Generating Content-Compliant Training Data in Big Data Education

Robert Häusler, Daniel Staegemann, Matthias Volk, Sascha Bosse, Christian Bekel and Klaus Turowski
Magdeburg Research and Competence Cluster Very Large Business Applications, Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

Keywords: Data Generation, Content-Compliant Data, Big Data, Education, Machine Learning.

Abstract: In order to ensure adequate education and training in a statistics-driven field, large sets of content-compliant training data (CCTD) are required. Within the context of practical orientation, such data sets should be as realistic as possible concerning the content in order to improve the learning experience. While there are different data generators for special use cases, the approaches mostly aim at evaluating the performance of database systems. Therefore, they focus on the structure but not on the content. Based on formulated requirements, this paper designs a possible approach for generating CCTD in the context of Big Data education. For this purpose, different Machine Learning algorithms could be utilized. In future work, specific models will be designed, implemented and evaluated.

1 INTRODUCTION

Facilitated by the rapidly evolving information technologies and the growth of computing and storing capabilities, the amount of data to be stored from varying application areas is constantly increasing (Yin and Kaynak, 2015). Besides velocity, variety and variability, volume is usually regarded as the major Big Data characteristic (NIST Big Data Public Working Group, 2018). This phenomenon has not only gained huge interest amongst researchers as well as practitioners (Grover et al., 2018; Maroufkhani et al., 2019; Staegemann et al., 2019b), but its exploitation has also a proven financial value for the involved companies (Müller et al., 2018).

As a consequence, there is a need of well-skilled graduates as well as employees in the field of Big Data. Around this topic, education service providers try to face these challenges by developing teaching and learning environments (TLEs) – compositions of teaching material, an application or information system and a model organization to make system-based teaching as realistic as possible – and deploying them to a worldwide community (Häusler et al., 2019). In the Big Data context, selected application examples are, for instance, customer segmentation and social media analysis in the field of digital marketing, production forecast within an ERP system or the use of different reporting tools having a look at buying and selling figures. For creating (parts of) those different TLEs, various types of data are required and it is necessary to command huge quantities of data. Preferably, they reveal a high similarity in terms of not only structure but also content (Gray et al., 1994). Frequently required company related data comprises, for instance, information regarding products, inventories, production facilities, large sets of customer data as well as transactional data. While much of this data is publicly available and therefore comparatively easy to acquire, this does not apply to customer and transactional data. On the contrary, especially customer data is usually protected by laws (Federal Ministry of Justice and Consumer Protection, 2018) and also constitutes valuable assets to the companies that possess it (Jöns, 2016). Whereas companies having appropriate data are able to anonymize or pseudonymize it for testing and training purposes, others have to resort to either acquisition or synthetic creation of such data (Lang, 2012). While there are existing approaches for the generation of such data, the focus is usually on the structure and the volume of data, but not on its content (Gray et al., 1994; Houkjær et al., 2006; Bruno and Chaudhuri, 2005), which might be crucial in some circumstances. One of those specific above-mentioned examples, stemming from the area of Big Data education, is customer segmentation as the primary stage of

Häusler, R., Staegemann, D., Volk, M., Bosse, S., Bekel, C. and Turowski, K.
Generating Content-Compliant Training Data in Big Data Education.
DOI: 10.5220/0009513801040110
In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020) - Volume 1*, pages 104-110
ISBN: 978-989-758-417-6

Customer Relationship Management (CRM) (Namvar et al., 2010). In order to ensure adequate education in this statistics-driven field (Tsou and Huang, 2018), large sets of content-compliant customer data are required, since (customer) segmentation is not suitable in data generated by a probability distribution.

Furthermore, also the domain of Big Data testing still lacks in maturity, wherefore the provision of realistic test data might be an important advancement (Tao and Gao, 2016; Staegemann et al., 2019a; Gao et al., 2016), constituting an additional area of application. For the purpose of creating a foundation for the generation of content-compliant training data (CCTD), the following research question is addressed: *How can CCTD be created in large-scale?*

In order to answer this question, the publication is structured as follows. After the introduction that also provides the motivation for the publication at hand, the related work is showcased. Thereupon, requirements and a conceptual design for the creation of CCTD are provided. In the end, a conclusion is given and future research prospects are reflected upon.

2 RELATED WORK

In the following, a selection of existing approaches, serving as a foundation for the requirements and design principles that are to be developed, is presented. These depict the result of a qualitative literature review, which is to be extended in future work.

