management fraud is greater than for employee fraud, as management is frequently in a position to directly or indirectly
manipulate accounting records, present fraudulent financial information or override procedures designed to prevent similar frauds by other employees.

2.1.3. Traditional Methods for Fraud and Anomaly Identification

In a large company, millions of journal entries can be made every year and it is difficult to imagine how an audit can be conducted effectively using a manual review of all entries or even a statistical sample. Using means of technology, auditors can design a range of tests, including ones to identify the characteristics noted above.

Fraud detection is an examination of the facts to identify the indicators of fraud. Reviewing and improving internal control systems is the primary defense against fraud and abuse. Researchers and practitioners have made several attempts to identify fraud indicators and build fraud prediction procedures. The need to fight fraud has exerted strong pressure for auditors to assume this role. In this respect, auditors’ role to prevent and detect fraud is a very important part of their job, which in turn makes use of advanced and complex computer systems as important as ever. Economically, financial fraud is rapidly becoming an alarming situation, and effective detection of accounting fraud has always been an important but complex task for accounting professionals [57].

According to [53], in most cases fraud in enterprises is discovered by chance. They also mention that although enterprises appear to have control systems in place, in many cases they are ineffective, primarily due to management either overlooking controls, or colluding in circumventing them. Also, stockholders, audit committees and top management are more insistent about being informed of fraud and more likely to put the blame on auditors if it is not found and reported. Current methods that auditors use are:

*Analytical tools for testing*

SAS 99 requires the auditor to undertake a variety of analytical and planning tasks and substantive audit procedures to support detection of errors rising from fraudulent financial reporting. Tests like these are conducted based on a checklist provided from the ACFE which includes the “red flags”. Tests include functions such as to: Extract journal entries, Sample journal entries, Summarize amounts, Total count, Classify, Sort, Compare, Stratify, Merge, Split and Join, Calculate rations of accounts, Test Duplicated Payments, Test Aging Calculations, Produce graphs and charts. From above mentioned analytical tools graphs such as Figure 4 are derived, with which an auditor can easily examine a large data set for unusual amounts [54].

![Figure 4: Example of fraud outlier](image)

However, for this method, paper [61] found that
• auditors using long checklists tend to be inaccurate in assessing fraud risk;
• Auditors generally overweigh clues about management’s character, clues likely to be wrong;
• Auditors are often insensitive to new evidence regarding fraud risk;
• When auditors use procedures based on prior audits, these become predictable and less effective.

To conclude, based on paper [60] conducting an academic literature review for financial statement fraud, there is not enough evidence to claim that the use of checklists improves auditors’ ability to assess fraud risk. Much of the research reviewed suggests that using checklists may actually restrain the auditors’ generation of ideas.

**Benford’s Law**

In 1938, Frank Benford published his “Law of Anomalous Numbers” paper, presenting the digital patterns expected in natural data sets and showing that they are counter intuitive, as lower digits are assigned higher probabilities of occurring than higher digits. Benford’s Law is based on the fact that many numbers normally used in business (and elsewhere) are not random, but rather follow some ordered progression. For example, a chart showing wealth will show that it is not uniformly distributed; a few people have much wealth and many people have less wealth. Sales, inventory and disbursements are also not uniformly distributed. Benford’s Law uses this fact to help point to fraud, inefficiencies and other forms of data manipulation [56].

One example of Benford’s Law in accounting as a fraud indicator was found in New Zealand firms, where earning numbers did not agree to expected distributions. Rather, the numbers contained more zeros in the second digit positions than expected and fewer nines, thus implying that when a firm had earning such as $1.900.000 they rounded up to $2.000.000 [54]. Benford’s law, as applied to auditing, is not more than a complex form of digital analysis. It examines an entire account to determine whether the numbers fall into the expected distribution. If used properly, digital analysis conducted on transaction level data could enhance auditors significantly, by identifying specific accounts in which fraud might reside, so that they can then analyze the data in more depth.

Benford’s analysis is a particularly useful tool as it is conducted on specific accounts using all the data available, instead of aggregated data, thus improving capability of identifying specific accounts for further analysis and investigation. However, as stated in paper [54], it should only be applied to accounts conforming to Benford’s distribution. Additionally, it is stated that Benford’s analysis is likely to be useful only in certain account types and auditors must be cognizant of the fact that certain types of frauds will not be found using this analysis. Furthermore, book [57] states that although Benford’s Law could have useful application in detecting financial statement fraud, it might not be able to narrow the list of possible fraudulent journal entries down to a manageable size. For example, if there is an unusually large quantity of journal entries starting with a particular digit, there may be thousands of entries identified for analysis. That was also the reason that in paper [24] many false positive suspicious journal entries have been identified.

**2.1.4. Conclusion**

Fraud in journal entries is defined as an intentional act by one or more individuals among management, employees or third parties, resulting to misrepresentation of financial statements. It is divided in two categories: Misappropriation of Asset and Financial Statement Fraud. Financial Statement Fraud is
considered more severe of the two, as it is extremely difficult to trace, comparing to the latter. Specifically, about 50 per cent of frauds involve overstated revenues, either by reporting revenues prematurely or by creating fictitious revenue transactions. Other fraudulent account relocations included significant or unusual accrual transactions [60]. An example of financial statement fraud is WorldCom Inc., where reallocation of operating expenses accounts to capital expenditure accounts took place [24][54].

Studied literature presents two traditional procedures for fraud identification: Analytical Procedures and Benford’s Law. However, based on [60] [53], traditional procedures have had little success in identifying fraud. One of the reasons could be that management is in a position to hide account irregularities or explain away any unusual deviations in accounts. Therefore, it is suggested that even with the introduction of new technology-based tools in auditing, such as data mining software, continuous auditing and pattern-recognition software, this area would require constant examination to test the efficiency and effectiveness of new tools to detect fraud.

Based on the above information gathered from literature, financial statement fraud seems to be difficult to be identified with current tools. Data mining could be a technique enabling auditors to better trace irregularities, particularly regarding identification of the unusual account relocations which, as previously mentioned, is the most known financial statement fraud.

2.2. Association Rules Data mining

Association rule mining, one of the most well researched techniques of data mining, was first introduced in paper [3]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, the market, risk management and inventory control. In this section, an association rules analysis will take place and then various association mining techniques will be briefly described. Last, based on association rules limitations, further post-processing techniques will be presented. Based on our research problem, the algorithm that is going to be used, as well as limitations, will be decided.

2.2.1. Basic Concept

Association rule analysis has emerged as a popular tool for mining commercial databases. Goal is to find joint values of the variables $X = (X_1, X_2, \ldots, X_p)$ that appear more frequently in the database. It is most often applied to binary-valued data $X_i \in \{0, 1\}$, where it is referred to as “market basket” analysis.

In the context of “market basket”, the observations are sales transactions, such as those occurring at the checkout counter of a store. Variables represent all of the items sold in the store. For observation $i$, each variable $X_j$ is assigned one of two values; $x_{ij} = 1$ if the $j$th item is purchased as part of the transaction, whereas $x_{ij} = 0$ if it is not purchased. Variables frequently having joint values of one, represent items that are frequently purchased together. That kind of information can be quite useful for stocking shelves, cross-marketing in sales promotions, catalog design and consumer segmentation based on buying patterns [29].

In a more comprehensive format, paper [63] defined association rules as: Let $I = I_1, I_2, I_3, \ldots, I_m$ be a set of $m$ distinct attributes, $T$ be a transaction containing a set of items such that $T \subseteq I$, $D$ be a database with different transaction records $T_s$. An association rule is an implication in the form of $X \rightarrow Y$, where $X, Y \subseteq I$
are sets of items called itemsets, and \( X \cap Y = \emptyset \). \( X \) (Left hand side, LHS) is called antecedent while \( Y \) (Right hand side, RHS) is called consequent, the rule means \( X \) implies \( Y \).

In general, a set of items (such as the antecedent or the consequent of a rule) is called an itemset. The number of items in an itemset is called the length of an itemset. Itemsets of some length \( k \) are referred to as \( k \)-itemsets.

Generally, an association rules mining algorithm contains the following steps:

- Sets of candidate \( k \)-itemsets are generated by 1-extensions of the large \((k-1)\)-itemsets generated in the previous iteration.
- \textit{Support} for candidate \( k \)-itemsets are generated by a pass over the database.
- Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large \( k \)-itemsets.

This process is repeated until no more large itemsets are found.

The "support" of the rule \( \text{supp}(A \rightarrow B) \) is the fraction of observations in the union of the antecedent and consequent. It can be viewed as an estimate of the probability of simultaneously observing both item sets \( \Pr(A \cup B) \) in a randomly selected market basket. "Confidence" \( \text{conf}(A \rightarrow B) \) of the rule is its support divided by the support of the antecedent [29], which can be viewed as estimate of \( \Pr(B|A) \):

\[
\text{Conf}(A \rightarrow B) = \frac{\text{supp}(A \rightarrow B)}{\text{supp}(A)}
\]

According to [3], Confidence is not downward closed and was developed together with support by Agrawal et al. (the so-called support-confidence framework). Support is first used to find frequent (significant) itemsets exploiting its downward closure property to prune the search space. Then confidence is used in a second step, to produce rules from frequent itemsets that exceed a minimum confidence threshold.

A problem with confidence is its sensitivity to the frequency of consequent \( Y \) in the database. Due to confidence’s calculation manner, consequents with higher support will automatically produce higher confidence values even if there is no association between the items [3].

Therefore, besides support and confidence, highly integrated in the association rules is also the parameter Lift. Lift is a measure of targeting model’s (association rule) performance at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model. This is an estimate of the association measure [29]:

\[
\frac{\Pr(A \rightarrow B)}{\Pr(A) \Pr(B)}
\]

As an example, suppose the itemset \( K = \{\text{peanut butter, jelly, bread}\} \) and consider the rule \( \{\text{peanut butter, jelly}\} \rightarrow \{\text{bread}\} \). A support value of 0.03 for this rule means that peanut butter, jelly and bread appeared together in 3% of the market baskets. A confidence of 0.82 for this rule implies that when peanut butter and jelly were purchased, 82% of the time bread was also purchased. If bread appeared in 43% of all market baskets then the rule \( \{\text{peanut butter, jelly}\} \rightarrow \{\text{bread}\} \) would have a lift of 1.95.
2.2.2. Association Rules Algorithms

Association rules were first introduced in paper [3] where the problem of “mining” a large collection of basket data type transactions was introduced as well. The paper presented an efficient algorithm called AIS (Agrawal, Imielinski, Swami) for association rules between sets of items. They were interested in finding all rules that have a Minimum transactional support $s$ and a Minimum confidence $c$.

In AIS algorithm, only one item consequent association rules is generated, which means that the consequent of those rules only contain one item. For example, we only generate rules like $X \cap Y \rightarrow Z$ but not $Y \cap Z$. Due to the fact that the databases are scanned many times to get frequent itemsets in AIS, an estimation method was introduced to make this algorithm more efficient. Thus, they pruned those itemsets candidates that had no hope to be large, consequently the unnecessary effort of counting those itemsets can be avoided. Since all candidate itemsets and frequent itemsets are assumed to be stored in main memory, memory management is also proposed for AIS when it is not enough. A main drawback of the AIS algorithm is that too many candidate itemsets are generated that turn out to be very small, which requires more space and wastes much effort that turns out to be unnecessary, while at the same time needs too many passes over the whole database even to be that effective.

Thus, Agrawal in paper [64] introduced the Apriori algorithm which was a great improvement in the history of association rules [63]. Apriori uses pruning techniques to avoid measuring certain itemsets, while guaranteeing completeness. These are itemsets that the algorithm can prove will not turn out to be large. Therefore, Apriori algorithm is more efficient than AIS during candidate generation process. It reduces computation, I/O cost and memory requirement because of the new pruning technique [65]. However, there are two bottlenecks of the Apriori algorithm. One is the complex candidate generation process that uses most of the time, space and memory. The other is the multiple scans of the database. Figure 5 describes the Apriori algorithm.

![Figure 5: Apriori Algorithm [14]](image)

As was previously mentioned, Apriori employs a different candidate generation method and a new pruning technique. In Apriori, there are two processes to find out all the large itemsets from the database. Candidate itemsets are generated first, then database is scanned to check actual support count of the corresponding itemsets. In the first scanning, support count is calculated and large 1-itemsets are generated by pruning the itemsets falls below predefined threshold. Processes are executed iteratively until candidate / frequent itemsets empty. Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules [65].
Still, Apriori algorithm was being held back by multiple whole database scans. Therefore, many new algorithms were designed with several modifications or improvements. Based on [63] [14] [64], computational cost of Apriori can be reduced in four ways:
1. By reducing the number of passes over the database
2. By sampling the database
3. By adding extra constraints on structure of patterns
4. through parallelization.

In recent years, much progress has been made in all these directions. First improvements after Apriori, introduced by [65], were Apriori-TID and Apriori-Hybrid. Such algorithms are based on Apriori and try to improve efficiency by making several modifications, such as reducing the number of passes over the database; reducing the size of the database to be scanned in every pass; pruning the candidates by different techniques and using sampling technique.

Another approach in solving the first bottleneck of Apriori algorithm, which is the number of passes over the database, was introduced by the FP-Tree (Frequent Pattern Tree) [66]. It is a tree structure pattern mining algorithm generating frequent itemsets by scanning the database only twice without any iteration process for candidate generation. The first is FP-Tree construction process and the second is generation of frequent patterns from the FP-Tree through a procedure called FP-growth. FP-Tree scales much better than Apriori because, as the support threshold goes down, the number, as well as the length of frequent itemsets, increases dramatically. Candidate sets that Apriori must handle become extremely large, and pattern matching with a lot of candidates by searching through the transactions becomes very expensive. Frequent patterns generation process includes two sub-processes: constructing the FT-Tree, and generating frequent patterns from the FP-Tree. FP-Tree may be much faster than Apriori but it is difficult to use in an interactive mining system and not suitable for incremental mining [65]. Mining results of FP-Tree are the same as Apriori’s series algorithms [63]. A recent study [69] introduced the Rapid Association Rule Mining (RARM), which is based on the tree structure like FP-tree but it is much faster.

Sampling was used by [66] for association rule mining. The approach can be divided into two phases: During phase 1 a database sample is obtained and all associations in sample are found. Results are then validated against the entire database. To maximize effectiveness of the overall approach, the author makes use of lowered minimum support on the sample. Since the approach is probabilistic (i.e. dependent on the sample containing all relevant associations), not all the rules may be found in this first pass. Associations that were deemed not frequent in sample but were actually frequent in the entire dataset are used to construct the complete set of associations in phase 2. This technique is suggested if data comes as a stream flowing at a faster rate than can be processed [63].

Parallelization techniques have be also been applied in order to take advantage of higher speeds and greater capacity that parallel systems offer [63]. Transition to a distributed memory system requires partitioning of the database among the processors which is a procedure generally carried out indiscriminately. The most known algorithm using this technique is FDM algorithm [68]. FDM is a parallelization of Apriori to shared nothing machines, each with its own partition of the database. At every level and on each machine, database scan is performed independently on local partition. Then, a distributed pruning technique is employed.

To conclude, Apriori algorithm seemed to be a breakthrough in Association rules data mining domain, but had drawbacks when the database is large in terms of computational cost. Therefore, further studies
based on Apriori basics have reduced its computational cost. On the other hand, other algorithms were introduced that were not direct descendants of Apriori, but aimed to resolve its bottlenecks. The most known is FP-Growth, which scales much better than Apriori, scanning the database only twice. Thus, computational cost of FP-Growth is less than Apriori’s, while they produce the same results [63].

In this thesis, time execution and computation cost are irrelevant. Therefore, Apriori Algorithm will be applied in our case study.

2.2.3 Drawbacks and solutions of applying Association rules mining

A drawback of association rules mining is the need to be configured before executed. So, users have to give appropriate values for the parameters in advance (often leading to too many or too few rules) in order to obtain a good number of rules [70]. Most of these algorithms require users to set two thresholds: minimal support and minimal confidence, and find all rules exceeding thresholds specified by users. Therefore, users must possess a certain amount of expertise in order to find the right settings for support and confidence to obtain the best rules. However, the most serious problem in association rule discovery is that the set of association rules can grow to be unwieldy as the number of transactions increases, especially if support and confidence thresholds are small. As the number of frequent itemsets increases, the number of rules presented to users typically increases proportionately. Therefore, it is difficult for users to identify those rules that are of interest among the large number of rules that are produced. Many of these rules may be redundant, irrelevant or uninteresting. Therefore, several techniques have been identified, aiming to reduce the number of the association rules [14] [63] [70].

Such techniques are presented as post-processing tasks aiming to improve selection of discovered rules. Among the post-processing methods introduced, several phases are presented, including pruning, summarizing, grouping or visualization. The pruning phase, consists of removing uninteresting or redundant rules. In summarizing phase, summaries of rules are generated. Groups of rules are produced in the grouping phase while visualization phase is useful to have a better presentation [30]. In the Section (2.2) post-processing the techniques are presented and analyzed.

2.3. Post-Processing methods in association rules mining

Given the research problem and research questions, the resulted rules from the association rule mining are not enough to provide an overview of the account pairs neither the extraction of the unusual one. Especially, considering the fact that association rule mining leads to sets of very large number of rules, hard to comprehend. Therefore, a postprocessing step was introduced, and provided a basic rationale for postprocessing the patterns generated, by an association rule mining process. In this section, literature analysis for post processing techniques for the association rule mining algorithms is presented. Post processing refers to the pieces of knowledge extracted in the previous step that are further processed. Postprocessing procedures thus provide a kind of “symbolic filter” for noisy, imprecise or “non-user-friendly” knowledge derived by an inductive algorithm [71].

