Data Mining: a Potential Research Approach for Information System Research

A Case Study in Business Intelligence and Corporate Performance Management Research

Karin Hartl, Olaf Jacob
Department of Information Management
University of Applied Sciences Neu-Ulm (HNU)
Neu-Ulm, Germany
karin.hartl@hs-neu-ulm.de, olaf.jacob@hs-neu-ulm.de

***Abstract*—This paper investigates the opportunities of Data Mining applications for Information System research. Data Mining is a data driven statistical approach for knowledge discovery. Hypotheses and models do not have to be developed at the beginning of the research, which allows the detection of new and otherwise undiscovered patterns in a given dataset. Consequently, current challenges in Information System research can be investigated from a different angle. The Data Mining results may provide additional, surprising and detailed insights to an Information System problem. To prove these assumptions, this study applies Association Rule Discovery and cluster analysis to a questionnaire based data set. This data set has been collected to investigate the relationship between Business Intelligence and Corporate Performance Management. Even though this relationship has been explored at several stages in the Information System research literature, the results are often short on detail. This paper explores, if Data Mining methods can provide additional information to the subject. Both of the applied Data Mining methods provide promising results. Association rules and clusters have been identified, providing a different view on the connection between Business Intelligence and Corporate Performance Management. Therefore, Data Mining techniques offer an option to reuse questionnaire-based data and to gain new insights in Information System research.**

Keywords-Information Systems; Data Mining; Association Rule Discovery; Cluster Analysis; Business Intelligence.

I. INTRODUCTION

This research discusses the potentials of explorative Data Mining techniques for Information System (IS) related research and is the extended version of the previously published work of Hartl and Jacob [1] at the Data Analytics Conference 2016 in Venice.

In IS research, Explanatory Factor Analysis (EFA) and Structural Equation Modelling (SEM) are the commonly used research approaches. Before conducting any analysis, research assumptions and hypotheses are developed and afterwards confirmed with data collected for this specific research purpose. This approach has one major limitation, its reliance on human imagination for generating research theories and assumptions [2]. The theories and assumptions are typically based on findings and research results already accomplished in the field. To prove the pre-defined research assumptions, data sets are collected. These data sets may hold even more information regarding a research subject than anticipated and analysed. Nevertheless, non-logical connections are generally not investigated.

With Data Mining methods, interesting information and patterns can be discovered from various data types [3]. Instead of testing previously defined hypotheses, Data Mining applications are working up from the data [4]. This opens the opportunity to detect non-anticipated connections and hidden information in a given data set.

This research is a first approach in exploring the potentials of Data Mining methods for Information System research subjects [1]. Two descriptive Data Mining techniques are applied to a questionnaire-based data set, exploring the impact of Business Intelligence (BI) on Corporate Performance Management (CPM) [5]. Aim of the questionnaire was to make the business value of BI tangible. CPM is a suitable concept to explore the business value of BI, since a successful CPM requires data and up-to-the-minute information [1][5][6]. In recent researches, the business value of BI for CPM has been explored by extracting assumed connections from the literature and a subsequent testing of these connections with data collected from industry. Several pre-defined connections and interdependencies between BI and CPM have been proven. However, these results are often generic and lack detail.

As a result, the above mentioned research subject presents an interesting case study to investigate, if Data Mining techniques provide more insights to IS research subjects.

Association Rule Discovery and cluster analysis are renowned Data Mining methods and therefore have been identified as a suitable first approach in exploring the value Data Mining has for IS research. Association Rule Discovery searches for structural connections in a data set, formulates If-Then-Statements and can consider all available research criteria. For the presented case study, the results of the association analysis could allow conclusions on the specific BI capabilities accounting for a successful CPM. This may allow researchers and practitioners to focus on the most important BI features in order to improve CPM.

Cluster analysis explores the similarities between cases in a data set. The case study results could facilitate a better understanding of the connection between BI and CPM. Besides analysing the research variables – e.g., BI and CPM characteristics – cluster analysis facilitates the consideration of descriptive items - e.g., annual turnover, company size. This makes conclusions on the actual development of BI and CPM in German companies possible. Furthermore, the results

could reveal whether companies with a well-established BI solution likewise have a well-functioning CPM. In addition, the findings may allow conclusions on the impact, the successful use and implementation of BI systems has on a company's CPM.

The paper is structured as follows. Section II discusses the motivation for this research. It points out, why IS research can profit from Data Mining applications. In Section III, the research approach is discussed. Section IV describes the case study. First, the research background is introduced. The terms BI and CPM are described and the relationship between these two areas is discussed. An investigation of the subject related research highlights the opportunities of the Data Mining approach. Second, the data selection and pre-procession is described for both Data Mining analysis – FP-Growth and k-means clustering. Third, the analysis process is described, before the results are presented. Section IV closes with a discussion on the case study results and on how they can be useful in practice. In Section V, it is discussed if Data Mining presents itself as a suitable research method for IS research topics. Section VI points out the next steps and future research opportunities.

II. Motivation for the Data Mining Approach

In the IS research field regarding the connection between BI and CPM first EFA and second PLS-SEM are the commonly used approaches. The starting point of these two analysis is always an empirically provable theory, which is developed based on assumptions [7]. Afterwards, hypotheses and a theoretical model are developed. To empirical investigate the proposed research model, typically, a questionnaire is developed and real-life data collected. Subsequent, these data are organized by an EFA analysis. The EFA groups correlating items and joins them together in a factor [8]. Data can be structured and reduced this way.

Next, the structured data are analysed with the PLS-SEM method by seeking the optimal predictive linear relationship to assess the previously defined causal relationship [1][9][10]. The characteristics (items) evaluated in the questionnaire built the measurement construct of the PLS model [10]. Mainly reflective PLS models are used in IS research. This means that the measurement model, also called the outer model, is caused by the construct [11]. The measurement items are then interchangeable and generally, a further investigation of the connections between the separate items is missing.

The creation of factors for compacting information might be the right approach for many research subjects, but it must not be the only correct approach to explore connections in IS research. It is assumed that Data Mining can highly contribute to the subject. Data Mining methods can include all available research criteria and has no need for compacting questionnaire data. This may lead to research result that are more detailed.

Data Mining can be understood as an extension of statistical data analysis and statistical approaches [12]. Both approaches aim to discover structure in data, but Data Mining methods are generally robust to non-linear data, complex relationships and non-normal distributions [13]. Data Mining is a data driven approach and supports the discovery of new and sometimes unexpected knowledge [2]. Instead of only testing assumed hypotheses, with Data Mining otherwise undiscovered data attributes, trends and patterns can be explored [14]. Especially with explanatory Data Mining techniques, a good understanding of connections in the data set can be achieved [15].

Although Data Mining is often only considered suitable for large data sets, Natek and Zwilling [16] illustrate in their research that small data sets are not limiting the use of the tool. They even applied a predictive Data Mining model to a relatively small data set. Prediction needs the division of the data set into a training set and a testing set. For exploratory Data Mining methods, this division is not necessary. All available data cases can be part of the analysis. Therefore, exploratory Data Mining methods should be applicable to small data sets.

III. Research Approach

The Data Mining literature describes various approaches to Data Mining problems [3][15]. The Cross Industry Standard Process for Data Mining (CRISP-DM) is a well-known procedure and the methodology chosen for this research [17]. As the data set is comparatively small, no selection of the appropriate data set was considered necessary. Therefore, the starting point for the data analysis was the pre-processing of the data. Fig. 1 shows the steps followed in the case study.

Each Data Mining analysis is conducted to answer a specific research question. This research question is then the basis for the chosen data and analysis method. In general, this research asks for the information gain in IS research through the application of Data Mining methods.

Exploratory Data Mining methods analyse given data and extract information and patterns from this data. In particular, association analysis allows to take all available items of a data set into account and to identify co-occurrence relationships on item basis. Cluster analysis groups the cases in the data set according to their similarity. The results identify structures in the data, which allow conclusions on dependencies.

In the data selection phase, the questionnaire data has been evaluated and missing values identified. Afterwards, the data has been pre-processed by calculating the missing values and addressing conflicting values. Han et al. [3] and Cleve and Lämmel [15] suggest alternatives for dealing with missing values, depending on the data structure. The important items of the questionnaire used for the case study are formatted as Likert scale items and can be interpreted as metric data. Metric data can be pre-processed by replacing the missing values in the sample by the mean value of all item-based compiled answers. Alternatively, the mean values can be stated by contemplating the data case closest to the data case with the missing value. This idea follows the k-nearest neighbours (kNN) approach. Joenssen and Müllerleite [18] assess the kNN approach as practicable imputation method for missing Likert scale values in small data sets. Therefore, the missing values in the data set were imputed using the kNN approach.

After dealing with the missing values, the data needs to be transformed in the required format for the applicable Data Mining technique. The applied Association Rule Discovery algorithm - FP-Growth - needs binary data [15].

Consequently, the Likert scale items have been transformed into binary variables. This has been directly done in the RapidMiner Data Mining tool. The information loss created by transforming the data into binary variables has been accepted at this point of the research. The goal was to apply Data Mining techniques to a questionnaire based data set for the first time in IS research. A wide range of IS researchers should understand the results. Therefore, the results ought to be elementary and understandable.

The results are then evaluated and interpreted. As in every research, not all findings are valuable and of real-life meaning. Accordingly, interpretation and evaluation presuppose a subject knowledge background to ensure that only sensible research results are discussed and interpreted.

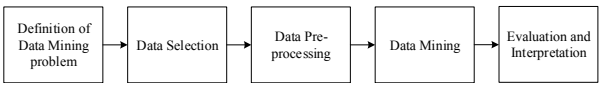

Figure 1. Research Approach (based on CRISP-DM) [17].

IV. CASE STUDY

The case study focuses on a currend IS research topic which investigates the relationship between BI and CPM. The aim of the initial research was to investigate how the business value of BI can be made tangible [5]. Therefore, a context model has been identified by conducting a in-depth literature study. To pove the context model, data has been collected with the help of a questionairre. This data set and this research study background are used as the basis for the following case study.

A. The Relationship Between BI and CPM

Nowadays, companies have to face a more challenging and continuously changing environment each day. Especially, globalization intensifies the competition and the struggle for success and existence. Additionally, the digitalization confronts companies with immense amounts of data. Transforming these data into information and using these information for the management of a company are assumed to retain a company's survival in these challenging times. BI is a process including applications for storing data and systems as well as methods for analysing these data and the business environment. The usage of BI promises companies support in their decision making process by acquiring, analysing and disseminating information from data significant to the business activities [4][19][20]. Consequently, BI is a source for data of high quality and actionable information. This indicates that the proper use and application of BI systems supports the successful management of companies [5].

IT investments are necessary, but since there are increasing continuously their return on investment is evaluated critically in companies. Therefore, BI projects and implementations need to be justified. Accordingly, the measurement of the BI business value is an important topic [21]. But capturing the business value of BI is a strategic challenge, due to the diverse nature of BI [22]. Generally, BI systems are not strictly amortized by saving costs after implementation. Instead, the main BI benefits are of an intangible nature and therefore hard to measure [21]. Williams and Williams [22] see the BI business value in the usage of the system and the contained data within the management processes of a company. Miranda [23] then put BI and CPM into context and identified CPM as the appropriate framework to prove the value proposition and benefits of BI.

CPM is defined by the Gartner Group as "an umbrella term that describes all processes, methodologies, metrics and systems needed to measure and manage the performance of an organization" [24]. Therefore, CPM presents the strategic deployment of the BI solutions and is born out of a company's need to proactively manage business performance [1][25]. Inferentially, CPM needs BI to work effectively on accurate, timely and high quality data and BI needs CPM for a purposeful commitment [1][24]. As a consequence, it is expected that the effectiveness of CPM increases with the effectiveness of the BI solution and therefore the company success improves as well [1][6].

B. Subject Related Research

The relationship between performance management and BI has been investigated in several studies during the last couple of years.

Williams and Williams [22] are one of the first to expose the necessity to investigate the usage of BI within a company, for the purpose of exploring and measuring the business value of BI. The authors show that the value created through the implementation and usage of BI is to be found particularly within the management processes of a company, which affect the operational processes. Additionally, the research suggest that the return on BI investment can be measured on the increased revenues and reduced costs within a company [1][22]. However, the value created through successful BI solutions is more than just monetary benefits. BI is a complex process and the quest to make the business value of BI tangible needs an in-depth evaluation of the connection between BI and a company's management processes.