Based on the book [14] [30], the postprocessing methods in association rules are categorized in:

- Redundancy
- Interestingness measures
- Summarization and grouping the rules in order to get a set more compact and generalized ones
Visualization

In the following subsections, a literature study is conducted for each postprocessing category. At the end of this section, postprocessing methods specifically for this research problem are selected.

2.3.1. Redundancy

In association rules algorithms, when resultant frequent itemsets are large, the former produce large number of rules. However, lots of the rules have identical meaning or are redundant. In fact, the number of redundant rules is much larger than expected. In most cases, redundant rules are significantly more than essential rules [9]. Another research [4] suggests that the removal of redundant rules can affect the quality of the information presented. Therefore, numerous frameworks have been proposed, most of which have several prior assumptions. Based on these assumptions, frameworks identify redundant rules and prune them subsequently [9]. Research [70] for example, suggests for the identical meaning rules to indicate the attributes that must or must not be present in the antecedent or consequent of the discovered rules as a solution.

Paper [8] marked rule r as redundant and eliminated it, in the presence of another rule R (consider r, \( R \in R' \), where \( R' \) is resultant ruleset) without considering whether rule R characterizes the knowledge of rule r. For example, the algorithm will mark rule \( AB \rightarrow C \) as redundant in the presence of rule \( A \rightarrow C \). However, it is apparent from this example that rule \( A \rightarrow C \) is not fully characterized the knowledge of rule \( AB \rightarrow C \).

Therefore, paper [9] proposed a method that removes redundant rules from the resultant ruleset without losing any important knowledge. The proposed methods mark a rule as redundant when it finds a set of rules that also convey the same knowledge. For example, the proposed method will mark rule \( A \rightarrow BC \) as redundant, if and only if rules such as \( A \rightarrow B \) and \( A \rightarrow C \) are present in that set. Research [9] proposed two methods to eliminate redundant rules: removing redundant rules with fixed antecedent rules, and with fixed consequent rules. To find redundant rules with fixed antecedent rules they proposed the following theorem:

**Theorem 1 [9]:** Consider rule \( A \rightarrow B \) satisfying the minimum confidence threshold such that antecedent A has i items and consequent B has j items where \( i \geq 1 \) and \( j > 1 \). The rule \( A \rightarrow B \) is said to be redundant if and only if n number of rules such as \( A \rightarrow e_1 \), \( A \rightarrow e_2 \), ... \( A \rightarrow e_n \) satisfy minimum confidence threshold where \( \forall e \subseteq B \) and \( n=j \).

**Example:** Let us apply this theorem to a ruleset R that has three rules such as \{\( ABX \), \( AB \rightarrow Y \) and \( AB \rightarrow XY \)\}. Consider the rule \( AB \rightarrow XY \) has s% support and c% confidence. Then, the rules such as \( AB \rightarrow X \) and \( AB \rightarrow Y \) will also have at least s% support and c% confidence because \( X \subseteq XY \) and \( Y \subseteq XY \). Since \( AB \rightarrow X \) and \( AB \rightarrow Y \) dominate \( AB \rightarrow XY \) both in support and confidence, for this reason \( AB \rightarrow XY \) is redundant.

To remove redundant rules with fixed Consequent rules follows theorem 2.

**Theorem 2 [9]:** Consider rule \( A \rightarrow B \) satisfying the minimum confidence threshold such that antecedent A has i items and consequent B has j items where \( i \geq 1 \) and \( j > 1 \). The rule \( A \rightarrow B \) is said to be redundant if and only if n number of rules such as \( e_1 \rightarrow B \), \( e_2 \rightarrow B \), ..., \( e_n \rightarrow B \) satisfy minimum confidence threshold where \( \forall e \subseteq A \) and each e has \( (i-1) \) items.

**Example:** Let us apply this theorem to a rule set R that has three rules, such as \{\( XY \rightarrow Z \), \( X \rightarrow Z \) and \( Y \rightarrow Z \)\}. Suppose rule \( XY \rightarrow Z \) has s% support and c% confidence. If n (i.e. number of items in the antecedent)
number of rules such as $X \rightarrow Z$ and $Y \rightarrow Z$ also satisfy $s$ and $c$ then, the rule $XY \rightarrow Z$ is redundant because it does not convey any extra information if rule $X \rightarrow Z$ and $Y \rightarrow Z$ are present.

Finally, paper [10] conducted a survey for rule pruning in association to rule mining for removing redundancy and defined the redundant association rules as:

Let $X \rightarrow Y$ and $X' \rightarrow Y'$ be two rules with confidence $cf$ and $cf'$, respectively. $X \rightarrow Y$ is said a redundant rule to $X' \rightarrow Y'$ if $X$ belong to $X'$; $Y'$ belong to $Y$, and $cf \leq cf'$.

This survey also investigated all the papers having some type of redundancy removed but not getting the interesting rule with non-redundant set. He concluded that there is need to generate an interesting rule set, comprehensible by end users, while providing it with a good decision.

2.3.2. Interestingness measures

As there is a huge number of association rules or patterns that are generated by most Association Rule Mining Algorithms, a need arises to prune away unnecessary and unwanted rules. Rules that are crucial and indispensable can therefore be presented to end users based on application of these “Interestingness Measures”. This is so that users can focus on rules that will provide better business understanding and intelligence. Lots of rules produced take time to process, and prove to be a hindrance to efficient processing. Space they occupy also becomes enormous [32]. However, there is a plethora of measures available today, and selecting the best amongst them requires a thorough research on each. The term “Interestingness Measure” unequivocally forms a very essential aspect of extraction of “interesting” rules from databases.

Study literature conducted for interestingness measures [18] stated that there is no agreement on formal definition of “interestingness”; this notion is best summarized as “record or pattern is interesting if it suggests a change in an established model.” This multidisciplinary concept portrays interestingness as an entity that captures the impression of "novel" or "surprising". However, further studies added more definitions to the term “interesting”. These are conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility, and actionability. A brief description for each definition is presented in Appendix B3.

Interestingness measures are distinguished in two categories: objective indicators and subjective indicators. Objective (data-driven) indicators do not take any background information into account and are therefore suitable if an unbiased ranking is required. Examples of such measures are lift, conviction, odds ratio and information gain [46]. These measure are also called Objective Measures Based on Probability. Probability-based objective measures are usually functions of a $2 \times 2$ contingency table. A contingency table stores the frequency counts that satisfy given conditions. Table 1 is a contingency table for rule $A \rightarrow B$, where $n (AB)$ denotes the number of records satisfying both $A$ and $B$, and $N$ denotes the total number of records [75].
In the table, $A$ and $B$ represent the antecedent and consequent of a rule, respectively, while $N$ denotes the probability of $A$; while $P(A)$ denotes the conditional probability of $B$, given $A$. These measures originate from various areas, such as statistics (correlation coefficient, odds ratio, Yule’s Q, and Yule’s Y), information theory ($J$-measure and mutual information), and information retrieval (accuracy and sensitivity/recall).

On the other hand, for subjective measures, access to users’ domain or background knowledge about the data is required. Access can be obtained by interacting with users during the data mining process or by explicitly representing users’ knowledge or expectations [75]. In the subjective, case users directly apply their own knowledge or beliefs, sometimes even without knowing the nature of the domain to find interesting rules. From that perspective, subjective interestingness may be unreliable and vary among users. This approach may be suitable to justify a particular users’ own belief, but may fail to discover some surprising rules that the users even cannot think of. One potential problem is that users’ subjective judgment may be unrealistic while applying those rules in the competitive business environment, comparing to objective measures where no prior knowledge is applied to mine interesting rules. Therefore, objective type of measure is likely to be reliable since no user’s biased preference is given while estimating the interestingness. However, one can still generate a large number of objectively interesting rules, but of little interest to another user [72]. To conclude, research [72] suggests that the Objective approach would be superior in the sense that it helps mining those unexpected rules that users cannot even think of. In the next section, objectives interestingness measures are described in more detail.

**Objective interestingness measures**

The first objective interesting measure introduced was conducted by the research [74], where statistical independence of rules was proposed as an interestingness measure. More methods have since been proposed, using different statistical approach. For example, research conducted by [50] developed a technique that discovered those rules that contradict a set of rules with strong support and confidence and are exceptions to existing knowledge, and therefore interesting. The authors are aware that exception rules typically have low levels of support and are computationally expensive to extract from large databases. However, exception rules often have strong levels of confidence similar to the commonsense rules.

Later on, another research for interesting subsets of rules was conducted by [73] in which the authors developed an optimized algorithm that was searching for the most interesting rules by integrating a variety of measures such as lift, support, confidence, entropy, Laplace values, Gini and chi-square measures. The rule discovery process was interactive, allowing users to interrogate and compare the set of rules against the several measures. Partial ordering of rules allows the characteristics of specific subsets to be revealed, characteristics that would have been missed using support and confidence levels alone.
In 2002, paper [33] conducted an extended study in objective interestingness measures introduced until that date. This Paper proposed the use of interestingness measures to select the most relevant rules, in addition to the support measure, eliminating uncorrelated and poorly correlated patterns. It concluded that there is no measure better than others in all application domains and that some indicators may be desirable for certain applications, but not for others. As an alternative to this problem, the paper suggested to match the desired properties of an application against properties of the existing measures by referring to domain experts and focusing on the most appropriate measure by comparing how well each measure agrees with their expectations.

Research [14] [75] states that objective measures rely on a user’s ability to select the right measure for a given scenario out of a huge set of available ones. They stated that some measures produce similar rankings while others almost reverse the order. This poses the problem of selecting the right measure for a given scenario. For a user, it is often unclear which measure to choose and how to link its results to his application scenario. Consequently, lots of rules deemed interesting will not be very useful. Therefore, selection strategies for probability-based objective measures were created due to the overwhelming number of interestingness measures.

Initially, paper [46] proposed a method to rank measures based on a specific dataset. In this method, users are first required to rank a set of mined patterns, and the measure that has the most similar ranking results for these patterns is selected for further use.

Another strategy was introduced in paper [34], which is based on using the MCDA method (Multiple Criteria Decision Aid) on some classical measures to help select a measure which is concordant with users’ objectives. This paper demonstrated that the MCDA method could prove helpful in selecting an appropriate interestingness measure. Focusing on relevance of the rules requires following the interestingness measure principle, defining a suitable metric to capture the dependencies among variables in a data set. Nonetheless, several such measures provide conflicting information about the interestingness of a pattern and the best metric to use.

In order to reduce the plethora of measurement choices, research [36] built a classification of measures based on correlation value between measures. They have noticed that measures consider two particular situations interesting. They consider it more interesting when either L or R is close to 0. This means that a rule is interesting when the antecedent almost always appears with the consequent (but the consequent does not have to always appear with the antecedent) or when the consequent almost always appears with the antecedent. They identified seven types. Figure 6 shows the classification of the objective measures [36].
After that, they selected one measure from each type and calculated correlation of the results produced by each measure given the combination L, R, and T in the range from 0 to 1 with increment of 0.01. Research [36] noted that most of the measures have a low correlation value (since they also are of different types), but certain measures were noticed to have slightly higher correlation between them. This means that they rank the rules in a similar fashion.

Further search in this area conducted comparisons between objective measurements and subjective measurements [15] [16]. The main reasons for that is the diversity of formal criteria and the fact that no measure wins in all criteria. In their research, authors evaluate different measures by means of expert interest. This is an important approach for pattern evaluation as it directly measures relation between a formal measure and an expert interest. Drawbacks of this approach, are the cost and diversity of datasets: for every dataset it is necessary to hire several experts, which is costly [17]. That is one reason why in later research [14], they suggested a framework for mining the interesting association rules based on both indicators. Figure 7 shows a detailed design for each step in order to get the interesting rules.

Unfortunately, asking experts to rank all tables manually is often impractical. A more practical approach is to provide experts with a smaller set of contingency tables for ranking and use this information to determine the most appropriate measure. In the empirical study [46], they used a methodology where first identified a small subset of contingency tables. The criteria where:
1. Subset must be small enough to allow domain experts to rank them manually. On the other hand, the subset must be large enough to ensure that choosing the best measure from the subset is almost equivalent to choosing the best measure when rankings for all contingency tables are available.

2. Subset must be diverse enough to capture as much conflict of rankings as possible among different measures.

The first criterion is usually determined by experts because they are the ones to decide the number of tables they are willing to rank. Therefore, the only criterion we could optimize algorithmically is diversity of the subset.

Another aspect that users have to take into consideration is which of the rules to select after an objective measurement has been applied. The objective interestingness measures described before give a value to the rules, which describes if the antecedent and consequent are positively correlated, negatively correlated or independent to each other. A former definition has been given in paper [78], which refers to positive association rules as being rules of the following type: given two items $X$ and $Y$, a positive association rule is a rule of the form $X \rightarrow Y$ ($X$ and $Y$ exist together frequently and $X \cap Y = \emptyset$). A negative association rule is one of the following: $\neg X \rightarrow Y$ or $X \rightarrow \neg Y$ (where $X$ means existence and $\neg X$ means absence in enough transactions). In simple words, positive association rule is 'If A, then B', whereas negative association rule technique generates the rule, 'If A, then not B.', or 'If not A, then B.' In general, an association rule is describes as:

- $P (S \land B) = P (S) \times P (B)$ => Statistical independence
- $P (S \land B) > P (S) \times P (B)$ => Positively correlated
- $P (S \land B) < P (S) \times P (B)$ => Negatively correlated

2.3.3. Summarization and Grouping

There are several traditional techniques used in order to improve the comprehensibility of discovered rules. For example, we can reduce the size of the rules by constraining the number of items in the antecedent or consequent of the rule. Simplicity of the rule is related with its size, and as such, the shorter the rule is, the more comprehensible it will be [70].

2.3.4. Visualization

The simplest way to represent a small number of association rules are textual descriptions, which can be examined with all the low level details such as the items contained in the LHS and RHS. However, since users must evaluate the rules in a sequential manner, they are not conducive to the analysis of complex data and large collections of association rules.

Ability of information visualization enabling users to gain an understanding of high dimensional and large-scale data can play a major role in the exploration, identification and interpretation of association rules. However, mining association rules often results in a very large number of found rules, leaving the analyst with the task to go through all the rules and discover interesting ones. Shifting manually through large sets of rules is time consuming and strenuous. Visualization has a long history of making large amounts of data better accessible using techniques like selecting and zooming. [39]

Visualization of association rules can provide immediate insight into the primary characteristics of set of rules. At least five parameters should be considered in the visualization of an AR: the set of the antecedent items, consequent items, association between antecedent and consequents, the rule’s support, and its confidence. Research on visualization of Association rules can be categorized into three main groups, depending on whether they are based on tables, matrices, or graphs.
Research [40] initially introduced a matrix-based visualization of the rules. Matrix-based visualization is limited in the number of rules it can visualize effectively, since large sets of rules typically also have large sets of unique antecedents/consequents. Therefore, they enhanced the matrix-based visualization using grouping of rules via clustering to handle a larger number of rules. Alternative research [41], visualized association rules using vertices and edges, where vertices typically represented items or itemsets and edges indicated relationship in rules. Interest measures are typically added to the plot as labels on the edges or by color or width of the arrows displaying the edges. Graph-based visualization offers a very clear representation of rules but they tend to easily become cluttered and thus are only viable for very small sets of rules.

Another visualization method for association rules mining was introduced by research [44] using Parallel Coordinates Plots. Parallel coordinates plots were designed to visualize multidimensional data where each dimension is displayed separately on the x-axis and the y-axis is shared. Each data point is represented by a line connecting the values for each dimension. Parallel coordinates plots were previously used to visualize discovered classification rules [43] and association rules [44]. In research [44], they displayed the items on the y-axis as nominal values and the x-axis represents the positions in a rule, i.e., first item, second item, etc.

Despite the advantages of previous works in visualizing association rules, the most common problem they encounter is their inability to handle a large collection of rules. In general, this results in occlusion and screen clutter problems due to the need to compress the whole visual representation into a single view. In other words, by presenting a large number of rules over many items in a single view, it is not easy for users to recognize relations between items and their interestingness measures. An alternate approach is to display different characteristics of the rules simultaneously in different views. However, the fundamental trade-off is that it is not possible for users to perceive and compare all of these characteristics at once, requiring them to switch between different views in order to see different features of a specific rule [39].

2.4. Conclusion

This chapter provided a literature analysis about Fraud in Accounting, Association Rule Mining and Post-Processing Method of Association Rules. The analysis has been conducted in order to reach our research goal and find all options available.

Considering the literature conducted for Fraud in Accounting, financial statement fraud seems to be difficult to identify with current tools. Data mining could be a technique enabling auditors to better trace irregularities. Particularly, regarding the identification of the unusual account relocations which, as it was previously mentioned, is the most known financial statement fraud. Therefore, coordinated with the research problem, which aims to provide methods that will support domain expert to find unusual financial transactions in terms of unusual account pairs, we consider to assist domain experts in Financial statement fraud category, where inappropriate account relocations is the main problem.

Based on literature of Association rules, Apriori seems to be the most known algorithm with the greatest improvement in the history of association rules. One of the reasons that other algorithms were developed after Apriori is its computational cost scanning big dataset. However, in this thesis, we did not encounter storage and memory problems, therefore in the modeling phase Apriori was used.

In the post-processing step, several methods were introduced in order to reduce the large number of rules and find the most interesting ones among them. Among the definitions of the term “interesting” is also the unexpected/unusual behavior from some rules. In this thesis, we are aiming to support auditors when
reviewing the journal entries by providing them with a holistic view of the dataset and also the unusual rules to be found in the dataset.