Miranda [23] suggested that CPM is an appropriate framework for BI applications. The authors describe the concept of CPM as a business management approach using business analysis to support the success and management of a company. Accordingly, CPM presents itself as a suitable framework to explore the business value of BI. Although, the research of Miranda [23] does not provide an empirical investigation of the subject, the article can be considered as the foundation for more detailed research in the field, including the following.

The connection between BI and CPM has been of research interest and empirically investigated mainly within the past 10 years. The differences and similarities between BI and CPM have been discussed by Aho [19] in form of a literature study and an action oriented research. The results support the conclusions of Miranda [23] and point out once more that BI and CPM need to be connected for an effective and efficient application. However, the empirical background does not deliver details on the configuration of the relationship between BI and CPM.

Yogev et al. [26] explore the business value gained through BI using a process-oriented approach. In the research model, key BI resources and capabilities are identified, which can explain the value created through the implementation and usage of a BI system. A hypotheses based theoretical model has been formulated and tested by the authors using EFA and SEM. The results demonstrate positive effects of BI on the strategic and the operational company level. Nevertheless, the empiricism does not provide any details about the BI related resources creating this positive effect.

Saed [27] investigates the relationship between BI and business success using regression and correlation analysis. First, hypotheses have been developed based on a literature review. Second, data has been collected and descriptively evaluated before correlation and regression analysis have been applied for hypotheses evaluation. Some of the hypotheses could be confirmed, but while these statistical techniques provide room for detailed results, only casual explanations have been provided [1].

Richards et al. [6] explore the connection between BI and CPM using EFA and PLS-SEM. Literature based hypotheses and a research framework have been developed. The framework supposes that BI directly influences and supports measurement, planning and analytics. The effectiveness of planning, measurement and analytics, again, influences the effectiveness of the company's processes. Through a large-scale survey, sample data has been collected. The number of variables in the questionnaire has then been reduces by applying an EFA. Afterwards, the factors have been converted into latent variables. The connections between these variables have then been evaluated with a Confirmatory Factor Analysis (CFA) using SmartPLS. Three of the seven hypotheses have been confirmed. Consequently, the research identifies a direction on how BI influences CPM, but the specific mechanisms who do so are not discussed or defined.

Hartl et al. [5] also developed a hypotheses based research framework. The framework assumed detailed relationships between BI and CPM. They expected that data quality and provision as well as pre-defined data analysis on the BI side of a company have a positive impact on the existence of closed-loop business processes on the CPM side of a company. Furthermore, technical data and method integration on the BI side of a firm is believed to have a positive impact on organizational alignment within the CPM. Eventually, the usage of extended collaborative and analytical functions is positively influencing CPM process effectiveness and process efficiency. All of these hypotheses have been confirmed, using an EFA first and a PLS-SEM analysis with SmartPLS second. Nevertheless, the detailed characteristics collected in the questionnaire regarding the development of BI and CPM in a company have only been used as the measurement construct of the PLS model. More detailed information contained in the data is not investigated.

This research project complements the subject related work and uses the questionnaire based data of Hartl et al. [5]. Instead of proving a pre-defined construct and compacting information from the collected data, this research considers all the research criteria. The aim is to get detailed insights on the relationship between BI and CPM on questionnaire item basis. This allows the identification of specific BI characteristics, which support a successful CPM. A first approach has been presented in the research of Hartl and Jacob [1]. The authors present an association rule analysis, which is extended in this research.

The aim of this case study is once more, to identify detailed information on the connection between BI and CPM. Exploratory Data Mining techniques are the used approach. As an alternative to SEM and grouping the questionnaire characteristics and measurement items describing BI and CPM together, all items are considered separately. It is supposed that this identifies patterns and structures that can explain the relationship between BI and CPM in more detail.

C. Defintion of the Data Mining Problem

Two main problems have been identified regarding the research project. Challenge number one is to identify the BI characteristics, which have a strong influence on CPM. Vice versa, it would be interesting to identify the CPM characteristics, which are strongly connected to BI.

The second challenge is to identify if companies with a high BI development also have a high CPM development. Additionally, it would be interesting to identify extra factors – e.g., company size, BI provider used –, which influence the development of CPM and/or BI in German companies.

Therefore, the following two research questions are defined:

> Research question 1: Which specific BI characteristics are most influential on CPM and which specific CPM characteristics are most affected by BI?
>
> Research question 2: Is a high development of BI related to a high development of CPM and are their certain characteristics that support a high development in both?

According to the defined research questions, two Data Mining methods have been identified for analyses – Association Rule Discovery and cluster analysis.

D. Data Selection

This research is based on the findings of Hartl et al. [5], where a set of criteria that is seen as suitable to represent CPM on one hand, and BI on the other hand, has been identified (Table I). A study has been conducted in Germany, to bring the criteria of both fields together and to clarify the relationship between BI and CPM. Therefore, the identified criteria have been transformed into questionnaire items, which had to be answered on a five-point Likert scale. The anchor points at the ends of the scale have been "does not apply" and "fully applies" and an additional definition "applies half and half" for the mid stage has been defined. The data collection has taken place from December 2014 until March 2015 using telephone interviews and an online questionnaire. Subjects were German companies who use BI for supporting their performance management. For this reason, decision makers from management, controlling and IT were addressed. In total 169 questionnaires were completed resulting in a response rate of 11.3%. The participating companies are mainly mid-sized

TABLE I: OVERVIEW OF THE BI AND CPM ITEMS [5]