The methods that are proposed were redundant reduction, interestingness measurements, visualization and summarizing. First, it is proposed to remove the redundant rules not providing extra information to domain experts. Another method is the visualization of the rules that have been generated. For the visualization method, several techniques have been proposed from literature for the visualization part, giving domain experts the opportunity to investigate a larger set of rules, compared to the usual text approach, which is more suitable for few rules.

Another post-processing method is applying interestingness measures. A large part of research conducted for association rules is focused on the interestingness measures applicable on the rules in order to extract the most interesting ones. These measures are divided into objective (probability-based) rules, which are generated from the data without previous knowledge, and subjective rules, which are determined from the domain knowledge depending on what the auditor is looking for. In this thesis, we are aiming to find interesting rules generated from data without directions from the auditors.

A drawback of the interestingness measures is that there seems to be no perfect measure among them. The selection of the measure that is going to be applied is based on the domain problem. One method proposed from research [46] was to provide experts with a smaller set of rules for ranking and use this information to determine the most appropriate measure. Unfortunately, all selecting criteria need domain experts’ interpretation, which is a time-consuming task, inapplicable in the real world and inappropriate for this thesis. In addition to the mentioned drawbacks, evaluation of the resulted rules can also not be determined by Recall or Precision which counts the ration of the true positive elements to the total number of true elements. These methods, require prior knowledge about which journal entries are fraudulent/suspicious, for example a labelled dataset. In our case, there a fraudulent or suspicious dataset is not available.

Research [36] tried to clear the scope of the measures by classifying all measures into categories, according to similarity in results given. In this thesis, although we do not aim to find which measure is the best suited for our research problem, we are going to apply a measure per measure category in order to capture all differences and same results. This approach was also suggested in paper [36]. Then, based on the foundation that the selection of a measure is based on the specific problem, the most suitable measure for our unusual definition is selected.

Therefore, in the interestingness measures section, the following measures are selected in this thesis: f-coefficient, Lift, Conviction, Cosine, certainty. These measures are described in more detail in the related section.

Finally, further research [78] conducted regarding association rules, studied the degree of correlation between the antecedent and the consequent. Positively correlated rules have been naturally selected as the “interesting” ones but later studies added that negatively associated rules might also be “interesting”, depending on the topic of the research. In this thesis, we are not aiming to discover rules that follow the rule ‘If A, then not B.’, or ‘If not A, then B.’; therefore negative association rules and independent association rules are excluded from the results. In the respective section, a more detailed analysis about the unusual rules selection is explained.
3. DATA UNDERSTANDING, SELECTION AND TRANSFORMATION

In the previous chapters, the research goal of this thesis is defined and, in addition, an understanding of the application domain and relevant prior knowledge is developed. Before determining the data that will be selected and used for the knowledge discovery phase, a data understanding step is first introduced. This process is an important element of data mining in developing knowledge of the properties of the subject of investigation. Therefore, in this chapter, the data included in a general ledger is initially described in detail. Afterwards, based on the research goal, two steps are executed, the data selection and the creation of a dataset on which discovery is performed. The two steps include the selection of the attributes that will be considered as important for the research goal and the creation of the dataset step includes the data transformation needed that would be the base for constructing the model. In this case, the model is the Association rules produced by the dataset. As mentioned in paper [36], these steps are crucial for the success of the entire project and are project-specific. The following sections provide the results of described steps for our case study. Finally, this chapter intends to provide an answer to the research question A2.

3.1. Data Understanding

A simple data model of two journal entries example is given in Table 2. The execution of financially relevant transactions in an ERP (Enterprise Resource System) creates specific data records. Every execution of a financial transaction creates a posting document. A posting document is a data record that represents a journal entry. Each document contains at least two journal entry items including the accounts that have been involved in the transactions. In the example in Table 2, the journal entry with the Document ID 1 has the simplest format (two lines). The journal entry with the Document ID 2 is slightly more complex. The accounts are represented by an Account ID and an Account Description. Besides these attributes, there might be multiple other metadata including Country, Money Currency, Posted Date, User, Value Category, Business Unit, Transaction Code, etc. We have to mention that the attributes that a journal entry contains vary from one company to another.

<table>
<thead>
<tr>
<th>DocumentID</th>
<th>Country</th>
<th>Account ID</th>
<th>Account Text</th>
<th>PL</th>
<th>BS</th>
<th>User</th>
<th>Value</th>
<th>Credit/Debit</th>
<th>Trans. Code</th>
<th>Post Date</th>
<th>Currency</th>
<th>Other metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>xxx</td>
<td>XXXYYYY</td>
<td></td>
<td>1</td>
<td>0</td>
<td>User 1</td>
<td>500</td>
<td>D</td>
<td>xxx</td>
<td>Dd/mm/yyyy</td>
<td>$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>xxx</td>
<td>XXXYYYY</td>
<td></td>
<td>0</td>
<td>1</td>
<td>User 1</td>
<td>-500</td>
<td>C</td>
<td>xxx</td>
<td>Dd/mm/yyyy</td>
<td>$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>xxx</td>
<td>XXXYYYY</td>
<td></td>
<td>0</td>
<td>1</td>
<td>User 2</td>
<td>500</td>
<td>D</td>
<td>xxx</td>
<td>Dd/mm/yyyy</td>
<td>$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>xxx</td>
<td>XXXYYYY</td>
<td></td>
<td>1</td>
<td>0</td>
<td>User 2</td>
<td>200</td>
<td>D</td>
<td>xxx</td>
<td>Dd/mm/yyyy</td>
<td>$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>xxx</td>
<td>XXXYYYY</td>
<td></td>
<td>1</td>
<td>0</td>
<td>User 1</td>
<td>-300</td>
<td>C</td>
<td>xxx</td>
<td>Dd/mm/yyyy</td>
<td>$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>xxx</td>
<td>XXXYYYY</td>
<td></td>
<td>1</td>
<td>0</td>
<td>User 3</td>
<td>-300</td>
<td>C</td>
<td>xxx</td>
<td>Dd/mm/yyyy</td>
<td>$</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Manual journal entries in the ledger

Information about the dataset provided for this thesis is given in Table 3. The database contains 370,721 lines representing the 150,639 financial transactions (journal entries) that took place. In addition, 76 columns represent the variables (metadata) which describe these transactions. In Appendix A1, the available metadata from the dataset are described in detail.
3.2. Data selection
The definition of unusual behavior of financial transactions in this research has been defined in terms of account combinations and amount of money. The initial task for this purpose is to select the most appropriate attributes from our dataset and then construct a transactions database in order to apply the Apriori algorithm selected in the previous chapter.

3.2.1. Selection of attributes:

1. Account ID: A general ledger contains all different accounts ID that were recorded. These account IDs are related to a company's assets, liabilities, owners' equity, revenue, and expenses. Account ID is represented in our case study from 7 digit figure. The first three digits from the left represent the class of the accounts in terms of revenue, expenses, assets, equity, liability, income. And the other 4 digits represent more detailed information about the accounts. In Appendix A3, there is an example of an Accounting tree [11]. In our case study, we have 287 different Account IDs.

2. Journal Entry (Document ID): This attribute represents the unique journal entry that is recorded in the database. As it was mentioned before, each journal entry is represented uniquely by a number and includes minimum two item lines with accounts that have credit or debit values.

3. Value Category: this attribute contains eight already defined categories in the given dataset with a certain range of money that the journal entry has posted. The Value categories are:

<table>
<thead>
<tr>
<th>Category</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-100</td>
<td>0-100</td>
</tr>
<tr>
<td>101-1000</td>
<td>101-1000</td>
</tr>
<tr>
<td>1001-10.000</td>
<td>1001-10.000</td>
</tr>
<tr>
<td>10.001-50.000</td>
<td>10.001-50.000</td>
</tr>
<tr>
<td>50.001-100.000</td>
<td>50.001-100.000</td>
</tr>
<tr>
<td>100.001-1.000.000</td>
<td>100.001-1.000.000</td>
</tr>
<tr>
<td>1.000.001-10.000.000</td>
<td>1.000.001-10.000.000</td>
</tr>
<tr>
<td>&gt;10 millions</td>
<td>&gt;10 millions</td>
</tr>
</tbody>
</table>

4. Credit/Debit: this attribute shows whether the amount of money has been credited or debited.

Someone could argue to choose the “Value” attribute which contains the exact value that the account posted (for example a continuous data such as 123 euro). However, this is one of the limitations in association rules, continuous data need to be scaled or labeled into categories before applying association rules. Since we already have the Value category in our initial dataset, and the categorization of the amounts are already defined by experts, we did not choose to categorize otherwise the value attribute, but rather select the attribute “Value Category”.

3.3. Transformation to transaction database
In order to apply the Apriori algorithm, we had to construct a transactional database. A transactional database contains a set of transactions and each one of these transactions has a list of items that refers to the activities that took place. In our case, a transaction is a journal entry and each transaction itemset contains the accounts posted in that journal entry. In addition to the accounts, the Value category of the journal entry is included as an item in the itemset. For example, if one journal entry posted two Account
IDs with value -50 euro (Credit) and 50 euro (Debit), then this journal entry is included in the Value Category “0-100”.

First, considering the large number of different Account IDs, a Chart of Accounts has been created. The Chart of Accounts transformation is also recommended and applied in the literature [20] [21] [22]. The author of these papers developed an internal taxonomy to represent a generic chart of accounts. Typically, a chart of accounts is a list of accounts that is created to define each class of items. It aggregates information into an entity’s financial statements. The chart is usually sorted in order by account ID, to ease the task of locating specific accounts.

This transformation was motivated by the fact that the amount of different Account IDs (287 in total) will create 287 different Boolean attributes. The Apriori Algorithm, then, might produce many rules as infrequent that, in a higher level, might be related with each other. Thus, many rules might “flagged” for further inspections, but in a sense, they are doing the same thing in a aggregated way. An example of Account IDs mapped to one Account Category is the following: the account ID “Trade receivables- Local Customer” and the Account ID “Trade Receivables-Payment Card”, are mapped to the Account Category “Trade receivables”.

Fortunately, for the case study in this research, a Chart of Account was already structured in a different dataset. It was structured in seven levels of description. Each Level described the Account ID starting from a generic category, such as asset, liability, expense, revenue (Level 1), to the most detailed description (Level 7).

In this thesis, we mapped the Account IDs in Level 3 of description, since also in KPMG, this level is used for analysis purposes and an additional reason is that not all Account IDs have in Level 4,5,6,7.

Therefore, we mapped each Account ID to its Level 3 level of Description and created an additional column in our original dataset (the original dataset did not include the Account Category attribute, the Level of Description of the Accounts IDs were stored in another dataset in KPMG, as it was mentioned previously). This required SQL Join clauses to combine records from the two datasets. At the end, our dataset consisted of 41 distinct Account Categories. Appendix A2 enumerates each one of them with a brief description of what they represent. In order to understand each aggregated Account Categories, a hierarchical tree of the Account Categories, based on the Level of description dataset, was constructed and is described in Appendix A4.

After this step, we combined attribute Account Category with attribute Credit/Debit. The reason behind this step is that there might be account categories that normally credit/debit money, thus finding the opposite behavior might indicate that a journal entry is unusual and needs further investigation. In addition, with this kind of transformation, in the visualization step, auditors could observe the flow of the money between the account categories.

With the tool R, the transformed database can be read as a transaction database and creates from that a transactions object. Specifically, it creates a “basket” format, each line in the transaction data file represents a transaction where the items (item labels) are separated. The collection of itemsets then are represented as binary incidence matrices with columns corresponding to the items (Account categories (Credit/Debit), Value Category) and rows corresponding to each journal entry of our dataset. The tool R provides the infrastructure for representing, manipulating and analyzing transaction data and patterns [38]. As it was mentioned in the Chapter 1, we selected this tool among the other data mining
tools (Rapidminer, Weka etc), because it includes, in the Association rules mining, packages with more capabilities in term of analyzing the transaction database and rules compared to other mentioned tools. In addition, we were more familiar with this tool and in terms of time, this is beneficial. Table 4 shows a small subset of the transaction database created by the dataset. Each line represents a transaction, in our case a journal entry, and inside the brackets the set of items of each transaction is displayed, in our case the account categories credited or debited in the journal entry and journal entry Value Category.

<table>
<thead>
<tr>
<th>Transaction Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.001-10.000</td>
</tr>
<tr>
<td>Total_liquid_assets_Credit</td>
</tr>
<tr>
<td>0-100</td>
</tr>
<tr>
<td>Total_liquid_assets_Debit</td>
</tr>
<tr>
<td>0-100</td>
</tr>
<tr>
<td>Total_liquid_assets_Debit</td>
</tr>
<tr>
<td>101-1.000</td>
</tr>
<tr>
<td>Bad_debts_Credit</td>
</tr>
<tr>
<td>0-100</td>
</tr>
<tr>
<td>Total_liquid_assets_Debit</td>
</tr>
<tr>
<td>101-1.000</td>
</tr>
<tr>
<td>COGS(own manufactured)_Credit</td>
</tr>
<tr>
<td>0-100</td>
</tr>
<tr>
<td>Total_receivables_Credit</td>
</tr>
<tr>
<td>10.001-50.000</td>
</tr>
<tr>
<td>Total_receivables_Credit</td>
</tr>
<tr>
<td>0-100</td>
</tr>
<tr>
<td>Total_receivables_Credit</td>
</tr>
<tr>
<td>0-100</td>
</tr>
<tr>
<td>COGS(own manufactured)_Credit</td>
</tr>
</tbody>
</table>

Table 4: Transaction Database of MJE example of 10 Journal Entries

One of the limitations in applying association rules algorithm in journal entries is that each account could not be stated multiple times in the transaction database, if it was executed many times in one journal entry in the real ledger. This might lead to false interpretation of rules, because it might result that a pair of accounts is infrequent, whereas they might be actually executed many times in one journal entry.

In this phase, we did not require to transform our data further in means of data clearing. The dataset that we analyzed did not include any missing values in any selected attribute. The transactional database that was created includes:

<table>
<thead>
<tr>
<th>MJE transaction database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transaction lines</td>
</tr>
<tr>
<td>Account Categories</td>
</tr>
<tr>
<td>Account Categories with Credit Debit</td>
</tr>
<tr>
<td>Attributes representing Account Categories with Credit Debit and Money Category in total</td>
</tr>
<tr>
<td>Maximum length of Itemset</td>
</tr>
<tr>
<td>Minimum length of Itemset</td>
</tr>
</tbody>
</table>

Table 5: Transaction Database

The transaction lines represent the total number of journal entries in the ledger. In total, 41 Account Categories were found, however, not all of the Account Categories were at some point credited or debited in journal entries. That is the reason why 78 distinct attributes where found in total in the created transaction database. In addition to the 78 attributes which represented the Account Categories (Credit/Debit), 8 distinct Value Categories in the transaction database were included. In total, 86 different attributes were counted in the transaction database.
An overview of the frequency of each item in our transactional dataset is depicted in the chart in Figure 8.

Figure 8 shows that the majority of the journal entries that have been posted involves values “0-100”. The higher the Value Category the less frequent it occurs in the dataset. For example, Value category “>10,000,000” is the less frequent Value category in the database. In addition, out of the 41 Account Categories, there are only a few of them that occur frequently in the journal entries. The majority of the Account Categories have been posted a few times in a total.
4. ASSOCIATION RULES MINING

In this chapter, the selected Association rules Algorithm Apriori is applied in the transactional dataset in R. For this purpose, the setting of the parameters are decided for the Apriori algorithm. Finally, we describe the results that the algorithm derived. This chapter intends to provide an answer to the research question A2.

4.1. Generating the association rules

Before applying the Apriori algorithm, the parameters setting such as support and confidence must be decided. In this thesis, considering the amount of transactions (approx. 150,000 transactions) and the goal of this research, all kinds of account associations are needed, frequent and infrequent, therefore the support parameter is set very low as is set also the confidence parameter. Specifically:

- As explained in Chapter 2, the support of the rule is the fraction of observations in the union of the antecedent and consequent (T (A → B)). It can be viewed as an estimate of the probability of simultaneously observing both item sets Pr (A and B). In our case, study we set the support parameter equal to 0.00001. Setting the parameter to this value will produce all rules that were found more than once (0.00001*150,000=1.5). Finding the transactions that occurred once in the dataset is a straightforward task which does not require any data mining algorithm (in R the function is the unique()) to be discovered. In addition, setting the support parameter to 0 would result to rules that contain items which are not found in the dataset together.

- Confidence, C (A → B), of the rule is its support divided by the support of the antecedent. Confidence was developed together with support as a second threshold to prune the search space. Since, we aim to find all the co-occurrences of Account Categories in combination with the Value Category, we set the Confidence parameter set at 0.

- apriori function generates rules with minimum length 2(one attribute LHS one attribute at RHS) and maximum length 10. In this research, considering that in each transaction the itemset contains minimum two account categories (credit/debit) and the amount of money category, we set minimum length parameter for the rules equal to 3.

- As also stated in paper [38], rules with more than one item in the consequent (RHS) gives no true benefit but rather a considerable cost from having to deal with larger number of rules. Therefore, in this research rules, with one consequent are generated, and more specifically, the LHS contains the Account categories (Credit/Debit). The RHS contains only the Value Category of the transaction.

The results for the specific parameters that were set in the function apriori produced the following results:

- 1148 rules was generated from the algorithm with the above settings.