BI items	CPM items
BI1_1: Clear roles and responsibilities for operating the BI systems	CPM1_1: Business management processes are transparent and traceable for managers
BI1_2: Data consistency ("Single Version of the Truth")	CPM1_2: Business management process are documented throughout the company
BI1_3: 24/7 operation of the BI systems	CPM1_3: Business management processes are communicated throughout the company
BI1_4: Only compulsory BI tools are used	CPM2_1: Business management processes base on a common database
BI1_5: Data integrity during simultaneous use	CPM2_2: Management methods are fully automated and linked without manual support
BI1_6: Clear roles and responsibilities for the BI-development between the company's departments and the IT throughout the whole enterprise	CPM2_3: Data in business management processes are complete
BI1_7: BI-architecture is documented	CPM2_4: Decision makers manual expenditure to edit reports is marginal
BI1_8: Master data changes are traceable	CPM3_1: Data in business management processes are relevant
BI1_9: BI relevant master data can be saved in various versions	CPM3_2: Data in business management processes are current
BI2_1: Use of feature set for predictive forecasting	CPM3_3: Effective use of external data (market data)
BI2_2: Use of feature set for describing data analysis	CPM4_1: Alignment of business management processes across all business functions
BI2_3: Use of feature set for information visualization	CPM4_2: Alignment of business management processes across all business units
BI3_1: Use of applications for scenario modelling	CPM4_3: Alignment of strategic and operational planning
BI3_2: Use of applications for statistical analysis	CPM5_1: Use of measurable indicators in all business functions
BI4_1: Each BI project is carried out using a standardized procedure model	CPM5_2: Use of measurable indicators in all business units
BI4_2: Each BI project bases on a standardized design method	CPM5_3: Use of measurable indicators in all operational business processes
BI4_3: Documentation standards for BI projects are clearly defined	CPM5_4: Use of measurable indicators in all strategic business processes
BI4_4: BI projects use agility	CPM6_1: Existence of feedback loops in operational business processes (e.g., complaint management)
BI5_1: Use of applications for adding describing comments	CPM6_2: Existence of feedback loops in strategy development (adjustment of vision, mission and the company's strategy to environmental changes)
BI5_2: Use of applications for sharing comments throughout the enterprise	CPM6_3: Existence of feedback loops in strategic planning processes
BI5_3: Use of applications for automatic text processing and Text Mining	
BI6_1: Denotations and spellings are standardized in the BI databases	
BI6_2: BI tools for strategic business management are interoperable	
BI6_3: Manual expenditures for ensuring standardized spelling and denotations are marginal	
BI7_1: Applications for mobile usage of the BI Systems are available	
BI7_2: Applications for the mobile usage of the BI Systems are used	
BI8_1: Use of BI applications for implementing alerts linked to automated workflow data in operational business processes	
BI8_2: Use of BI applications for implementing alerts linked to automated workflow data in strategic business processes	

E. Data Pre-processing and Data Mining

Before applying the analysis, the data has been screened for outliers and missing values. In total, the proportion of missing values is at 12%. As the dataset contains of many items and the missing values are balanced across the data set, it was decided to impute the missing values using the kNN approach. The procedure described in Section 3 has directly been implemented in RapidMiner. Additionally, no severe outliers have been detected and all cases in the data set could be included in the analysis.

1) Association Analysis [1]: Association Rule Discovery is a popular pattern discovery method [2]. With association rules, co-occurrence relationships between data items can be discovered, taking into account as many research items as needed and available [15]. This indeed can lead on the upside to results that are more detailed and on the downside to an enormous amount of discovered association rules. Unmanageable amounts of association rules easily can be organized by instating measures to evaluate and select rules based on their potential interestingness for the researcher [2]. These interestingness measures include *Lift*, *Support* and *Confidence* [15][28].

To generate association rules, many algorithms are available. The FP-Growth algorithm is the classic procedure used in RapidMiner. The algorithm works in two main steps [3]. In the first step, an FP-tree is generated. The FP-tree has a root node, which is usually marked by Null. Then, a separate node is built for each item. The algorithm calculates the relative frequency of the occurrence of the items in the data set. Afterwards, all cases are viewed and the items are ordered

in a tree structure. In the second step, frequent item sets are directly extracted from the FP-tree. For each item a separate FP-tree is generated, which is evaluated from the leaves towards the tree root. Only items meeting the previously defined minimum *Support* are then recapped as a rule (please refer to Han et al. [3], pp. 257 for a detailed graphical description of the FP-Growth algorithm). After generating the frequent item set, association rules can be generated through a Rapid Miner operator. The overall rule interestingness is measured through the *Confidence* measure [15].

The FP-Growth algorithm needs binary data. Therefore, the data has to be transformed into binary variables. RapidMiner can do this transformation directly. The questionnaire characteristics "does not apply" to "applies half and half" (1-3) have been transformed to *does not apply* and "does apply" and "fully applies" (4-5) to *does apply*.

Furthermore, the minimum levels for the interestingness measures have been defined. Only association rules (X→Y) with a minimum *Support*≥0.6 have been considered as interesting. This means that in at least 60% of all the cases in the data set the rule has to show [28]. The confidence level has been set at *Confidence*≥0.7. This determines that in at least 70% of the cases in the data set where the first part of the rule (X) is shown, the second part of the rule (Y) has to show as well [16]. The measure *Lift* needs to be *Lift*>1 to indicate a positive correlation between the items of a rule [3]. Regarding the minimum settings of the measures, 103 association rules have been discovered.

Association rules do not imply causality. They find items that imply the presence of other items [2]. As the research focus is on the benefits of BI for CPM, the attention lies on association rules beginning with BI items, leaving 52 association rules for evaluation. The association rules with the highest *Support* are shown in Table II.

It is conspicuous that especially the CPM items *Data in business management processes are relevant* and *Data in business management processes are current* apply in a company, if specific BI items apply too. In more detail, these two CPM items most likely apply in a company, if in addition to the items in Table II the following BI items apply as well:

- *Data consistency ("Single Version of the Truth"),*
- *Only compulsory BI tools are used,*
- *Master data changes are traceable,*
- *Clear roles and responsibilities for the BI-development between the company's departments and the IT throughout the whole enterprise.*

In addition, the BI item *Use of applications for automatic text processing and Text Mining* is found in combination rules with the item *Clear roles and responsibilities for operating with the BI system* and *Data integrity during simultaneous use* (Table III). Different from these two, *Use of applications for automatic text processing and Text Mining* has the characteristic does not apply. Still, the CPM items *Data in business management processes are relevant* and *Data in business management processes are current* apply, indicating that data currency and relevance is not influences by the usage of Text Mining and context processing tools.