An example of the rules produced by Apriori are shown in the Table 6:
Table 6: Example of generated association rules

<table>
<thead>
<tr>
<th>Rules</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Tech_accounts_Credit,  Total_short_term_payables_Debit} =&gt; {0-100}</td>
<td>3.32E-05</td>
<td>0.3333333333</td>
</tr>
<tr>
<td>{Tech_accounts_Debit,    Total_short_term_payables_Credit} =&gt; {0-100}</td>
<td>3.98E-05</td>
<td>0.352941176</td>
</tr>
<tr>
<td>{Consumer_price_reductions&amp;marketing_allow_Credit,</td>
<td>0.00020579</td>
<td>0.96875</td>
</tr>
<tr>
<td>Prepayments_and_accrues_income_Debit} =&gt; {0-100}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{Accruals_and_deferred_income_Credit,</td>
<td>1.33E-05</td>
<td>1</td>
</tr>
<tr>
<td>Non_perf_trade_allowence_Debit} =&gt; {0-100}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{Consumer_price_reductions&amp;marketing_allow_Debit,</td>
<td>0.000112853</td>
<td>0.459459459</td>
</tr>
<tr>
<td>Prepayments_and_accrues_income_Credit} =&gt; {0-100}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{Prepayments_and_accrues_income_Credit,  Total_liquid_assets_Credit} =&gt; {0-100}</td>
<td>0.000119491</td>
<td>0.45</td>
</tr>
<tr>
<td>{Prepayments_and_accrues_income_Debit,</td>
<td>1.99E-05</td>
<td>0.09375</td>
</tr>
<tr>
<td>Total_receivables_Credit} =&gt; {0-100}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{Total_short_term_payables_Credit,  Trade_asset_related_expenses_Credit} =&gt; {0-100}</td>
<td>1.33E-05</td>
<td>0.01980198</td>
</tr>
<tr>
<td>{Total_receivables_Debit,</td>
<td>1.99E-05</td>
<td>0.022556391</td>
</tr>
<tr>
<td>Trade_asset_related_expenses_Credit} =&gt; {0-100}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{Prepayments_and_accrues_income_Debit,</td>
<td>0.00020579</td>
<td>0.8611111111</td>
</tr>
<tr>
<td>Total_liquid_assets_Debit} =&gt; {0-100}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6 shows a small subset of the rules generated by the Apriori algorithm, from the settings applied in the parameters. For example, the first rule shows that Account Category “Par_Acc_Credit” and “Par_Acc_Debit” have been infrequently found together (based on the support parameter) in the transaction database, however these two Account Categories have been found only in the Value Category “>10.000.000” (based on the parameter confidence=1). As it was expected, given also the parameters settings explanation, the generated rules are many in number in order to be manually checked (1148 rules) for unusual behavior. This behavior of association rules was explicitly mentioned in many papers as analyzed in Chapter 2. Therefore, the next chapter will describe the post processing steps applied in the produced rules in order to visualize the rules and extract the unusual rules particular for the thesis research problem.
5. POST PROCESSING

In this chapter, the selected post-processing techniques for the mining rules are presented and applied. As mentioned in the previous chapter, given the huge size of the mined rules, it may be very tedious for analysts to find interesting knowledge and even strong correlations between attributes are not always obvious from the discovered rules.

In chapter 2, the post-processing methods in association rules were introduced and described. In this thesis, redundancy method, visualization method and interestingness measure are specific selected and applied in regards to the research goal. Finally, this Chapter intends to provide an answer to the research question A4 and A5.

5.1. Grouping and Redundant rules

In this section, the methodology for removing the redundant rules is described. Based on the preliminaries in chapter 2, the Theorem 2 is selected for the redundant rules removing task. As a reminder, an example of the Theorem 2 is the following: A rule set $R$ has three rules, such as \( XY \rightarrow Z, X \rightarrow Z \) and \( Y \rightarrow Z \). Suppose rule \( XY \rightarrow Z \) has $s\%$ support and $c\%$ confidence. If $n$ (i.e. number of items in the antecedent) number of rules such as \( X \rightarrow Z \) and \( Y \rightarrow Z \) also satisfy $s$ and $c$ then, the rule \( XY \rightarrow Z \) is redundant because it does not convey any extra information if rule \( X \rightarrow Z \) and \( Y \rightarrow Z \) are present.

In our case study, the selection of this theorem is derived due to the fact that the consequence (RHS) is always the value Category. An example in our set of rules is the following:

1: \{Trade receivable Credit, Trade receivable Debit, Inventory Debit\} \( \rightarrow \) \{0-100\}
2: \{Trade receivable Credit, , Inventory Debit\} \( \rightarrow \) \{0-100\}
3: \{Trade receivable Credit, Trade receivable Debit \} \( \rightarrow \) \{0-100\}

If rule 2 and rule 3 have also the same support and confidence as rule 1, then rule 1 is redundant because it does not convey any extra information, if rule 2 and rule 3 are present.

The Theorem 2 is applicable when not only there is one consequent in the rule, but also the consequent must be the same. Therefore, rules have been grouped per Value categories and then in each value category the redundant theorem is applied. The grouping method is also applied in order to improve the comprehensibility of the discovered rules. After grouping the rules, there following commands were applied in order to remove the redundant rules from the original 1148 rules. The deletion of the redundant rules resulted to the following rules per Value category:

<table>
<thead>
<tr>
<th>Grouping total rules</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Subrules(“0-100”)</td>
<td>60</td>
</tr>
<tr>
<td>Subrules(“101-1.000”)</td>
<td>85</td>
</tr>
<tr>
<td>Subrules(“1.001-10.000”)</td>
<td>102</td>
</tr>
<tr>
<td>Subrules(“10.001-50.000”)</td>
<td>93</td>
</tr>
<tr>
<td>Subrules(“50.001-100.000”)</td>
<td>37</td>
</tr>
<tr>
<td>Subrules(“100.001-1.000.000”)</td>
<td>77</td>
</tr>
<tr>
<td>Subrules(“1.000.001-10.000.000”)</td>
<td>25</td>
</tr>
<tr>
<td>Subrules(“&gt;10.000.000”)</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 7: Rules per Value Category
The redundant rule deletion methodology reduced the 1148 rules to 492 rules in total. From the results, we can observe that the two last categories have the least amount of rules, which means that the diversity of account combinations are less, the more money is involved the less account categories are involved. On the other hand, in the value category “1.001-10.000” are discovered the highest number of rules. That means that the complexity of account combinations are higher or more account categories are involved in this value category.

5.2 Visualization
In this section, the visualization tool is developed, which aims to provide an overview of the pairs of accounts. Since the aim is to support auditors to identify potential unusual behavior, the inspection and analysis of the whole dataset through a visualization tool could be a supplementary procedure to the already defined auditing procedure (Appendix B2). The intention of this task is to provide auditors a tool that could provide a view of all account pairs in journal entries based on the whole dataset, in addition to the journal entry’s Value Categories.

One of the visualization tools that were introduced in the chapter 2 is via graphs. Graph visualization is the most easy to interpret and offers a very clear representation of rules based on paper[41], however, if the number of rules is large, it can easily become a complex graph from which no information about the transactions can be derived[41]. Therefore, the graphs should be as clear as possible without decreasing their insights.

Although the rules are reduced to 492, there is still a large number of rules to be visually displayed. For this purpose, eight different graphs were constructed for each Value category. Without this grouping, the graph displaying all 492 rules would be too complex to analyze. Figure 15 shows the graphs developed for value category “>10.000.000” which is the least complex category in our dataset.
All Account categories occurred in the Value category (">10,000,000") are shown in the Figure 9 (all attributes in the transactional database such as account categories and the Value Category are displayed with a green cycle). The other colored cycle indicate the rules that were found from the dataset in the Value Category >10,000,000. That is the reason why these colored cycles pointing to the particular Value Category (since all rules have as RHS the ">10,000,000" attribute). The Account categories pointing to the same colored cycle indicate that they have found together in the same rule as a LHS. The color of cycles indicates the Lift measure, the size of the cycle indicates the frequency (support) of the pair of attributes found in the transaction database.

In our case, if two Account Categories are pointing to the same colored cycle then the accounts occurred in the same journal entry (the simplest form of journal entry, one debit item one credit item). Furthermore, if three Account categories are pointing to the same cycle then they occurred in the same journal entry (more complex journal entry) etc. Given this graph, an auditor following the arrows can check which accounts are associated with each other for this particular Value category. Thus, even if one rule is not flagged as unusual, from a visualization graph, auditors can have a holistic view of the money flow inside the journal entries and determine if something is unusual. For example, it could be the case that one frequent behavior in one Value Category, which would not be extracted as "unusual" in the next step, is for auditors unusual. Therefore, through the graphs, auditors would have the opportunity to analyze both frequent and infrequent account pairs in each Value Category, and, based on their domain knowledge, determine which need further inspections.
However, for value categories that produced more rules, such as “1.001-10.000”, the graphs got more complex and it is difficult to analyze them. All graphs are included in Appendix C1. Observing the graphs, it seems that in value categories such as “>10.000.000”, “1.000.001-10.000.000”, “50.001-100.000” there is not large diversity in the account pairs, on the contrary, the value category “>10.000.000” seems to have account pairs that are really frequently found together in the dataset and, in addition, these pairs are strong correlated (red cycles). This means that this value category has account pairs that are only mostly found in these categories rather in the other. On the hand, the value categories such as “101-1.000”, “1.001-10.000”, “10.001-50.000” have rather complex graphs difficult to interpret. The majority of the account combinations are infrequently found (mostly small cycles) in the dataset and in addition they are not strongly correlated in the specific value categories (mostly yellow cycles), but are rather spread in all value categories.

Due to the drawback previously mentioned, the visualization graphs are supplementary to the objective measures that are applied in the next section.

5.3 Interestingness measures
In this section the unusual rules are discovered from the set of rules (492) that derived in the section (5.1) and later on used in the visualization graphs. The intention of this task is first to define what an unusual rule is and extract these rules by applying interestingness measures.

In Chapter 2, the interestingness measures were divided in two categories, objective measures and subjective measures. Objective (data-driven) measures are Probability-based measures that generate rules based on the data and do not take into account previous knowledge, whether subjective measures generates rules depending on the domain expert knowledge.

In this thesis, and since it is an explorative research we aim to find unusual rules that derived from the stored data and not from previous knowledge. Therefore, at the beginning of this section, the selection and the description of the Objective measures are analyzed in more detail and then the selected objective measures are applied on the discovered set of rules.

5.3.1. Objective measures Selection and Description
As explained in Chapter 2, many objective measures have been introduced as interestingness measures. Among these measures, none is ranked as the best objective measure among them, but rather the selection of the measure depends on the problem. Unfortunately, selecting the most appropriate measure needs domain experts’ interpretation, which is a time-consuming task and not a suitable method for a thesis. Therefore, book [36] provided a classification of the objective measures into categories. Each category included measures that produced almost the same results. Therefore, we selected one objective measure of each category, analyzed its definition, and applied it to the produced rules to extract the “unusual” among them. The selected measures are:

- Cosine
- Phi-coefficient
- Lift
- Conviction
- Certainty

A brief explanation is followed for the selected objective measure:
Cosine [36]

Cosine is defined by the type \( \frac{P(X|Y)}{\sqrt{P(X)P(Y)}} = \cosine(X \rightarrow Y) \). The closer \( \cosine(X \rightarrow Y) \) is to 1, the more transactions containing item X also contain item Y and vice versa. On the contrary, the closer \( \cosine(X \rightarrow Y) \) is to 0, the more transactions contain item X without containing item Y, and vice versa. The equality shows that transaction not containing neither X nor Y have no influence on the result of \( \cosine(X \rightarrow Y) \).

Lift [36]

Lift is defined by the type \( \frac{P(X|Y)}{P(X)P(Y)} = \mathit{lift}(X \rightarrow Y) = \frac{\mathit{conf}(X \rightarrow Y)}{P(Y)} \). If the Lift measure is equal to 1 this means that the occurrence of X and the occurrence of Y in the same transaction are independent events, hence X and Y are not correlated. Lift above 1 indicates a positive correlation and below 1 a negative correlation.

Phi-coefficient [36]

The phi coefficient is a measure of the degree of association between two binary variables. Phi is a chi-square based measure of association and depends on the strength of the relationship and size of the data. Given two binaries, for example LFS and RHS, the phi-coefficient is calculated from the type:

<table>
<thead>
<tr>
<th>LHS</th>
<th>RHS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>a</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
</tr>
</tbody>
</table>

\[ \phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \]

If the value lies below zero then it indicates a negative correlation. If the value lies near zero then it denotes very little or no correlation. Last, if the value lies above zero positive correlation. The closer the value lies near 1 the stronger the correlation is.

Conviction [36]

Conviction is defined by the type \( \frac{P(X)P(\overline{Y})}{P(X|\overline{Y})} \). Conviction was developed as an alternative to confidence which was found to not capture direction of associations. Conviction compares the probability that X appears without Y if they were dependent with the actual frequency of the appearance of X without Y. In that respect, it is similar to lift, however, in contrast to lift it is a directed measure since it also uses the information of the absence of the consequent.

Based on the type and the above description, conviction strength relies on the absence of the RHS when the LFS occurs. This type of measure is not in our interest for this thesis and, therefore, it is not applied.

Certainty [45]

Certainty is defined by the type \( \frac{P(Y|X)-P(Y)}{1-P(Y)} = \mathit{certainty}(X \rightarrow Y) \). The certainty factor is a measure of variation of the probability that Y (RHS) is in a transaction when only considering transaction with X (LHS). An increasing Certainty means a decrease of the probability that Y is not in a transaction that X is in. Negative Certainty has a similar interpretation. Again the values ranges from -1 to 1, with 0 indicating independence between X and Y.
5.3.2. Flagged unusual rules

The selected objective measures described in the previous section, produce three types of rules, positive correlated, negative correlated and independent rules. We are not interested in the negative ($\neg X \rightarrow Y$ or $X \rightarrow \neg Y$) rules or independent (the occurrence of one does not affect occurrence of the other) rules in this thesis, since the unusual behavior is defined as the account combinations containing an unusual amount of money compared to their frequent behavior. This means that the rules must have co-occurrences in the dataset. Therefore, we consider only the positive correlated rules.

However, among the positive correlated rules, there are also rules that have high support and high confidence. Consequently, these positive correlated rules includes also the account pairs that have been frequently found in the dataset in the same set and in the same Value Categories. These rules are irrelevant for this thesis, since they are not unusual.

Therefore, specific criteria have to be met for the rules to be flagged. For this purpose, we followed the skeptical of paper [81] which defines as unusual the rules that have low support and high confidence. This lies also with our unusual definition, because we aim to find infrequent behavior. Then, we discussed with the auditors if this definition of “unusual” is indeed interesting to be further investigated from an audition perspective, and the response was positive. Therefore, the flagged “unusual” rules are those that have been identified as positive correlated by the objective measures, but also follows the framework “low support, high confidence”.

The defined two criteria for the extraction of rules as “unusual” that need further inspection by auditors are:

1. Positive correlated rules
2. High confidence and Low support

For the second criteria, a minimum threshold is decided for the confidence and a maximum threshold for the support. There is no methodology that could give as the optimal minimum and maximum thresholds. In this research, we set the minimum confidence=0.7 and maximum support=0.05.

Figure 10 illustrates the rules that have been extracted as unusual from phi- coefficient based on the criteria stated previously. We extracted the rules that were fallen inside the red area. For the phi-coefficient measure the confidence set 0.7 and positive correlated rules are those above the 0. In Appendix C2, the plot for the other 3 objective measures and their red area containing the unusual rules are displayed.
Each measure produced a different number of unusual rules based on the two criteria. This deviation in the number is derived from the fact that each measure considers different concept as positive or negative correlated. From the plot, we can see that also high support rules have been extracted (red points) among the low support rules. Thus, the rules that were extracted were sorted by support. Rules that were above the maximum threshold were not considered in next step where we combine in one set the rules that were extracted as unusual from all measures.

After carefully inspecting the extracted rules, we observed that the measures produced almost the same results. In Appendix C2, the produced rules are displayed for the selected measures. The results from cosine or lift which are less than phi or certainty are a subset of the phi/certainty results. Phi-coefficient and certainty produced the same rules, except the one extra rule that certainty had in addition. Since the aim is to provide auditors with a set of rules that needs further inspection to determine if unusual behavior has indeed occurred, the certainty measure is selected which includes both the rules produced from the other measures but contains also a few more.

Therefore, based on the criteria previously defined, the unusual rules from the dataset which correspond to unusual combination of Accounts posted in a Value category are 64 and are displayed in Appendix C2(Certainty).
5.4. Summary

In this chapter, we have illustrated how account pairs and more complex account combinations that derived from Apriori mining, can be visualized. Prior to that, grouping techniques and redundant rules reductions methods were applied to the results to reduce the large number of the rules. Although these methods decreased the number of the resulted rules, value categories containing much and varied data had complex visualization graphs, difficult to analyze. Afterwards, objective measures applied on the whole transactional database constructed rules from the data. These rules provide information on whether the LHS and the RHS are positive correlated, negative correlated or independent. These measures have been developed as additional probabilistic measures to the support and the confidence due to their limitations. Based on our research, problem we are only interested in the positive correlated rules which also include the “low support, high confidence” criteria. Therefore, we applied 4 objective measures and extracted the rules that fell into our defined criteria.
6. EVALUATION

The purpose of this chapter is to evaluate our proposed approach given the research problem. The research goal is to support auditors to discover unusual financial transactions from the whole dataset. Association rule mining was the approach followed in this thesis. Therefore, we build two supplementary tools to achieve the research goal. The visualization of the rules and, on top of that, the extraction of the unusual rules. Rules in both tools represented accounts in journal entries in addition to the journal entry’s value category. Since the thesis goal has a supportive character in the auditing procedure and is not intended to replace the current auditing procedure, we evaluated our results based on the effectiveness, understandability and potential usage of this method.

The evaluation was performed in two stages. First, for the visualization graphs that aim to support auditors by providing a holistic view of the financial transactions (account combinations inside journal stored in the general ledger). Secondly, for the rules flagged as unusual, which represent the account combinations in respect to their Value category. In this chapter, initially an evaluation methodology for visualization results is introduced and then an evaluation methodology for the flagged unusual rules. The details of the evaluations are presented in the following sections. Finally, this chapter intends to provide an answer to the research question A6.

6.1. Evaluation of Visualization tool

For the visualization evaluation purpose, the Seven Guiding Scenarios of Information Visualization Evaluation is followed [76]. The seven scenarios guidelines is the first paper that conducted a structured survey of visualization evaluation methodologies and provided a guidance in respect to how to choose between the evaluation approaches. The research [76] was based on extensive literature analysis over 800 papers where the most common evaluation scenarios had been identified. In Appendix D1, the scenarios and methods are described. The general methodology that the paper suggest is

Step 1: Setting a goal
Step 2: Picking suitable scenarios based on the goal.
Step 3: Considering applicable approaches based on the selected scenarios
Step 4: Creating evaluation design and planned analyses based on the research goal.

After carefully studying all the scenarios and methods for each scenario, the visualization evaluation for this thesis is described in the following table.

<table>
<thead>
<tr>
<th>Visualization Evaluation in this case study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Goal</strong></td>
</tr>
<tr>
<td><strong>2. Suitable Scenario for the research goal</strong></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Table 9: Visualization Evaluation methodology

<table>
<thead>
<tr>
<th>3. Applicable approaches in this case study</th>
<th>Interviews/Questionnaire</th>
<th>Interviews/Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Evaluation design</td>
<td>Interviews based on Scenarios which are suitable for our research goal. Questions to the interviewees are based on some of the Scenario Questions. The selected questions are then developed along the guidelines for semi-structured interviews provided by paper [77].</td>
<td></td>
</tr>
</tbody>
</table>

6.1.1. Evaluation Setup
We interviewed three experienced auditors who are auditing the particular dataset that this thesis analyzed. The evaluation session in the interviews consisted of two parts; in the first half the research goal and research problem is briefly introduced, then the visualization results are demonstrated and explained. The second half is a question discussion session to receive feedback from the auditors on the results. The questions for the evaluation are based on the visualization evaluation methodology described in the previous sections. The interviews were semi-structured. Semi-structured refers to an interview in which the interviewee is asked a few predefined questions on several predefined topics and the interviewer has room to ask more questions in addition to the questions prepared. That means that we had prepared a set of questions before the interview took place but were not strict to these questions only. Semi-structured interviews setup provides an opportunity for the interview to ask more detailed questions if it is necessary and for the stakeholder to elaborate more on the topic.

Questions asked:

1. What kinds of visualization tools are currently in use? How do they help you to have an overview of the underlying data?
2. Is this visualization graph understandable after explanation (UE)?
3. What features are seen as useful (UE)?
4. What features are missing (UE)?
5. Is the graph adopted easily (CTV)?
6. Do you thing that with graphs displaying the whole dataset unusual behavior would be identified more easily (UE)?
7. Will this tool free you from a physical or mental effort in the auditing procedure (CTV)?
8. To what extent do you thing that an auditor may use this kind of tool (UE)?

6.1.2. Results
The overall feeling about the visualization graphs of the account categories and the way they display the flow of the credit/debit was positively expressed by the auditors. Specifically since there is no other tool that is currently used providing insights for the whole dataset. However, doubts have been expressed whether this kind of graphs will be adopted easily by the auditors. Additional recommendations on the visualization tool were in respect to the attributes used. Specifically, attributes such as user information or specific amount of money posted by accounts would be a good next step for implementation. A summary of the evaluation feedback is provided below:
Expert 1: Supervisor auditor (interview)

No visualization tools are currently used. The current tools used for inspecting MJE are pivot tables (excel) and the IDEA (interactive Data Extraction and Analysis) tool which extracts journal entries that meet specific risk criteria’s (Appendix B2). Inspections are based on subsets. The graphs need detailed explanation to be understood. After the explanation, the graph can be analyzed and also explained to other auditors. The more complex the graph gets the less value it gives to the auditors (graph for Value Category “1.001-10.000”). He thinks that aspect that seen as useful is the holistic overview of the flow of the money in each Value Category. The generalization to Account Categories was seen as a positive attribute in order to have an overview. The frequency of the journal entries was also seen as a positive feature to the graphs as also the Lift feature (although there were a difficulty to digest the meaning of the color). Furthermore, the auditor identified as a missing value the exact value of each account. An interactive graph seemed to him also as a necessary next step for a visualization tool, not only for displaying the exact value that each category has posted but also for getting the exact journal entry number from the frequency cycles.

Although he seemed positive about the visualization graphs, he had doubts whether a new way of inspecting MJE dataset will be adopted. He was skeptical whether auditors can adapt new things when they already are used to certain other methods. Training for this purpose is a necessary step. However, he stated that acquiring a holistic view of the MJE through graphs could lead to better reassurance about the financial statements. In addition, he stated that additional features, such as exact money value of account categories, could potential lead to less mental or physical effort, but not this tool.

Finally, he mentioned that these graphs could be a basis and with additional features they could be used as a tool by auditors. The graphs should be developed in a way that will be interactive and not complex. Training sessions are needed.

Expert 2: Senior Auditor (Report)

As the previous auditor mentioned, no visualization tool is currently used. He claimed that the graphs were difficult to understand, both the simple graphs but also the complex graphs. The features that were seen as useful where that all account categories in each Value categories were displayed. However, the features that he stated as missing were the count of items, which means that he did not understand the meaning of the cycle, one of the two features (the other is the Lift which is described by the color) that is seemingly more easy to understand. Based on the previous answers, he does not believe that the graphs are easily adaptive. However, he thinks that with easier visualization graphs the might be better possibilities to identify unusual transactions. Contrary to his previous answer, he does not believe that visualization tools will be used by auditors.

Expert 3: Auditor

After the description of the graphs, the auditor claimed that the process and the graphs were very clear to him. The features that he found useful were the clear insight in how transactions are flowing through the organization, and also, that they provide an indication of which accounts are connected. Specifically, he claimed that this way, the flows of the transactions between accounts are visible. Therefore, it can provide an indication whether logical transactions were executed. However, he mentioned that the graphs will provide a direction for the audit but might not be direct evidence. A feature that was seen as missing was the interactivity. The auditor stated that it would be nice to have buttons and inspect for
each account its connection. Furthermore, he believed that the tool can be adaptive if the process is explained and had been used once or twice. In addition, he was very positive that graphs can help auditors to identify unusual transactions. In the question whether the tool will free him from a physical or mental effort, he stated that it would give a more visible and viable insight in the journal entry process and, also, it might add additional consideration for the journal entry testing. Finally, he believed that this tool will be used in the future when more attributes could be used and the data would be derived from different kind of systems (SAP, JDE, Exact etc.) for the analysis.

6.2. Evaluation of the unusual account combinations

In our case we encountered difficulties to evaluate the effectiveness of our method. This difficulty arises from two reasons. First we do not had prior “unusual” data, and as mentioned in the Chapter 2, there is no knowledge among auditors about what fraud looks like. In case we had such dataset, we could evaluate to what extent our model would return the correct results by measuring the performance with recall (True Positives divided by the number of True Positives and the number of False Negatives), precision (True Positives divided by the number of True Positives and False Positives) or F-score (conveys the balance between the precision and the recall).

Secondly, a proper evaluation of the flagged set as unusual, needs the expertise of people with domain knowledge. This is a costly task in terms of time for auditors and impractical for a research with time restrictions as this one. It needs dedication of third parties (auditors), validating that indeed these rules need further inspection. Especially, considering the possibility that each auditor considered different rules as unusual or not unusual, iterative evaluations would have to be performed.

One evaluation measure for our methodology that we can have is the accuracy of the rules which can be evaluated by the Certainty objective measure [45]. This is very convenient to our thesis since our extracted rules are based on the certainty measure. Therefore, we can claim that our results are accurate in terms of the criteria (positive correlated rules, low support high confidence) we defined in the section 5.3.2. However, in order to evaluate the robustness of a model, accuracy alone is typically not enough information to make this decision.

Therefore, first, we evaluated the results of this tool with an alternative approach. We applied a different methodology for an unsupervised outlier identification for categorical data proposed in paper [82]. The proposed method considers number of categories inside categorical variables for outlier identification instead of distance. In our case, we considered the number of occurrences of account categories in each Value Category. Motivated also by [19] [24], where distribution of the value in each account category was used to discover the “suspicious” transactions, we evaluated our results using the distribution of the account categories in each Value Category. We examined one subset of the extracted rules whether they are indeed under the line of their relative frequency distribution. Appendix D2 explains in detail this approach and shows the results.

However, we realized that the alternative evaluation approach was not enough to measure whether we extracted unusual rules based on the terms we defined in Chapter 1 (Unusual financial transactions in this thesis are defined as account combinations inside journal entries containing an unusual amount of money compared to their frequent behavior).

Therefore, we constructed a synthetic dataset to evaluate our methodology. An evaluation based on synthetic dataset shows if the specific methodology applied in the section 5.3 also holds on practice. Based on paper [83], synthetic dataset is created to meet certain needs or conditions or test real data. The synthetic dataset has to mimic the real dataset in order to be effective. Thus, we developed a dataset that
included the attributes shown in Table 10. The size of the synthetic dataset and the attributes used in total are shown in Table 11.

<table>
<thead>
<tr>
<th>Credit</th>
<th>Debit</th>
<th>Value Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_Cr</td>
<td>A_De</td>
<td>0-100</td>
</tr>
<tr>
<td>B_Cr</td>
<td>B_De</td>
<td>101-1.000</td>
</tr>
<tr>
<td>C_Cr</td>
<td>C_De</td>
<td>1.001-10.000</td>
</tr>
<tr>
<td>D_Cr</td>
<td>D_De</td>
<td>10.001-100.000</td>
</tr>
<tr>
<td>E_Cr</td>
<td>E_De</td>
<td>&gt;100.001</td>
</tr>
<tr>
<td>F_Cr</td>
<td>F_De</td>
<td></td>
</tr>
<tr>
<td>I_Cr</td>
<td>I_De</td>
<td></td>
</tr>
<tr>
<td>K_Cr</td>
<td>K_De</td>
<td></td>
</tr>
<tr>
<td>G_Cr</td>
<td>G_De</td>
<td></td>
</tr>
</tbody>
</table>

Table 10: attributes created

---

The developed synthetic dataset was small compared to the original one (approximately 2%), but it mimics the behavior of the original dataset. Before constructing the dataset with the characteristics shown in Table 8 and Table 9, we constructed first, one random generated transaction database, but the transactions did not followed the general ledger behavior (the original dataset). For example, the maximum confidence among the rules was 0.19 and the rules had approximately the same support. Therefore, we did not used it in the evaluation, instead we manually generated one dataset that followed the behavior of the real dataset.

Since the size of the synthetic dataset was really small compared to the original we used less attributes representing the data. We included 9 categorical data (41 in the real dataset) that represent Account Categories (A, B, C, D, E, F, I, K, G) and 5 categorical data that represent the Value Categories. Each Account category can be both credited and debited as it is also the case in the real dataset. In total, in the synthetic transactional database, we had 23 attributes.

Apart the transactions included in the synthetic dataset, we manually added a set of unusual transactions. The goal is to measure to which extent the added unusual transactions are among the extracted rules.

To do that, first, we applied the Apriori algorithm in the synthetic dataset (without the unusual transactions). We set the support parameter equal to 0.001 and the confidence parameter equal to 0. The transactions implemented as unusual were decided based on the support. Figure 19 shows the top 10 highest support rules.
The unusual transactions that we added in the synthetic dataset were developed based on the rules shown in the Table 12. Since these rules have the highest support, we added transactions that had infrequent behavior compared to their frequent behavior. After the addition of these transactions the synthetic dataset had in total 2844 transactions. Table 13 shows the added as “unusual” transactions.

The Apriori Algorithm generated 75 rules from the synthetic transactional database. Then we applied the Certainty measure to the rules, as we did in the section 5.3.2 to the real dataset to see whether the unusual transactions that we added have been extracted. The rules resulted from Apriori, in addition to their support, confidence and Certainty measure are displayed in Appendix D3.

Since the number of the rules is not large (75 rules), we analyzed them manually from the Table (Appendix D3). We searched through the rules the unusual transactions that we added in the dataset. Table 14 shows the Certainty measure for each of the unusual transactions.
The unusual transactions that we added in the dataset could all be extracted if we chose to extract the negative correlated rules. In contrast to the criteria that we set in Chapter 5.3.2, where we extracted positive correlated rules that had high confidence and low support. The criteria that we set are based on the literature analysis that we have conducted, however, it seems that in the synthetic dataset that we developed, the unusual transactions are not extracted from the defined criteria but rather from the negative associated rules.

Since the synthetic dataset had labeled transactions as “unusual”, we were able to apply performance measures. The performance measures were applied for the extracted negative correlated rules which included the unusual transactions.

In Appendix D5 the plot for the Certainty objective measure is displayed and the red area containing the unusual rules. Table 15 shows the extracted negative correlated rules that were extracted after the Certainty measure was applied.

<table>
<thead>
<tr>
<th>Negative correlated rules</th>
<th>Labeled as unusual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 {B_Cr,F_De}=&gt;0-100</td>
<td>yes</td>
</tr>
<tr>
<td>2 {C_Cr,I_De}=&gt;100.001</td>
<td>yes</td>
</tr>
<tr>
<td>3 {C_Cr,B_De}=&gt;1.001-10.000</td>
<td>yes</td>
</tr>
<tr>
<td>4 {A_Cr,E_De}=&gt;100.001</td>
<td>yes</td>
</tr>
<tr>
<td>5 {E_Cr,K_De}=&gt;1.001-10.000</td>
<td>yes</td>
</tr>
<tr>
<td>6 {G_Cr,I_De}=&gt;1.001-10.000</td>
<td>yes</td>
</tr>
<tr>
<td>7 {A_Cr,K_De}=&gt;100.001</td>
<td>yes</td>
</tr>
<tr>
<td>8 {I_Cr,B_De}=&gt;1.001-10.000</td>
<td>yes</td>
</tr>
<tr>
<td>9 {F_Cr,A_De}=&gt;100.001</td>
<td>yes</td>
</tr>
<tr>
<td>10 {K_Cr,C_De}=&gt;100.001</td>
<td>no</td>
</tr>
<tr>
<td>11 {D_Cr,K_De}=&gt;0-100</td>
<td>no</td>
</tr>
<tr>
<td>12 {C_Cr,D_De}=&gt;10.001-100.000</td>
<td>no</td>
</tr>
<tr>
<td>13 {I_Cr,A_De}=&gt;0-100</td>
<td>no</td>
</tr>
<tr>
<td>14 {K_Cr,G_De}=&gt;0-100</td>
<td>no</td>
</tr>
<tr>
<td>15 {D_Cr,C_De}=&gt;0-100</td>
<td>no</td>
</tr>
<tr>
<td>16 {F_Cr,E_De}=&gt;100.001</td>
<td>no</td>
</tr>
<tr>
<td>17 {G_Cr,C_De}=&gt;100.001</td>
<td>no</td>
</tr>
<tr>
<td>18 {B_Cr,A_De}=&gt;1.001-10.000</td>
<td>no</td>
</tr>
</tbody>
</table>

Table 15: Extracted unusual rules

For the performance evaluation we applied the confusion matrix by measuring the standard metrics that are commonly used [84]. The confusion matrix is shown in Figure 11.
In the Confusion Matrix, TN (True Negative) indicates the number of correct prediction that an instance is irrelevant, in our case, TN are 66 rules. FP (False Positive) indicates the number of incorrect predictions that an instance is relevant, in our case, 9 rules are FP. FN (False Negative) indicates the number of incorrect predictions that an instance is irrelevant, in our case we have 0 FN. TP (True Positive) indicates the number of correct predictions that an instance is relevant, in our case, 9 rules are TP. Standard Performance Measures [84]:

\[
\text{Accuracy} = \frac{TN + TP}{TN + FN + FP + TP} = \frac{66 + 9}{66 + 0 + 9 + 9} = \frac{75}{84} = 0.89
\]

\[
\text{Precision} = \frac{TP}{FP + TP} = \frac{9}{18} = 0.50
\]

\[
\text{Recall} = \frac{TP}{FN + TP} = \frac{9}{0 + 9} = 1
\]

\[
F - \text{Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} = \frac{2 \times 1 \times 0.5}{2 + 1} = 0.66
\]

The performance measures results showed that extracting the negative correlated rules can extract all the unusual transactions. However, the method extracted also other rules that were also irrelevant (or were unusual in terms we defined in Chapter 1 but were not manually added). On the other side, none unusual transaction, among the added ones, were missing from the extracted results.

A possible explanation for this result, which contradicts the criteria that we set, is that the measure Certainty takes into account both confidence of the rule and the support of the RHS. For example, a negative correlated rule from the Certainty measure is when Conf(A→B) < Supp(B), which means:

\[
\frac{\text{Supp}(A \rightarrow B)}{\text{Supp}(A)} < \text{Supp}(B).
\]

In our case, we created the unusual financial transactions from the highest support rules, which means that Supp(A) increased more is and the Supp(A→B) is a really small number. Since the Value categories (the B), which we included in the unusual transactions, have relative high support then, probably, that is the reason why all unusual financial transactions that we added had negative correlated rules. To conclude, the results of the evaluation might have derived from the format of our synthetic dataset or the defined criteria that we set in section 5.3.2 do not produce the estimated results for the specific problem.