Furthermore, the association rules illustrate that:

- *Use of BI applications for implementing alerts linked to automated workflow data in strategic business processes* does not apply,
- *Use of applications for sharing comments throughout the enterprise* does not apply,
- *Use of applications for adding describing comments* does not apply,
- *Use of applications for automatic text processing and Text Mining* does not apply

TABLE II: STRONGEST ASSOCIATION RULES (RULE BODY CONTAINS OF BI ITEMS ONLY) [1]

BI items		**CPM items**	**Interestingness**
Data integrity during simultaneous use=**applies**	→	Data in business management processes are relevant=**applies**	Support=0.83 Confidence=0.92
Clear roles and responsibilities for operating the BI systems=**applies**	→	Data in business management processes are relevant=**applies**	Support=0.80 Confidence=0.93
Data integrity during simultaneous use=**applies**	→	Data in business management processes are current=**applies**	Support=0.78 Confidence=0.86
Data integrity during simultaneous use=**applies** *AND* Clear roles and responsibilities for operating the BI systems=**applies**	→	Data in business management processes are relevant=**applies**	Support=0.76 Confidence=0.93
Data integrity during simultaneous use=**applies**	→	Data in business management processes are relevant=**applies** *AND* Data in business management processes are current =**applies**	Support=0.74 Confidence=0.83
Clear roles and responsibilities for operating the BI systems=**applies**	→	Data in business management processes are current=**applies**	Support=0.74 Confidence=0.86
Clear roles and responsibilities for operating the BI systems=**applies**	→	Data in business management processes are relevant=**applies** *AND* Data in business management processes are current =**applies**	Support=0.73 Confidence=0.84
24/7 operation of the BI systems=**applies**	→	Data in business management processes are relevant=**applies**	Support=0.70 Confidence=0.91

the CPM item *Management methods are fully automated and linked without manual support* does not apply as well (Table III).

2) Cluster Analysis: Clustering attempts to find patterns and groups in research criteria [2]. The data are organized without previous knowledge of potential groups and are arranged by means of their similarity [15]. The objects belonging to one group are as much as possible homogenous [15]. The groups, however, are as heterogeneous as possible [15]. In clustering, all attributes available can be used in parallel. This offers a detailed view of the cluster features and enables a thorough view on the relations between BI and CPM. Clustering can be done by defining a similarity and distance measure, which is also known as proximity measure. Interesting results regarding the research data are believed to be accomplished by using the k-means algorithm as it is a well-known partitioning algorithm [29]. The algorithm works in 5 main steps [3][30].

1. From a data set k ojects are identified as the centre of k clusters. Each k object is one data case (k is the number of clusters to be estracted from a data set).
2. Every other data case from the dataset is then assigned to the k object - also calles cluster centre - it is most similar to. This can be measured based on the Euclidean distance between the data case and the cluster mean (for further details on the Eucledian Distance or alternative proximity measures, please refer to Han et al. [3] or Cleve and Lämmel [15]).
3. After assigning each data case to one cluster, a new cluster centre is calculated.
4. All data cases are then reassignes to one of the new cluster centres and new clusters are built.
5. The above process is continued until there are no more changes in the cluster centres and the compilation of the clusters.

The number of clusters k is to be chosen beforehand by the researcher.

In the case study, the k-means algorithm has been applied using k=2, k=3 and k=4, resulting in the most interesting output using k=3. Regarding, the data set has been divided into 3 clusters – Cluster 1, Cluster 2 and Cluster 3. An extract of the results is visually displayed in Fig. 2. The horizontal axes of the graph shows the items evaluated in the questionnaire. The vertical axes shows the encoded answers to these questions. For the Likert scale formatted BI and CPM items, number 1 is encoded as "does not apply", 3 as "applies half and half" and 5 as "totally applies". The numbers in between stand for middle stages. For the supporting questions, each number encodes an answering option. Table IV explains the codes and their meaning in detail.

Cluster 1 contains of 43 cases. The status and development in CPM and BI is assessed as mainly positive. Besides the characteristic *Management methods are fully automated and linked without manual support (CPM2_2)* CPM is weighed as continuously well developed and only the BI areas

TABLE III: SECOND AND THIRS SET OF STRONG ASSOCIATION RULES (RULE BODY CONTAINS OF BI ITEMS ONLY)

Second Set of Association Rules			
BI items		**CPM items**	**Interestingness**
Use of BI applications for implementing alerts linked to automated workflow data in strategic business processes=**does not apply**	→	Management methods are fully automated and linked without manual support=**does not apply**	Support=0.64 Confidence=0.85
Use of applications for sharing comments throughout the enterprise=**does not apply**	→	Management methods are fully automated and linked without manual support=**does not apply**	Support=0.63 Confidence=0.85
Use of applications for adding describing comments=**does not apply**	→	Management methods are fully automated and linked without manual support=**does not apply**	Support=0.60 Confidence=0.85
Use of applications for aromatic text processing and Text Mining=**does not apply** *AND* Use of applications for sharing comments throughout the enterprise=**does not apply**	→	Management methods are fully automated and linked without manual support=**does not apply**	Support=0.60 Confidence=0.86
Third Set of Association Rules			
BI items		**CPM items**	**Interestingness**
Clear roles and responsibilities operating the BI system=**applies** *AND* Use of applications for automated text processing and Text Mining=**does not apply**	→	Data in business management processes are relevant=**applies**	Support=0.68 Confidence=0.92
Clear roles and responsibilities for operating the BI systems=**applies** *AND* Data integrity during simultaneous use =**applied** AND Use of applications for automated text processing and Text Mining=**does not apply**	→	Data in business management processes are relevant=**applies**	Support=0.64 Confidence=0.92
Clear roles and responsibilities operating the BI system=**applies** *AND* Use of applications for automated text processing and Text Mining=**does not apply**	→	Data in business management processes are relevant=**applies** *AND* Data in business management processes are current =**applies**	Support=0.61 Confidence=0.82