6.3. Conclusion
The overall evaluation of this approach, the two supplementary methods for supporting the auditors, had good results. During the evaluation of the first method, the visualization tool, we involved domain experts. The feedback provided by them was regarding the usability, effectiveness and the degree the tool can be
learned easily. The size of the participants is relatively small (3 domain experts) but provides meaningful insights about the tool capabilities and the intention of usage.

First, since no visualization tool was previously used, the idea of a visualization tool for obtaining a holistic overview of all manual journal entries was well perceived by the participants and all agreed that it could be a potential support method for reassurance. The graphs that were developed as visualization tool were difficult to understand with only a brief explanation, it needed detailed description of the graphs with small examples in paper so that they could understand the meaning of support and Lift displayed in the graphs (size and color of the edges respectively). Additional features that they felt that could enhance the capabilities of this tool were mentioned by all experts, such as the exact value of Account categories and exact number of frequent journal entries. An interactive graph was also proposed for further research, such as buttons that could extract the account combination only for specific Accounts or for displaying the flows from specific Accounts. Furthermore, they also agreed that this kind of tool would not spare them time in the auditing procedure or reduce their effort, rather it could be a supplementary method for reassurance or additional consideration for the journal entry testing. However, from all experts we received a doubt whether a visualization tool will be actually used by experts. They showed a hesitation whether new methods in the auditing procedure will be adopted by auditors.

The extraction of the 64 rules as unusual was impossible to be evaluated because we did not had transactions flagged as unusual to measure the effectiveness of the method. In addition, an evaluation by domain experts is impractical for this thesis since it requires a significant amount of time. Therefore, we constructed a synthetic dataset that behaved similar to the original one. We added unusual transactions inside the synthetic dataset and measured the performance of our approach. The results showed that the unusual transactions produced negative correlated rules when the measure Certainty was applied. This is in contrast to our defined criteria in Chapter 5.3.2, where unusual rules were defined as positive correlated rules with low support and high confidence. The criteria were selected from the literature, where a negative association rule is one of the following: ¬X → Y or X → ¬Y (where X means existence and ¬X means absence in enough transactions) whether the Certainty measure defines negative correlated rules when the Supp(B) is greater than the Conf(A→B).

Furthermore, the standard performance measures, such as Accuracy, Recall, Precision and F-Measure, were applied to the negative correlated rules. The results showed that all unusual transactions that were added in the synthetic dataset were extracted if we selected as criteria of extraction the negative correlated rules.
7. CONCLUSION

In this chapter, we discuss our findings and reflections of the research and answer the main research question defined in the Chapter 1. Furthermore, limitations of this research is discussed, and recommendation for further research of this thesis is defined.

7.1. Research Conclusion

Considering the current auditing procedure, auditors rely on means such as data filtering (running query commands) and sample reviewing of the journal entries. Therefore, since the journal entry testing is extracted from predefined criteria derived from previous knowledge and not from knowledge of the underlying dataset, there is a lack in the quality of the overview which might lead to unexplored suspicious transactions.

The goal of this research is providing KPMG with a tool that would increase the quality of review of the general ledger and, in hence, would support auditors in the identification of unusual financial transactions. Therefore we proposed data mining journal entries in the general ledger. We focused in this research specifically in the Manual Journal Entries excluding the system generated journal entries. After the initial phase, the investigation of the literature regarding data mining journal entries and discussing KPMG current auditing procedure, we also identified also the literature gap. We focus in this research on account pairs (sets) included in journal entries and the journal entries value. A method that in literature is not yet investigated and is proposed for further research. In addition, we defined what an unusual financial transaction is in our thesis. A definition that is subjective and derived after discussion with domain experts. Since the co-occurrences of pairs is the main focus, we selected to apply Association rule algorithms.

Therefore we defined our main research question as: “How can association rule techniques support domain experts in finding unusual financial transactions from the whole dataset of a general ledger?”

To answer the main question several sub-questions were defined. The first step was analyzing the literature conducted in the field of Fraud in Accounting, Association rule mining and the Post-processing methods identified for the Association rules mining algorithms. This chapter provided us with the basis to answer our sub-questions.

First we decided to apply the Apriori algorithm, and then, we selected the appropriate attributes from the general ledger regarding our research problem. After choosing the attributes that we believed were the most appropriate for our research problem, we transformed them and created a new dataset that had the format of a transaction database and simultaneously did not lose information about the journal entries. The next step was the application of the Apriori algorithm. Since we wanted to have insights about the whole dataset, we selected the most possible minimum threshold as a parameters. The algorithm produced initially 1148 rules from the dataset, providing information about both for the support and confidence of each rule. The Apriori produced many rules, and therefore, impossible to be investigated by auditors in order to identify unusual financial transactions.

Association rules provides, through the post-processing step, the opportunity to analyze the rules further. Our first concern was to reduce the rules as much as possible without losing any insight of the dataset. Thus, we removed the redundant rules, keeping 492 rules out of 1148. Then, we created two supplementary tools, the visualization tool and the extraction of unusual rules tool. These tools were intended to support auditors in finding unusual financial transactions. We defined them as supplementary to each other because some of the visualization graphs are complex to analyze and it also depends on the auditors’ eye not to miss any unusual behavior. On the other hand, the extraction of unusual rules by itself would limit the possibility to find something as unusual only to the extracted rules.
The visualization tool displayed the rules discovered in the dataset in graphs. For each Value Category a graph was created. The rules corresponded to account co-occurrences in journal entries. Domain experts, after evaluating this tool, agreed that it could be a supportive tool to discover unusual financial transactions. Since no visualization tool are currently in use, this tool provided an overview of the flow of the transactions and could add journal entries for testing that were not previously considered. However, they also agreed that the tool as it is now, could not be used yet, but can be perceived as a basis. The main reason for that is that they believed an interactive tool is more user friendly to the domain experts and that some complex graphs are difficult to analyze.

The extraction of unusual rules could provide a supplementary method to discover unusual financial transactions. Due to the fact that our dataset contained many journal entries, we created eight Value Category graphs, some of which were complex to be manually inspected. Therefore, we applied objective measures in the resulted rules, another post-processing step in Association rule mining. We defined 2 criteria (positive correlated rules with low support and high confidence) in our thesis which corresponded to the unusual rules. After the application of the objective measures, out of 492 rules, 64 were extracted as unusual. However, in the evaluation phase, testing a synthetic dataset, we observed that the unusual transactions were extracted from the negative correlated rules. Therefore, additional tests in other synthetic or real dataset should be made to validate whether negative correlated ruled can indeed extract the unusual financial transactions or the reason behind the evaluation results was the created synthetic dataset.

To conclude, Association rules mining provided us with the opportunity to analyze the whole dataset of a general ledger, and through the post-processing methods that are applicable to the rules, could support domain experts in the auditing process. Association rules mining was applied because we focused on account pairs (set of accounts), in contrast to previous literature that aimed to discover suspicious behavior in individual accounts. Discovering the association between the accounts in a general ledger was interesting both from a business perspective (experts can follow the flow of the money and observe whether unusual behavior took place) but also from a literature perspective. The combination of our two supplementary tools that were developed from the post-processing phase in Association rules mining can be seen as one support tool for the auditors in identifying unusual financial transaction (accounts in a journal entry). We cannot claim that it is a complete tool, but rather a basis for further research, both for the visualization tool and for extracting the unusual financial transactions.

7.2. Limitations
Some limitations that we encountered during this thesis are described in this section:

- At the beginning of our research we had to make an assumption concerning the definition of the “unusual” transactions. Unusual or suspicious transactions in journal entries do not have specific characteristics and auditors do not know what fraud looks like.
- Considering the structure of a journal entry, it might be the case that a pair of accounts appears many times inside a few journal entries. In that sense then, it can be considered as a frequent account pair. Apriori, however, lies its results to a Boolean matrix that counts whether an attribute is found or not. Therefore, in this case, the account pair will be found infrequent in the dataset.
- Despite the fact that we used multiple methods reducing the rules to a manageable number, value categories such as “1.001-10.000” still produced a large number of rules. Consequently, the graphs displaying the journal entries in these value categories where complex to analyze.
- We had no guidance in the parameters selection in the second criteria “High confidence, Low support”. The parameters were based on logic. A different setting would result to more rules or fewer rules.
• Due to the fact that we analyzed MJE, the majority of them were infrequent by nature (see Figure 14) in the general ledger, thus the criteria measures applied in the rules resulted many flagged rules as unusual.

• The evaluation of the visualization was conducted by interviewing three domain experts. The evaluation results might be different if more domain experts were involved.

• Due to time constraints and the large number of the rules that were extracted as unusual from the interestingness measures, we could not evaluate whether indeed, from an accounting perspective, and the unusual rules corresponded to unusual journal entries that need further inspection.

• The developed synthetic dataset was small compared to the original. Therefore less attributes were used in the dataset. The evaluation of the second tool was based on the synthetic dataset and despite the fact that it produced good results, we cannot know whether it will have the same performance in large datasets.

7.3. Future research

Data mining journal entries for detecting unusual behavior is a challenging task and is still not thoroughly investigated. This research was performed solely for the specific case study provided from KPMG. However, conducting several case studies could develop a more comprehensive approach in the definition and identification of unusual behavior of journal entries. This thesis is considered as an explorative research and, in this direction, future work such as the following can be conducted:

• Other unsupervised data mining could be applied to the general ledger. Clustering algorithms, for example, could be applied to accounts in addition to their credit/debit amount. Outliers could be discovered from the dataset, either as outlier point representing an unusual amount of value or as an outlier class representing a set of values as outliers. Other attributes, such as users or date could also be considered. Especially the month/day of the month that an account has been posted is something that auditors are concerned about. Therefore, clustering algorithms in terms of value and month or value and users could also be a future research topic for detecting unusual behavior.

• Another direction of future research is to measure which of the objective measures produced the more accurate results regarding unusual account pairs. However, this research needs constant interaction and evaluation by domain experts to rank each measure and define which one of them has the best results.

• Future work of our research can address the complexity of the visualization tool. This can be achieved by visualizing specific account categories from all value categories in one graph. Another way is to develop an interactive tool that includes aggregated information about the account pairs and values. Removing redundant rules in a more specific manner could also produce simpler and more understandable graphs. For example, visualizing accounts that were only found infrequent in the dataset or accounts that have an immediate effect on the revenue.
LITERATURE


[26] Qi Liu, The application of exploratory data analysis in auditing, PhD dissertation Graduate School-Newark, 2014


[29] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The elements of statistical Learning, Data mining, Inference and Predictions, Book, 2010


[33] Pang-Ning Tan, Selecting the right interestingness measure for association patterns, Information Systems 29, pages 293–313, 2014


[49] Fernando Berzal, Juan-Carlos Cubero, Nicolas Marin, Anomalous Association Rules, IEEE ICDM workshop alternative techniques for data mining and knowledge, 2004


[55] Cindy Durtschi, William Hillison,Carl Pacini,Use of Benford’s Law to Assist in Detecting Fraud in Accounting Data Use of Benford’s Law to Assist in Detecting Fraud in Accounting Data, Journal of Forensic Accounting,Vol.V, pages 17-34,2004

[56] Gogi Overhoff, The impact and reality of Fraud Auditing Benford’s Law: Why and how to use it, 22nd Annual ACFE Fraud Conference and Exhibition, 2011


[66] Han, J. and Pei, J., Mining frequent patterns by pattern-growth: methodology and implications. ACM SIGKDD Explorations Newsletter 2, pages 14-20, 2000

[67] Ivan Bruha, A. (Fazel) Famili, Postprocessing in Machine Learning and Data Mining, ACM SIGKDD Explorations Newsletter - Special issue on “Scalable data mining algorithms”, Volume 2 Issue 2, 2000


[71] Ivan Bruha, A. (Fazel) Famili, Post processing in Machine Learning and Data Mining, ACM SIGKDD Explorations Newsletter, Scalable data mining algorithms Volume 2 Issue 2, Pages 110-114, 2000


APPENDIX A

Appendix A1: Attributes

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandant</td>
<td>Business Unit</td>
</tr>
<tr>
<td>Fiscal year</td>
<td>Posting period</td>
</tr>
<tr>
<td>Account type</td>
<td>Account type text</td>
</tr>
<tr>
<td>Posting key</td>
<td>Posting key text</td>
</tr>
<tr>
<td>Document Item Nr</td>
<td>Document Header Text</td>
</tr>
<tr>
<td>Entry time</td>
<td>Last change date</td>
</tr>
<tr>
<td>Entry User Type</td>
<td>Entry User Name</td>
</tr>
<tr>
<td>Clearing Date</td>
<td>Clearing Document</td>
</tr>
<tr>
<td>Material Nr</td>
<td>Profit Center</td>
</tr>
<tr>
<td>GR/IR Account</td>
<td>Allowed manual posting</td>
</tr>
<tr>
<td>Line item</td>
<td>G/L account assigned automatically</td>
</tr>
<tr>
<td>Value(Doc)</td>
<td>Currency(CC)</td>
</tr>
<tr>
<td>Transaction Key</td>
<td>Business Transaction</td>
</tr>
<tr>
<td>Reversal doc Year</td>
<td>Reversal Doc Nr</td>
</tr>
<tr>
<td>Open item mngm</td>
<td>Status</td>
</tr>
<tr>
<td>Noted item without balance update</td>
<td>Bach input</td>
</tr>
<tr>
<td>Manual bank payment</td>
<td>Local currency</td>
</tr>
<tr>
<td>Entry 8pm-6am</td>
<td>000 or 999 value</td>
</tr>
<tr>
<td>Posting within same g/l account</td>
<td>All line items of posting automatically created</td>
</tr>
</tbody>
</table>

Table A.1: Attributes description

Appendix A2: Chart of Account

<table>
<thead>
<tr>
<th>Number Represents:</th>
<th>Account Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Accruals and deferred income</td>
<td>Accrual income: another form of receivable that we will invoice and collect at some point in the future. Deferred: removes the liability no longer needed</td>
</tr>
<tr>
<td>2</td>
<td>Actuarial gains / losses reserve</td>
<td>the need to make assumptions about the future rate of salary increases, the length of employee tenure</td>
</tr>
<tr>
<td>3</td>
<td>Bad debts</td>
<td>A bad debt is an amount owed by a debtor that is unlikely to be paid</td>
</tr>
<tr>
<td>4</td>
<td>Bad goods expenses</td>
<td>Bad goods expenses often refers to the loss that a company experiences because it sold goods or provided</td>
</tr>
</tbody>
</table>

61
| 5. | Cash discounts to customers | Services and did not require immediate payment |
| 6. | Consumer price reductions & marketing allowance | - |
| 7. | Cost of goods sold (own manufactured) | The cost of goods manufactured is the cost assigned to units either completed or still in the process of being completed at the end of an accounting period |
| 8. | Cost of goods sold (purchased IG) | - |
| 9. | Current taxes (as per tax return) | - |
| 10. | Deferred taxes on temporary differences | Deferred tax assets can arise due to net loss carry-overs, which are only recorded as asset if it is deemed more likely than not that the asset will be used in future fiscal periods. |
| 11. | Financial assets | A financial asset is an intangible asset whose value is derived from a contractual claim, such as bank deposits, bonds, and stocks |
| 12. | Fixed distribution expenses | Costs like advertisement, salaries etc |
| 13. | Functional accounts | Fictional accounts classify transactions according to the end use or purpose for which the obligation, expenditure, or collection is made |
| 14. | Gains/(losses) on real estate operations | - |
| 15. | IG expenses / income | - |
| 16. | Impairment of assets (excluding goodwill) | Seeks to ensure that an entity’s assets are not carried at more than their recoverable amount |
| 17. | Interest expense on employee benefits | Employee benefits includes short-term benefits, post-employment benefits, other long-term benefits, termination benefits |
| 18. | Inventories | Refers to the goods and materials that a business holds |
| 19. | Litigation (other than income tax) & unfound claim | - |
| 20. | Net general license fee | - |
| 21. | Net interest expense/ income | The cost incurred by an entity for borrowed funds. Interest expense is a non-operating expense shown on the income statement. It represents interest payable on any type of borrowings – bonds, loans, |
| 22. | Nonperformance trade allowances | Allowances given to a customer that are not related to performance and have limited or no perceived value for trade |
| 23. | Operational IG Loans | - |
| 24. | Other product fixed expenses | Fixed operating expenses include many different costs that a business is |
obligated to pay and cannot decrease over the short run without major surgery on the human resources and physical facilities of the business.

| 25 | Other variable expenses | Virtually every business has *variable expenses*, which move up and down in tight proportion with changes in sales volume or sales revenue e.g. Cost of goods sold, commissions paid, transportation costs |
| 26 | Parallel accounts (total must be 0) | This enables you to perform valuations and closing preparations for a company code according to the accounting principles of the group as well as other accounting principles |
| 27 | Prepayments and accrued income | Accrued income is income which has been earned but not yet received. Prepaid expense is expense paid in advance but which has not yet been incurred. |
| 28 | Production related overheads | Indirect expenses associated with processes used to produce a good or service. Production overhead may include expenses such as stationery, utilities, support staff salaries, and rent or other facilities costs |
| 29 | Retained earnings | *Retention earnings* is the percentage of net earnings not paid out as dividends, but retained by the company to be reinvested in its core business or to pay debt |
| 30 | Taxes on actuarial gains / losses | An actuarial gain or loss is a gain or loss arising from the difference between estimates and actual experience in a company’s *pension plan* |
| 31 | Technical accounts for IS (ttl must be 0) | - |
| 32 | Total liquid assets | An asset that can be converted into cash quickly and with minimal impact to the price received |
| 33 | Total medium/long term debts | Long-term debt consists of *loans* and financial obligations lasting over one year |
| 34 | Total medium/long term provisions | - |
| 35 | Total product fixed marketing expenses (NNS) | Fixed expenses are costs that typically remain the same regardless of changes or fluctuations in production levels or sales volumes. |
| 36 | Total receivables | The *total* money owed to a company by its customers, minus the money owed that will likely never be paid |
| 37 | Total reserves | *Sum of all deposits that a depository institution (bank, building society, credit union, finance company, insurance company) is allowed to take* |
into account as a part of its legal reserve requirements

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>38</strong></td>
<td>Total short term payables</td>
<td>all obligations to pay out cash at some date in the near future, including amounts that a firm owes to trade creditors and bank loans etc.</td>
</tr>
<tr>
<td><strong>39</strong></td>
<td>Trade asset related expenses</td>
<td>outflow of cash or other valuable assets from a person or company to another person or company</td>
</tr>
<tr>
<td><strong>40</strong></td>
<td>Trade spend related expenses</td>
<td>outflow of cash or spending from a person or company to another person or company</td>
</tr>
<tr>
<td><strong>41</strong></td>
<td>VDE finished goods transport &amp; handling items</td>
<td>-</td>
</tr>
</tbody>
</table>

Table A.2: Chart of Accounts

Appendix A3: Accounting tree

Figure 12: Accounting tree [11]
Appendix A4: Level of Description

Figure 13: Level of Description of Accounts IDs
APPENDIX B
Appendix B1: Classification of data mining algorithms

Figure 14: Classification of data mining algorithms [12]
Appendix B2 Current Auditing procedure

Figure 15: Current auditing procedure for Manual Journal Entries

Appendix B3

**Generality/Coverage.** A pattern is general if it covers a relatively large subset of a dataset. Generality (or coverage) measures the comprehensiveness of a pattern, that is, the fraction of all records in the dataset that matches the pattern.

**Conciseness.** A pattern is concise if it contains relatively few attribute-value pairs, while a set of patterns is concise if it contains relatively few patterns.

**Novelty.** A pattern is novel to a person if he or she did not know it before and is not able to infer it from other known patterns.

**Reliability.** A pattern is reliable if the relationship described by the pattern occurs in a high percentage of applicable cases.
**Peculiarity.** A pattern is peculiar if it is far away from other discovered patterns according to some distance measure. Peculiar patterns are generated from peculiar data (or outliers), which are relatively few in number and significantly different from the rest of the data.

**Diversity.** A pattern is diverse if its elements differ significantly from each other, while a set of patterns is diverse if the patterns in the set differ significantly from each other. Diversity is a common factor for measuring the interestingness of summaries.

**Surprisingness.** A pattern is surprising (or unexpected) if it contradicts a person’s existing knowledge or expectations.

**Utility.** A pattern is of utility if its use by a person contributes to reaching a goal. Different people may have divergent goals concerning the knowledge that can be extracted from a dataset.

**Actionability/Applicability.** A pattern is actionable (or applicable) in some domain if it enables decision making about future actions in this domain [75].
APPENDIX C

Appendix C1: Visualization graphs

Figure C.1: Graph “0-100”
Figure C.2: Graph for “1.000.001-10.000.000”
Figure C.3: Graph for “50.001-100.000”
Figure C.4: Graph for “1.001-10.000”
Figure C.5: Graph for “10.001-50.000”
Figure C.6: Graph for “1.001-10.000”
Appendix C2: Objective measures plots

Figure C.7: Lift extraction plot
Figure C.8: Cosine extraction plot

Figure C.9: Certainty extraction plot
Appendix C3: Extracted rules

<table>
<thead>
<tr>
<th>Rule ID</th>
<th>Description</th>
<th>Weight</th>
<th>Support Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Total_receivable_Credit, Total_receivable_Debit</td>
<td>2.697566e-11</td>
<td>0.846744e-06</td>
<td>(10.00000, 95.0)</td>
<td>0.789752</td>
<td>(10.00000, 65.0)</td>
<td>0.789752</td>
<td>(10.00000, 95.0)</td>
</tr>
<tr>
<td>2</td>
<td>Total_liquidity_asset_Credit, Total_receivable_Credit</td>
<td>1.68978e-11</td>
<td>0.721399e-06</td>
<td>(10.00000, 95.0)</td>
<td>0.693375e-05</td>
<td>(10.00000, 65.0)</td>
<td>0.693375e-05</td>
<td>(10.00000, 95.0)</td>
</tr>
<tr>
<td>3</td>
<td>Total_liquidity_asset_Credit, Total_receivable_Debit</td>
<td>9.274212e-12</td>
<td>0.520194e-06</td>
<td>(10.00000, 95.0)</td>
<td>0.520194e-06</td>
<td>(10.00000, 65.0)</td>
<td>0.520194e-06</td>
<td>(10.00000, 95.0)</td>
</tr>
<tr>
<td>4</td>
<td>Total_liquidity_asset_Credit, Total_receivable_Debit</td>
<td>6.863856e-16</td>
<td>0.394667e-09</td>
<td>(10.00000, 95.0)</td>
<td>0.394667e-09</td>
<td>(10.00000, 65.0)</td>
<td>0.394667e-09</td>
<td>(10.00000, 95.0)</td>
</tr>
<tr>
<td>5</td>
<td>Total_liquidity_asset_Credit, Total_receivable_Debit</td>
<td>7.275672e-15</td>
<td>0.771651e-09</td>
<td>(10.00000, 95.0)</td>
<td>0.771651e-09</td>
<td>(10.00000, 65.0)</td>
<td>0.771651e-09</td>
<td>(10.00000, 95.0)</td>
</tr>
<tr>
<td>6</td>
<td>Total_liquidity_asset_Credit, Total_receivable_Debit</td>
<td>1.397163e-13</td>
<td>0.358174e-08</td>
<td>(10.00000, 95.0)</td>
<td>0.358174e-08</td>
<td>(10.00000, 65.0)</td>
<td>0.358174e-08</td>
<td>(10.00000, 95.0)</td>
</tr>
<tr>
<td>7</td>
<td>(Non_perform_asset_Credit, Total_liquidity_asset_Credit)</td>
<td>3.171798e-14</td>
<td>0.831269e-09</td>
<td>(10.00000, 95.0)</td>
<td>0.831269e-09</td>
<td>(10.00000, 65.0)</td>
<td>0.831269e-09</td>
<td>(10.00000, 95.0)</td>
</tr>
<tr>
<td>8</td>
<td>(Non_perform_asset_Credit, Total_liquidity_asset_Credit)</td>
<td>2.129005e-14</td>
<td>0.521893e-09</td>
<td>(10.00000, 95.0)</td>
<td>0.521893e-09</td>
<td>(10.00000, 65.0)</td>
<td>0.521893e-09</td>
<td>(10.00000, 95.0)</td>
</tr>
<tr>
<td>9</td>
<td>(Non_perform_asset_Credit, Total_liquidity_asset_Credit)</td>
<td>3.477807e-16</td>
<td>0.381131e-09</td>
<td>(10.00000, 95.0)</td>
<td>0.381131e-09</td>
<td>(10.00000, 65.0)</td>
<td>0.381131e-09</td>
<td>(10.00000, 95.0)</td>
</tr>
<tr>
<td>10</td>
<td>(Non_perform_asset_Credit, Total_liquidity_asset_Credit)</td>
<td>1.827762e-14</td>
<td>0.486126e-09</td>
<td>(10.00000, 95.0)</td>
<td>0.486126e-09</td>
<td>(10.00000, 65.0)</td>
<td>0.486126e-09</td>
<td>(10.00000, 95.0)</td>
</tr>
<tr>
<td>11</td>
<td>(Non_perform_asset_Credit, Total_liquidity_asset_Credit)</td>
<td>2.087906e-14</td>
<td>0.481131e-09</td>
<td>(10.00000, 95.0)</td>
<td>0.481131e-09</td>
<td>(10.00000, 65.0)</td>
<td>0.481131e-09</td>
<td>(10.00000, 95.0)</td>
</tr>
</tbody>
</table>

Figure C.10: Cosine extracted rules
Figure C.11: Lift extracted rules

<table>
<thead>
<tr>
<th>No.</th>
<th>Rules</th>
<th>Lift</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Accrued and deferred income credit, Accrued and deferred income debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>2</td>
<td>(Financial assets Credit, Taxes on accrual basis income debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>3</td>
<td>(Accrued gains/ losses reserve Credit, Accrued gains/ losses reserve Debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>4</td>
<td>(Accrued gains/ losses reserve Credit, Financial assets Credit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>5</td>
<td>(Accrued gains/ losses reserve Credit, Functional accounts)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>6</td>
<td>(Functional accounts, Functional accounts)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>7</td>
<td>(Financial assets Credit, Functional accounts)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>8</td>
<td>(Prepayment and advance income Credit, Trade expenses related expenses Debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>9</td>
<td>(Inventories Credit, Other expenses related expenses Debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>10</td>
<td>(Inventories Debit, Production related expenses Credit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>11</td>
<td>(Functional accounts, Trade expenses related expenses Debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>12</td>
<td>(Accrued and deferred income credit, Total short term payable Credit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>13</td>
<td>(Accrued and deferred income debit, Total liquid assets Credit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>14</td>
<td>(Accrued and deferred income credit, Total short term payable Debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>15</td>
<td>(Functional accounts, Other expenses related expenses Credit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>16</td>
<td>(Net interest expense/ income Debit, Total liquid assets Debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>17</td>
<td>(Net interest expense/ income credit, Total liquid assets Credit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>18</td>
<td>(Accrued and deferred income credit, Technical accounts Credit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>19</td>
<td>(Accrued and deferred income credit, Functional accounts)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>20</td>
<td>(Net interest expense/ income credit, Total liquid assets Debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>21</td>
<td>(Net interest expense/ income credit, Total liquid assets Credit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>22</td>
<td>(Cash discounts to customers Credit, Total short term payable Debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>23</td>
<td>(Cash discounts to customers Credit, Total liquid assets Credit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>24</td>
<td>(Deferred taxes on temporary differences Debit, Total medium/long term debt Debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>25</td>
<td>(Taxes on accrual basis income Debit, Total medium/long term debt Debit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>26</td>
<td>(Net interest expense/ income credit, Total liquid assets Credit)</td>
<td>&gt; 120,000.00</td>
<td>3.023220E-03</td>
<td>0.4911230</td>
</tr>
<tr>
<td>Account Code</td>
<td>Description</td>
<td>Balance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>---------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Total receivables, loans, total current receivables</td>
<td>13,980,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Total current assets, total current assets</td>
<td>14,350,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Total current liabilities, total current liabilities</td>
<td>13,500,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Total liabilities, total liabilities</td>
<td>14,100,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Total equity, total equity</td>
<td>14,100,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Total assets, total assets</td>
<td>14,350,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: The table above represents the financial statements of a company, showing the balances of major financial accounts.*
<table>
<thead>
<tr>
<th>Rule Number</th>
<th>Condition</th>
<th>Action</th>
<th>Support</th>
<th>Confidence</th>
<th>Rule Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Net_variable_expenses_Credit, Total_long_term_payables_Credit)</td>
<td>$0.00</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>Rule 1: Net_variable_expenses_Credit, Total_long_term_payables_Credit.</td>
</tr>
<tr>
<td>2</td>
<td>(Functioning_assurance_Credit, Total_long_term_payables_Credit)</td>
<td>$0.00</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>Rule 2: Functioning_assurance_Credit, Total_long_term_payables_Credit.</td>
</tr>
</tbody>
</table>

Figure C.12: Phi extracted rules
Figure C.14: Certainty extracted rules

49 \( \text{Production related overheads Credit,} \)
\( \text{Total product fixed market expenses Debit,} \)
\( \Rightarrow (1.001-10.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

50 \( \text{Accruals and deferred income Debit,} \)
\( \text{Total liquid assets Credit,} \)
\( \Rightarrow (10.001-50.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

51 \( \text{Accruals and deferred income Credit,} \)
\( \text{Total liquid assets Debit,} \)
\( \Rightarrow (10.001-50.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

52 \( \text{Bad debts Credit,} \)
\( \text{Other product fixed expenses Credit,} \)
\( \Rightarrow (10.001-50.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

53 \( \text{Bad debts Credit,} \)
\( \text{Production related overheads Debit,} \)
\( \Rightarrow (10.001-50.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

54 \( \text{Bad debts Credit,} \)
\( \text{Functional accounts Debit,} \)
\( \Rightarrow (10.001-50.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

55 \( \text{Other product fixed expenses Credit,} \)
\( \text{Production related overheads Debit,} \)
\( \Rightarrow (10.001-50.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

56 \( \text{Other product fixed expenses Credit,} \)
\( \text{Total product fixed market expenses Debit,} \)
\( \Rightarrow (10.001-50.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

57 \( \text{Net general licence fees Credit,} \)
\( \text{Total short term payables Debit,} \)
\( \Rightarrow (50.001-100.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

58 \( \text{Functional accounts Credit,} \)
\( \text{Retained earnings Debit,} \)
\( \Rightarrow (100.001-1.000.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

59 \( \text{Other product fixed expenses Credit,} \)
\( \text{Prepayments and accruals income Credit,} \)
\( \Rightarrow (100.001-1.000.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

60 \( \text{Total medium/long term provision Credit,} \)
\( \text{Total reserves Debit,} \)
\( \Rightarrow (100.001-10.000.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

61 \( \text{Actual real gains/ losses reserve Credit,} \)
\( \text{Total reserves Debit,} \)
\( \Rightarrow (100.001-10.000.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

62 \( \text{Taxes on actual gains/losses Debit,} \)
\( \text{Total reserves Debit,} \)
\( \Rightarrow (100.001-10.000.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

63 \( \text{Actual real gains/ losses reserve Credit,} \)
\( \text{Total medium/long term provision Credit,} \)
\( \Rightarrow (100.001-10.000.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

64 \( \text{Functional assets Credit,} \)
\( \text{Total medium/long term provision Credit,} \)
\( \Rightarrow (100.001-10.000.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)

65 \( \text{Far ROA Credit,} \)
\( \text{Far ROA Debit,} \)
\( \Rightarrow (>10.000.000) \)
\( 1.327677e-05 \)
\( 1.000000 \)
## APPENDIX D

### Appendix D1: Visualization Evaluation Scenarios

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Description</th>
<th>Methods</th>
</tr>
</thead>
</table>
| 1. Evaluating environments and work practices (EWP) | The output of studies in this category are often design implications based on a more holistic understanding of current workflows and work practices, the conditions of the working environment itself, and potentially current tools in use. Questions in this scenario usually pertain to identifying a set of features that a potential visualization tool should have. For example:  
- What is the context of use of visualizations?  
- In which daily activities should the visualization tool be integrated?  
- What types of analyses should the visualization tool support?  
- What are the characteristics of the identified user group and work environments?  
- What data is currently used and what tasks are performed on it?  
- What kinds of visualizations are currently in use? How do they help to solve current tasks?  
- What challenges and usage barriers can we see for a visualization tool? | • Field Observation  
• Interviews  
• Laboratory Observation |
| 2. Evaluating visual data analysis and reasoning (VDAR) | Outputs are both quantifiable metrics such as the number of insights obtained during analysis or subjective feedback such as opinions on the quality of the data analysis experience. Even though VDAR studies may collect objective participant performance measurements, studies in this category look at how an integrated visualization tool as a whole supports the analytic process, rather than studying an interactive or visual aspects of the tool in isolation process. Questions in this scenario usually considering how a visualization tool supports:  
- Data exploration? How does it support processes aimed at seeking information, searching, filtering, and reading and extracting information?  
- Knowledge discovery? How does it support the schematization of information or the (re-)analysis of theories?  
- Hypothesis generation? How does it support hypothesis generation and interactive examination?  
- Decision making? How does it support the communication and application of analysis results? | • Case study:  
• Domain experts interacting with the visualization to answer questions. It contains interviews or surveys, automated logging to assess user performance. Interviews conducted weekly to discuss data insights and participant experience with the tool.  
• Laboratory experiments |
| 3. Evaluating communication through visualization (CTV) | Visualizations in this category have the goal or purpose to convey a message to one or more persons, in contrast to targeting focused data exploration or discovery. Their effectiveness is usually measured in terms of how effectively such a message is delivered and acquired. Studies in CTV are often interested in quantifying a tool’s quality through metrics such as learning rate, information retention and accuracy, and qualitative metrics such as interaction patterns of the way people absorb information or approach the tool. | • Controlled Experiments  
• Field Observation and Interviews: direct observation and interviews are common evaluation techniques |
Questions thus pertain to the quality with which information is acquired and the modalities with which people interact with the visualizations. Examples of questions are:

- Do people learn better and/or faster using the visualization tool?
- Is the tool helpful in explaining and communicating concepts to third parties?
- How do people interact with visualizations installed in public areas? Are they used and/or useful?
- Can useful information be extracted from a casual information visualization?

### 4. Evaluating Collaborative Data Analysis (CDA)

Evaluations in the CDA group study whether a tool allows for collaboration, collaborative analysis and/or collaborative decision making processes. Collaborative data analysis differs from single-user analysis in that a group of people share the data analysis experience and often have the goal to arrive at a joint conclusion or discovery.