- *Use of applications for adding describing comments (BI5_1),*
- *Use of applications for sharing comments throughout the enterprise (BI5_2)* and
- *Use of applications for automatic text processing and Text Mining (BI5_3)*

are evaluated with a critical tendency. Over 55% of the

TABLE IV: CODING OF THE SUPPORTING VARIABLES

Pos	Position in the company	1 – Management 2 – Middle Management 3 – Lower Management
Ber	Area of work in the company	1 – Company Management 2 – Managerial Accounting 3 – Financial Accounting 4 – Sales and Distribution 5 – Information Technology
Um	Company revenue	1 – Below 10 million € 2 – Between 10 and 50 million € 3 – Between 50 and 125 million € 4 – Between 125 and 500 million € 5 – Above 500 million €
Ma	Number of full-time Employees	1 – Below 50 Employees 2 – Between 50 and 250 Employees 3 – Between 250 and 1000 Employees 4 – Between 1000 and 5000 Employees 5 – Above 5000 Employees
ERP	Throughout the company, ERP software from the same provider	1 – Yes 2 – No
BI	Throughout the company, BI software from the same provider	1 – Yes 2 - No

questioned companies in the cluster stated to have above 1000 full-time employees. Almost half of these 55% even have above 5000 full-time employees. The annual turnover is exceeding 125 million Euro. The ERP and BI software used in the companies is primarily from the same producer and over 70% of the cases only use BI software from the same producer across the whole enterprise.

Cluster 2 can be described as the least developed in both BI and CPM. It contains of 36 cases. The CPM items *Data in business management processes are current (CPM3_1)* and *Data in business management processes are relevant (CPM3_2)* have a positive tendency in their development. The other CPM items are rather not distinctive. The same can be found in regards to the BI items. Only the items *Clear roles and responsibilities in operating with the BI Systems (BI1_1)*, *24/7 operation of the BI Systems (BI1_3)* and *Data integrity during simultaneous use (BI1_5)* are positively distinctive in the sample. The cases of this cluster comprise companies of all sizes. The same can be said for the annual revenue:

- 22% of the cases have an annual revenue below 125 million Euros,
- 33% of the cases have an annual turnover between 125 million - 500 million Euro,
- 33% have an annual revenue above 500 million Euro and
- 11% did not specify their annual turnover.

Predominantly, this cluster shows non-uniform BI software throughout the company. The ERP and BI software used throughout the company are mainly from different producers as well.

Cluster 3 consists of 90 cases. The situation and development of BI and CPM in the companies is assessed more critically than in Cluster 1 but more positively than in Cluster 2 (Figure 2). The items are continuously ranked as medium developed, but with a positive tendency. Especially the first couple of BI items (item *BI1_1* until *BI1_9*), which can be grouped under the topic Data Quality and Provision

Figure 2. 3-Cluster solution using k-means clustring in RapidMiner

have an obvious positive tendency in their distinction. The BI items measuring

- Flexible Modelling and Analysis (items *BI3_1* and *BI3_2*),
- Rule-based implementation of BI-projects (items *BI4_1* to *BI4_4*),
- Information enrichment through unstructured Data (item *BI5_1* till item *BI5_3*) and
- Event-driven flow of Information (items *BI8_1* and *BI8_2*)

are weighted below *applies half and half*. They can be identified as the BI problem areas in Cluster 3. In CPM, an obvious problem area is the item *Management methods are fully automated and linked without manual support (CPM2_2)*. The cluster concentrates together medium-sized and big companies, with more than 70% of them having between 500 and 5000 employees. The enterprises in the cluster mainly use the same BI software throughout the company, but almost 65% specified that their ERP systems and their BI systems are not from the same producer.

F. Result Evaluation and Interpretation

CPM is a management strategy for decision support [31]. This support is achieved by using measures and Key Performance Indicators (KPI's) from data. Decision makers and managers use the information gained from data to monitor the companies target achievements. If necessary, the strategy, the processes and the goals are adjusted to ensure the company's survival and success.

1) Association Analysis [1]: Data is the quintessence in CPM but only useful if provided when needed and of high quality. The association rules show that if BI items related to the subject of data quality and data provision (e.g., *Data consistency*, *Data integrity during simultaneous use*) are well established in a company the *Data in the business management processes are relevant* and *current*. The rules illustrate the connection between data and the business management processes and therefore the connection between BI and CPM. Supported with high quality data, decision makers can act and rely on actionable information to manage the enterprise. The rules underline the function of BI as a decision support tool needed for a successful CPM. The concentration on business management processes as the CPM part of the rule highlights the understanding of CPM as a multiplicity of business management processes connected and integrated into each other [30]. If the processes in a company are managed, based on needed high quality data, it is the initial point for an overall effective CPM.

Nevertheless, a CPM strategy is not implemented in one run. The implementation is a slow process carried out in sub-steps [31]. The focus of the association rules on *Data in business management processes is relevant* and *current* supports this. The rules indicate that initially the attention has to lie on each management processes separately. Once a company is working on high quality data when needed, a good connection of the management processes throughout the enterprise is possible. The lack of association rules containing further CPM items might be an indicator for companies in Germany still working on implementing a thorough performance management.

The second set of association rules discovered that if no opportunity to use and share comments within an enterprise and no opportunities to use unstructured data are given, a full automation and linkage of the company's management methods without manual support is not given either. Management methods are ideally accepted process descriptions for dealing with certain issues (e.g., Balanced Scorecard) [31]. These methods can only be successful if goal-oriented, understood and used continuously [31]. Consequently, management methods need defined measures and data analysis. For all measures to be useful, a reference magnitude is needed, which can be supplied by adding and sharing comments. Furthermore, the association rules imply that fully automated management and planning methods are dependent on the use of comments for ensuring transparency as well. Only if supported by describing comments, automated management processes and planning methods are understandable throughout a company and the need for manual support is minimalized.

Text Mining enables knowledge discovery from semi-structured or unstructured data. This is a rather advanced analysis method of BI and the rules indicate that if there is no or not much *usage of automatic text processing and Text Mining, Data in business management processes* are still *relevant* and *current* but the *Management methods* are not fully *automated or linked without manual support*. Text Mining is an advanced research method used to gain new information from texts. The association rules suggest that this feature of BI is not important for data currency and relevance in business management processes. Therefore, it might to be ignored in the establishment process of CPM. However, it seems to be interesting once an automation of management methods without manual support wants to be achieved.