For the CDA evaluation of such systems any of or a combination of the following questions may be relevant to address:

- Does the tool support effective and efficient collaborative data analysis?
- Does the tool satisfactorily support or stimulate group analysis or sensemaking?
- Does the tool support group insight? [78]
- Is social exchange around and communication about the data facilitated?
- How is the collaborative visualization system used?
- How are certain system features used during collaborative work? What are patterns of system use?
- What is the process of collaborative analysis? What are users’ requirements?

- Context of group work
- Log analysis
- Field of Laboratory Observation

### 5. Evaluating User Performance (UP)

User performance is predominantly measured in terms of objectively measurable metrics such as time and error rate, yet it is also possible to measure subjective performance such as work quality as long as the metrics can be objectively assessed. The most commonly used metrics are task completion time and task accuracy. Outputs are generally numerical values analyzed using descriptive statistics (such as mean, median, standard deviations, and confidence intervals) and modeled by such methods as Analysis Of Variance.

There are basically two types of questions:

- What are the limits of human visual perception and cognition for specific kinds of visual encoding or interaction techniques?
- How does one visualization or interaction technique compare to another as measured by human performance?

- Controlled experiments: quantitative and statistically results.
- Field Logs: analyzing the logs to draw usage statistics or single out interesting behavior.

### 6. Evaluating User Experience (UE)

Evaluation of user experience seeks to understand how people react to a visualization either in a short or a long time span. A visualization here may interchangeably be intended as an initial design sketch, a working prototype, as well as a finished product. The goal is to understand to what extent the visualization supports the intended tasks as seen from the participants’ eyes and to probe for requirements and needs. Evaluations in UE produce subjective

- Informal Evaluation. An informal user feedback evaluation is performed by demoing the visualization to a group of people, often and preferably domain experts. It is
results in that what is observed, collected, or measured is the result of subjective user responses.

The main question addressed by UE is: “what do my target users think of the visualization?” More specifically:
- What features are seen as useful?
- What features are missing?
- How can features be reworked to improve the supported work processes?
- Are there limitations of the current system which would hinder its adoption?
- Is the tool understandable and can it be learned?

7. Automated Evaluation of Visualizations (AEV)

Evaluations in the AEV group study the aspects of visualization that can be measured automatically by a computational procedure. This class of evaluation scenarios comprises all methods that employ an automatic computer-based evaluation of visualization. The results of studies in this group usually consist of a series of numbers that represent the visualization quality or efficiency.

Questions in this scenario usually pertain to the visual effectiveness or computational efficiency with which data is represented. Typical questions in this domain are:
- Is this layout algorithm faster than other state of the art techniques? Under what circumstances?
- How does the algorithm perform under different volumes of data and number of dimensions?
- What is the best arrangement of visual features in the visualization to optimize the detection of interesting patterns?
- What is the extent to which the current visualization deviates from a truthful representation of underlying data?
- What is the best ordering of visual items to speed up visual search?

The simplest kind of evaluation and it is, probably for this reason, extremely common.

- Usability Test. A usability test is carried out by observing how users perform a set of predefined tasks. For each session, the evaluators take notes of interesting observed behaviors, remarks voiced by the user, and major problems in interaction.
- Field Observation
- Questionnaire

- Algorithmic Performance Measurement
- Quality Metrics
<table>
<thead>
<tr>
<th>Paper Tags</th>
<th>EuroVis</th>
<th>InfoVis</th>
<th>IVS</th>
<th>VAST</th>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. People’s workflow, work practices</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>EWP</td>
</tr>
<tr>
<td>2. Data analysis</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>VDAR</td>
</tr>
<tr>
<td>3. Decision making</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>VDAR</td>
</tr>
<tr>
<td>4. Knowledge management</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>VDAR</td>
</tr>
<tr>
<td>5. Knowledge discovery</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>VDAR</td>
</tr>
<tr>
<td>6. Communication, learning, teaching, publishing</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>CTV</td>
</tr>
<tr>
<td>7. Causal information acquisition</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>CTV</td>
</tr>
<tr>
<td>8. Collaboration</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>CDA</td>
</tr>
<tr>
<td>9. Visualization-analytical operation</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>UP</td>
</tr>
<tr>
<td>10. Perception and cognition</td>
<td>15</td>
<td>16</td>
<td>13</td>
<td>1</td>
<td>UP</td>
</tr>
<tr>
<td>11. Usability/effectiveness</td>
<td>7</td>
<td>57</td>
<td>33</td>
<td>9</td>
<td>UP&amp;UE</td>
</tr>
<tr>
<td>12. Potential usage</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>UE</td>
</tr>
<tr>
<td>13. Adoption</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>UE</td>
</tr>
<tr>
<td>14. Algorithm performance</td>
<td>77</td>
<td>27</td>
<td>11</td>
<td>0</td>
<td>AEV</td>
</tr>
<tr>
<td>15. Algorithm quality</td>
<td>10</td>
<td>8</td>
<td>7</td>
<td>0</td>
<td>AEV</td>
</tr>
<tr>
<td>16. Proposed evaluation methodologies</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>17. Evaluation metric development</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Figure 16: Paper classified per purpose, the final scenario which were assigned [76]

Appendix D2: Relative Frequency Distribution

We evaluated the results of this tool with an alternative approach. We applied a different methodology for an unsupervised outlier identification for categorical data proposed in paper [82]. The proposed method considers number of categories inside categorical variables for outlier identification instead of distance. In our case, we considered the number of occurrences of account categories in each Value Category. Motivated also by [19][24], where distribution of the value in each account category was used to discover the “suspicious” transactions, we evaluate our results using the distribution of the account categories in each Value Category. Therefore, first, we created an item frequency plot for each Value Category. The relative frequency of a data variable is a:

\[
\text{Relative Frequency} = \frac{\text{Frequency}}{\text{Transaction Database size}}
\]

The line indicates the mean value of each item in the whole dataset. This approach provides the opportunity of inspecting the item frequency distribution of the items based on the dataset. Then, we selected a random subset of our resulted “unusual” rules and inspected whether the rules account categories fall into the margins of our definition as unusual rules: “account combinations inside journal
entries containing an unusual amount of money compared to their frequent behavior”. Figure 19 shows the relative frequent distribution of all Value Categories.
The inspected subset of the flagged rules as unusual from the section 5.6.2 is the following:
<table>
<thead>
<tr>
<th>Unusual Rules based on Interestingness Measures</th>
<th>Unusual based on the “unusual” definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production related overheads Credit, Total product fixed markt expenses Debit=&gt;1.001-10.000</td>
<td>Yes, both accounts are slightly below the average line, which means they frequently posted to other Value category</td>
</tr>
<tr>
<td>Bad debts Credit, Production related overheads Debit =&gt; 10.001-50.000</td>
<td>Not clear, Production related overheads Debit is above the average line, however Bad debts Credit is below the average line.</td>
</tr>
<tr>
<td>Total medium/long term provision Credit, Total reserves Debit=&gt;1.000.001-10.000.000</td>
<td>Yes, both accounts found below average line.</td>
</tr>
<tr>
<td>Par Acc Credit, Par Acc Debit =&gt;10.000.000</td>
<td>No, both accounts found above average line</td>
</tr>
<tr>
<td>Inventories Debit, Production related overheads Credit=&gt;100.001-1.000.000</td>
<td>Yes, Inventories Debit found much below the line, although it is one of the frequent account categories.</td>
</tr>
<tr>
<td>Net interest expense/income Debit, Total short term payables Credit =&gt;0-100</td>
<td>Yes, both account found below their average line.</td>
</tr>
<tr>
<td>Accruals and deferred income Credit, on perf trade allowance Debit =&gt;0-100</td>
<td>Not clear, Accruals and deferred income Credit is below the average line, Non perf trade allowance Debit on the line</td>
</tr>
<tr>
<td>Functional account Credit, Total medium/long term debts Debit=&gt;1.001-10.000</td>
<td>Not clear, Functional accounts credit on the average line, Total medium/long term debts Debit below the average line.</td>
</tr>
</tbody>
</table>

Table 16: Inspected rules subset

Based on the random subset that we have tested, the rules extracted are either unusual (based on our criteria) or it is not clear enough from the diagram. From the subset, only one rule had identified unusual whether both accounts were above the mean value. The rule “Par Acc Credit, Par Acc Debit =>10.000.000” had a strong positive correlation (found only in the particular Value Category) and this pair of accounts was found infrequently in the dataset. Thus, due to the criteria that we set, positive correlated and “low support high confidence”, the rule was extracted as unusual but in the real case this is its only behavior. To summarize, we tested the 12% of the results (8/63, the first 2 rules where above the maximum support threshold).The 4/8=50% of the rules was correctly extracted, 37.5% not clear, and 12.5% no correct rule extracted.
## Appendix D3: Rules from the synthetic dataset

<table>
<thead>
<tr>
<th>rules</th>
<th>support</th>
<th>confidence</th>
<th>certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Credit=D_Cr,Debit=D_De} =&gt; {Value=10.001-100.000}</td>
<td>0.005274262</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=C_Cr,Debit=D_De} =&gt; {Value=10.001-100.000}</td>
<td>0.00140647</td>
<td>0.081633</td>
<td>-0.0003205</td>
</tr>
<tr>
<td>{Credit=K_Cr,Debit=D_De} =&gt; {Value=&gt;100.001}</td>
<td>0.008438819</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=G_Cr,Debit=D_De} =&gt; {Value=1.001-10.000}</td>
<td>0.023909986</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=C_Cr,Debit=D_De} =&gt; {Value=1.001-10.000}</td>
<td>0.012658228</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=I_Cr,Debit=A_De} =&gt; {Value=&gt;100.001}</td>
<td>0.007032349</td>
<td>0.833333</td>
<td>0.78206897</td>
</tr>
<tr>
<td>{Credit=I_Cr,Debit=B_De} =&gt; {Value=1.001-10.000}</td>
<td>0.012496483</td>
<td>0.959459</td>
<td>0.94698975</td>
</tr>
<tr>
<td>{Credit=I_Cr,Debit=I_De} =&gt; {Value=&gt;100.001}</td>
<td>0.005274262</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=I_Cr,Debit=K_De} =&gt; {Value=&gt;100.001}</td>
<td>0.017229255</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=I_Cr,Debit=F_De} =&gt; {Value=&gt;100.001}</td>
<td>0.012658228</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=D_Cr,Debit=E_De} =&gt; {Value=10.001-100.000}</td>
<td>0.015827289</td>
<td>0.918367</td>
<td>0.91108263</td>
</tr>
<tr>
<td>{Credit=D_Cr,Debit=C_De} =&gt; {Value=10.001-100.000}</td>
<td>0.02151899</td>
<td>0.887324</td>
<td>0.87726898</td>
</tr>
<tr>
<td>{Credit=B_Cr,Debit=C_De} =&gt; {Value=10.001-100.000}</td>
<td>0.012658228</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=K_Cr,Debit=G_De} =&gt; {Value=1.001-10.000}</td>
<td>0.015822785</td>
<td>0.918367</td>
<td>0.89690796</td>
</tr>
<tr>
<td>{Credit=K_Cr,Debit=G_De} =&gt; {Value=0-100}</td>
<td>0.00140647</td>
<td>0.081633</td>
<td>-0.3826558</td>
</tr>
<tr>
<td>{Credit=K_Cr,Debit=C_De} =&gt; {Value=&gt;100.001}</td>
<td>0.00281294</td>
<td>0.112676</td>
<td>-0.1602525</td>
</tr>
<tr>
<td>{Credit=K_Cr,Debit=C_De} =&gt; {Value=0-100}</td>
<td>0.02151899</td>
<td>0.887324</td>
<td>0.83035961</td>
</tr>
<tr>
<td>{Credit=K_Cr,Debit=A_De} =&gt; {Value=0-100}</td>
<td>0.012658228</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=K_Cr,Debit=K_De} =&gt; {Value=1.001-10.000}</td>
<td>0.005274262</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Credit,Debit</td>
<td>Value</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>K_Cr,F_De</td>
<td>0.012658228</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D_Cr,G_De</td>
<td>0.008438819</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D_Cr,C_De</td>
<td>0.00140647</td>
<td>0.081633</td>
<td>-0.3826558</td>
</tr>
<tr>
<td>D_Cr,A_De</td>
<td>0.012658228</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D_Cr,K_De</td>
<td>0.00281294</td>
<td>0.112676</td>
<td>-0.3359181</td>
</tr>
<tr>
<td>F_Cr,E_De</td>
<td>0.00140647</td>
<td>0.081633</td>
<td>-0.2008445</td>
</tr>
<tr>
<td>F_Cr,C_De</td>
<td>0.015822785</td>
<td>0.918367</td>
<td>0.87709727</td>
</tr>
<tr>
<td>F_Cr,A_De</td>
<td>0.024613221</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F_Cr,F_De</td>
<td>0.001758087</td>
<td>0.333333</td>
<td>0.12827586</td>
</tr>
<tr>
<td>F_Cr,F_De</td>
<td>0.003516174</td>
<td>0.666667</td>
<td>0.49814717</td>
</tr>
<tr>
<td>E_Cr,E_De</td>
<td>0.005274262</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G_Cr,E_De</td>
<td>0.012658228</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A_Cr,E_De</td>
<td>0.033403657</td>
<td>0.969388</td>
<td>0.96134049</td>
</tr>
<tr>
<td>A_Cr,E_De</td>
<td>0.001054852</td>
<td>0.040541</td>
<td>-0.254576</td>
</tr>
<tr>
<td>E_Cr,G_De</td>
<td>0.024964838</td>
<td>0.959459</td>
<td>0.93896384</td>
</tr>
<tr>
<td>E_Cr,G_De</td>
<td>0.008438819</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G_Cr,G_De</td>
<td>0.012658228</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G_Cr,C_De</td>
<td>0.00140647</td>
<td>0.148148</td>
<td>-0.1138697</td>
</tr>
<tr>
<td>C_Cr,C_De</td>
<td>0.005274262</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G_Cr,C_De</td>
<td>0.015822785</td>
<td>0.918367</td>
<td>0.90520079</td>
</tr>
<tr>
<td>B_Cr,A_De</td>
<td>0.015822785</td>
<td>0.918367</td>
<td>0.90520079</td>
</tr>
<tr>
<td>B_Cr,A_De</td>
<td>0.00140647</td>
<td>0.081633</td>
<td>-0.1597854</td>
</tr>
<tr>
<td>Rule</td>
<td>Value</td>
<td>Weight 1</td>
<td>Weight 2</td>
</tr>
<tr>
<td>------------------------------</td>
<td>----------------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>{Credit=A_Cr,Debit=A_De}</td>
<td>{Value=0_100}</td>
<td>0.00351614</td>
<td>0.666667</td>
</tr>
<tr>
<td>{Credit=A_Cr,Debit=A_De}</td>
<td>{Value=101_1000}</td>
<td>0.005274262</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=C_Cr,Debit=B_De}</td>
<td>{Value=0_100}</td>
<td>0.037271449</td>
<td>0.972477</td>
</tr>
<tr>
<td>{Credit=E_Cr,Debit=K_De}</td>
<td>{Value=101_1000}</td>
<td>0.015822785</td>
<td>0.918367</td>
</tr>
<tr>
<td>{Credit=A_Cr,Debit=B_De}</td>
<td>{Value=101_1000}</td>
<td>0.033052039</td>
<td>0.969072</td>
</tr>
<tr>
<td>{Credit=E_Cr,Debit=F_De}</td>
<td>{Value=100.001}</td>
<td>0.017229255</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=G_Cr,Debit=I_De}</td>
<td>{Value=0_100}</td>
<td>0.029887482</td>
<td>0.965909</td>
</tr>
<tr>
<td>{Credit=G_Cr,Debit=F_De}</td>
<td>{Value=0_100}</td>
<td>0.024613221</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=B_Cr,Debit=I_De}</td>
<td>{Value=101_1000}</td>
<td>0.021097046</td>
<td>0.714286</td>
</tr>
<tr>
<td>{Credit=B_Cr,Debit=F_De}</td>
<td>{Value=0_100}</td>
<td>0.008438819</td>
<td>0.285714</td>
</tr>
<tr>
<td>{Credit=C_Cr,Debit=E_De}</td>
<td>{Value=0_100}</td>
<td>0.054149086</td>
<td>0.980892</td>
</tr>
<tr>
<td>{Credit=C_Cr,Debit=I_De}</td>
<td>{Value=100.001}</td>
<td>0.0053094233</td>
<td>0.961783</td>
</tr>
<tr>
<td>{Credit=C_Cr,Debit=K_De}</td>
<td>{Value=100.001}</td>
<td>0.021097046</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=C_Cr,Debit=F_De}</td>
<td>{Value=0_100}</td>
<td>0.008438819</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=C_Cr,Debit=K_De}</td>
<td>{Value=100.001}</td>
<td>0.0201246321</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=E_Cr,Debit=F_De}</td>
<td>{Value=100.001}</td>
<td>0.001054852</td>
<td>0.223814936</td>
</tr>
<tr>
<td>{Credit=A_Cr,Debit=K_De}</td>
<td>{Value=100.001}</td>
<td>0.021097046</td>
<td>1</td>
</tr>
<tr>
<td>{Credit=A_Cr,Debit=K_De}</td>
<td>{Value=100.001}</td>
<td>0.001054852</td>
<td>0.223814936</td>
</tr>
<tr>
<td>{Credit=A_Cr,Debit=F_De}</td>
<td>{Value=100.001}</td>
<td>0.021097046</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 17: All rules from the synthetic dataset
Appendix D4: Certainty objective measure for negative correlated rules

Figure D.1: Certainty extraction plot for the synthetic dataset