The association rules discovered only comprise 3 different CPM related items *Data in business management processes are current*, *Data in business management processes are relevant* and *Management methods are fully automated and linked without manual support*. This awakes the awareness that BI is not the only information-technological support in companies. Enterprise Resource Planning Systems (ERP), Customer Relationship Systems (CRM) and Supply-Chain-Management System (SCM) also play an important role for a successful performance management. Before focusing on implementing a BI solution, the predominant step might be to focus on existing software first and afterwards built an effective BI solution on top.

2) Cluster Analysis: The cluster analysis divided the data set into 3 clusters. Cluster 1 contains the companies, which show a high development in both BI and CPM. The companies in this group mainly have above 1000 full-time

employees and are considered as big. The BI systems throughout the company are primarily uniform.

Cluster 3 incorporates medium-sized to big companies, but generally with less employees then the ones in Cluster 1. The BI and CPM development is mediocre. The BI systems are uniform but ERP and BI software are mostly from different manufacturers.

Cluster 2 comprises of a mixture of all company sizes but mainly medium-sized enterprises. They have by trend a less developed BI and CPM in common. BI and ERP software as well as the BI software generally are not uniform within the company.

The results imply a connection between BI and CPM. Clusters with a bad development in one discipline also show a low distinction in the other discipline and vice versa. It seems that long established big companies are further in implementing both CPM and BI. Usually, companies with more resources invest earlier in new technologies than other enterprises. As well, they can employ experts to help them implement and use new technologies. This might be an explanation for the high development in both BI and CPM in Cluster 1.

The results further show that medium-sized companies are generally not as experienced in BI. BI characteristics dealing with data quality and data provision are well developed as well, but the medium to low distinction of characteristics including supportive BI techniques and tools indicate no current usage. This is reflected on the CPM side with a low development of the overall connection between the business management processes.

In addition, the patterns in the data set also indicate that in Germany, there are companies of all sizes who struggle with CPM and BI. All clusters discovered differ in the handling of BI and ERP software, which could be a reason for the struggle. Companies in Cluster 2 use different software from various manufacturers for BI and CPM. Additionally, the majority of the cases in this cluster declared that they are not using uniform BI software throughout the company (more than 76%). Instead, the majority of companies in Cluster 1 claimed to use ERP and BI software from the same brand as well as uniform BI software across the whole enterprise. Uniform BI software reduces unnecessary interfaces and ensures that throughout the company the same data and numbers are used for decision support. Inferentially, this facilitates advantages for CPM by providing transparency, which supports the formation of feedback loops in operational and strategic processes.

The results of the Data Mining analysis answer the research questions and can support practitioners in building a successful BI solution. The identified association rules point out the BI characteristics, which are most related to a high development of certain CPM characteristics. The cluster analysis identified three different clusters with a different distinction of BI and CPM development. Therefore, a clear connection between BI and CPM can be recognized as well as additional items influencing a high development of both – BI and CPM.

V. DISCUSSION

The main goal of the previously published research by Hartl and Jacob [1] and this extended version has been to explore if Data Mining techniques can provide more detailed insights to IS research subjects. Association Rule Discovery and cluster analysis have been applied to the questionnaire based data set collected by Hartl et al. [5]. The aim was to explore, if these Data Mining procedures can extract even more information from the given data set.

In fact, information that is more detailed has been extracted from the data set. While in a PLS analysis all research items seem to matter equally, the association rule analysis showed that there are a handful of characteristics on BI and CPM side, which seem to be most related. For example, the business management processes do profit especially from a well-functioning BI - in detail, the existence of clear roles and responsibilities for operating the BI system and data integrity during simultaneous use. With this background, companies can analyze and improve their regarding BI development to ideally support the organizational alignment of their business processes. The results indicate that this is one of the first steps to a well-functioning CPM supported through BI.

In addition, the cluster analysis offered interesting results contained in the data set. Besides the BI and CPM related characteristics, the questionnaire contained general information about the company too (e.g., company size, software difference and similarities between ERP and BI systems used throughout the company). With cluster analysis, this information can be put into context with the CPM and BI characteristics evaluated. The results showed that a good development in BI and CPM is rather given, if the BI and ERP tools used within a company are from the same software developer.

In comparison to the subject related research, the results of the Data Mining approach show different and more detailed information about the connection of BI and CPM. Besides simply proving a positive relationship, the research outcomes allow conclusions on a path of action for practitioners. Although these inferences still need further investigation in practice, it has been possible to identify the BI and CPM items with the strongest connection. Furthermore, an already collected data set has been re-used and more insights could be gained about a previously investigated subject.

The Data Mining approach presents itself as a suitable addition to exploring the connection between BI and CPM. Inferentially, the results support the assumption that Data Mining methods are in general suitable for IS research subjects. In addition to reanalyze previously collected questionnaire data, Data Mining could support IS research without collecting data especially for the research purpose.

Data Mining methods like Text Mining and Web Mining might allow a totally different approach to IS research topics.

VI. Conclusion and Future research

The Data Mining approach to the research area of BI and CPM has been successful. Nevertheless, this research still presents an early attempt in exploring the usage of Data Mining in IS research. It might be possible, that different clustering and/or association algorithms are a better fit to explore data in a Likert scale format. Hence, future research should evaluate and compare the results of different algorithms when applying Data Mining to questionnaire based data sets. Additionally, this research only applies Association Rule Discovery and cluster analysis to a relatively small data set. Due to the nature of Data Mining it is well possible that even slightly bigger data sets (200 and more cases) could lead to improved and even more results.

A next step in continuation of this research will be the application of Data Mining to other IS related research topics. It will be interesting to see, if similar results can be obtained. Therefore, given data sets will be evaluated using similar Data Mining applications. If indeed more information can be extracted from each given data sets, many existing researches can benefit.

Another interesting future research will be the application of Data Mining applications as a first approach towards an IS research topic. Text Mining and Web Mining offer the extraction of information from existing sources. In Text Mining the goal is to extract high quality data from texts [3]. In the IS research field the major part of information is available in text form, for example in the form of articles, digital libraries and books. The use of Text Mining could help to gain an overview of a research topic, structure information and detect new research fields. With Web Mining techniques, structured as well as unstructured online data can be analyzed [3]. Existing websites, media data and web usage data can be explored. Both applications offer IS researchers a huge opportunity by providing insights to IS related subjects without relying on test persons and questionnaires. Therefore, Text Mining and Web Mining applications should be tested to search for answers regarding current research questions.

References

[1] K. Hartl and O. Jacob, "Using Data Mining Techniques for Information System Research Purposes - An Exemplary Application in the Field of Business Intelligence and Corporate Performance Management," The Fith Conference on Data Analytics (Data Analytics 2016) IARIA, Oct. 2016.

[2] K.-M. Osei-Bryson and O. Ngwenyama, eds. Advances in Research Methods for Information Systems Research. Springer-Verlag: New York, 2014.

[3] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. vol. 3. Waltham, USA: Morgan Kaufmann Publishers, 2012.

[4] S. Rouhani, A. Ashrafi, A. Z. Ravasan, and S. Afshari, "The Impact Model of Business Intelligence on Decision Support and Organizational Benefits," Enterprise Information Management, vol. 29(1), pp. 19-50, 2016.

[5] K. Hartl, O. Jacob, F.H. Lien Mbep, A. Budree, and L. Fourie, "The Impact of Business Intelligence on Corporate Performance Management," The 49th Hawaii International Conference on System Science (HICSS 2016), Jan. 2016, pp. 5041-5051.

[6] G. Richards, W. Yeoh, A.Y.L. Chong, and A. Popovič, "An Empirical Study of Business Intelligence Impact on Corporate Performance Management," The Pacific Asia Conference on Information Systems (PACIS 2014), 2014, Paper 341.

[7] R. Weiber and D. Mühlhaus, Structural Equation Modelling. 2nd ed., Berlin Heidelberg: Springer-Verlag, 2014.

[8] K. Backhaus, B. Erichson, W. Plinke, and R. Weiber, Multivariate Analysis Methods. An Application-oriented Introduction. vol. 13. Berlin Heidelberg: Springer-Verlag. 2011.

[9] N. Urbach and F. Ahlemann, "Structural Equation Modeling in Information Systems Research Using Partial Least Squares," Journal of Information Technology Theory and Application, vol. 11(2), Article 2, 2010.

[10] E. Vinzi, W. W. Chin, J. Henseler and H. Wang, eds. Handbook of Partial Least Squares: Concepts, Methods and Applications. Berlin: Springer-Verlag, 2010.

[11] J. F. Hair, M. Sarstedt, L. Hopkins, and V.G. Kuppelwieser, "Partial leats squares structurl equation modeling (PLS-SEM)," European Business Review, vol. 26(2), 2014.

[12] J. Jackson, "Data Mining: A Conceptual Overview," Communications of the Association for Information Systems, vol. 8, 2002. 8: pp. 267-296.

[13] A. J. Stolzer and C. Harlford, "Data Mining Methods Applied to Flight Operations Quality Assurance Data: A Comparison to Standard Statistical Methods," Journal of Air Transportation, vol. 12(1), pp. 6-24, 2007.

[14] M. L. Gargano and B.G. Raggad, "Data mining - a powerful information creating tool," OCLC Systems & Services: International digital library perspectives, vol. 15(2), pp. 81-90, 1999.

[15] J. Cleve and U. Lämmel, Data Mining. vol. 2. Berlin: Walter de Gruyter GmbH, 2016.

[16] S. Natek and M. Zwilling, "Data Mining for Small Student Data Set – Knowledge Management Sytem for Higher Education Teachers," The International Conference for Management, Learning and Knowledge, Jun. 2013, pp. 1379-1398.

[17] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, et al., "CRISP-DM 1.0," SPSS Inc., 2000, Available from: https://www.the-modeling-agency.com/crisp-dm.pdf.

[18] D. W. Joenssen and T. Müllerleile, "Missing Data in Data Mining," HMD Praxis der Wirtschaftsinformatik, vol. 51(4), pp. 458-468, 2014.

[19] M. Aho, "The Distinction between Business Intelligence and Corporate Performance Management - A Literature Study Combined with Empirical Findings," The Mini Conference on Scientific Publishing (MCSP 2010), 2010.

[20] M. Hannula and V. Pirttimäki, "Business Intelligence: Empirical Study on Top 50 Finnish Companies," Journal of American Academy of Business, vol. 2(2), pp. 593-599, 2003.

[21] S. Negash, "Business Intelligence," Communications of the Association for Information Systems, vol. 13(1), pp. 177 - 195, 2004.

[22] S. Williams and N. Williams, "The Business Value of Business Intelligence," Business Intelligence Journal, vol. 8, pp. 30-39, 2003.
[23] S. Miranda, "Beyond BI: Benefiting from CPM Solutions," Financial Executive, vol. 20(2), 2004.
[24] J. Becker, D. Maßing, and C. Janiesch, "An evolutionary process model for introducing Corporate Performance Management Systems," Data Warehousing, pp. 247-276 2006.
[25] F. H. Lien Mbep, O. Jacob, and L. Fourie, "Critical Success Factors of Corporate Performance Management (CPM): Literature Study and Empirical Findings," The Sixth International Conference on Business Intelligence and Technology (BUSTECH 2015), March 2015.
[26] N. Yogev, L. Fink, and A. Even, "How Business Intelligence Creates Value," The European Conference on Information Sytems (ECIS 2012), Paper 84, 2012.
[27] R. A. Saed, "The Relationship between Business Intelligence and Business Success: An Investigation in Firms in Sharjah Emirate," American Journal of Business and Management, vol. 2(4), pp. 332-339, 2013.
[28] R. M. Müller and H.-J. Lenz, Business Intelligence. Berlin Heidelberg: Springer-Verlag, 2013.
[29] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. vol. 2. Berlin Heidelberg: Springer-Verlag, 2011.
[30] M. Lang, ed. Handbook of Business Intelligence: Potentials, Strategies, Best Practices. vol. 1. Düsseldorf: Symposium, 2015.
[31] K. Oehler, Corporate Performance Management with Business Intelligence Tools. Carl Hanser Verlag: München, 2006.