ToXgene, a data generator by Barbosa et al. (2002), creates XML files that comply with a pre-defined structure. By doing so, they provide the user with a means of testing the performance of databases and evaluating algorithms. The focus mainly lies on its ability to quickly generate huge databases. Data are distinguished into *simpleType* respectively *complexType* and can adhere, inter alia, to a normal or exponential distribution. Evaluating database performance, the *Simple and Realistic Data Generation* approach by Houkjær et al. (2006) focuses on the creation of huge volumes of data that have dependencies within and between the tables. For realizing the dependencies, a directed graph is used. The user can customize the parameters for the generation. Furthermore, different distributions and relations can be specified for certain dependencies. Gray et al. (1994) proposed a dispersed approach for the fast creation of database contents. While the data are supposed to correspond to statistical characteristics, they are filled with dummy information. For the distribution, the authors propose several approaches, like uniform or negative exponential distribution as well as Gauss or Poisson distribution. The data generation framework by Bruno and Chaudhuri (2005) is conceptualized to be as general as possible, since the usual individualistic design of most generators reduces their comparability. For this purpose, they introduce the "Data Generation Language", comprising scalars, rows, and iterators. For the sake of completeness, it has to be mentioned, that there are also online data generators like *GenerateData* and *Mockaroo*. Although they offer data generation functionalities, they are subject to severe restrictions regarding the amount of data to be produced and the configurability. Furthermore, it is not visible how exactly those data are created, hampering their evaluation.

3 REQUIREMENTS AND CONCEPTUAL DESIGN

While there are different data generators for special use cases, the approaches mostly aim at evaluating the performance of database systems. Hence, they are rather focusing on structure instead of its content. However, some use cases demand a realistic data set with semantic correctness and content compliance (Ceravolo et al., 2018; Janowicz et al., 2015; Bizer et al., 2012). In order to generate CCTD, additional requirements are needed. They are formulated in the following section.

Within the context of practice-oriented education and training (e.g. analytical CRM, OLAP and data mining with the aim of deriving knowledge and thus supporting processes), data sets should be as realistic as possible concerning the content in order to improve the learning experience. Therefore, CCTD should have the highest possible specificity. It is particularly important to map the complicated properties or attributes, such as school career or family relationships in regards of customer data. The *ToXgene* approach has already shown a possible differentiation of attributes into simple (e.g. name, age, residence) and complex (e.g. income, hobbies, health status) types. In order to generate a robust, consistent and meaningful data set, a certain set of attributes and dependencies has to be defined. The power of this set is also important for the use of machine learning (ML) algorithms, since an enlargement of the input data can lead in many cases to an exponentially increasing calculation effort when training the algorithms (Fernandes de Mello and Antonelli Ponti, 2018). In general, a data model should be designed as complex as necessary to minimize the risk of unpredictable dependencies between different attributes. Because as complexity increases, so does the risk of unnatural growth due to variety and diversity within this "mesh of relationships". A static data set contains constant data that does not change over time. For example, when static customer data is generated, the corresponding records are created only once. In contrast, a dynamic data set is variable and usually designed based on a static one. That means, attributes like personal interests or hobbies can change over time. However, a dynamic system may develop a chaotic behaviour (Bungartz et al., 2013). This can be prevented by using control parameters. Nevertheless, both static and dynamic approaches could be implemented by means of ML. Figure 1 shows a possible concept.

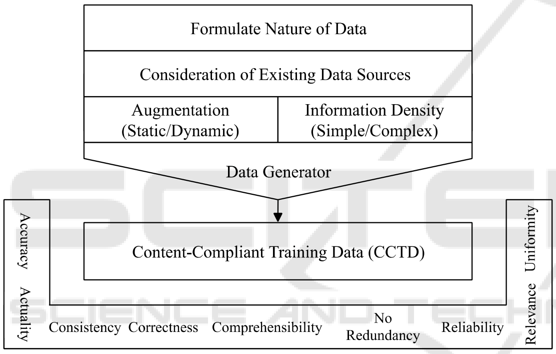
Formulate Nature of Data
Consideration of Existing Data Sources
Augmentation (Static/Dynamic)
Information Density (Simple/Complex)
Data Generator
Content-Compliant Training Data (CCTD)
Accuracy
Actuality
Consistency
Correctness
Comprehensibility
No Redundancy
Reliability
Relevance
Uniformity

Figure 1: Possible concept for generating CCTD.

Firstly, the nature of data has to be formulated. With the help of process analysis or expert interviews (in the education service provider context for example with the owner of a TLE or any data expert), one or more specific use cases has to be defined at the beginning. Additionally, the requirements of these specific cases should be recorded. It is important to determine, which data (types) are mandatory, which attributes must be contained and which technical systems are involved to deliver the correct data in the appropriate format. Based on this, heterogeneous but suitable data sources should be gathered. For example, statistic (web) services can be used for this purpose, such as Statista[1], federal statistical offices or any other sources providing actual data. This step is necessary for ML algorithms to analyse the structure of the original data in order to reproduce this data and generate it in large scale. Due to the mix of different requirements and data sources, augmentation and information density have to be defined. It has to be dealt with the question, whether a dynamic approach is possible with the existing data or if a static data set have to be created first. On the other hand, it must be determined which data can be mapped as *simple*, which as *complex* type and how the structure/relationships between these different data types should be defined/built. The identified requirements, the gathered data sources and the designed data model serve as input parameters for generating CCTD. As a generator, different ML algorithms, such as classification/clustering, Markov chains or neural networks, could be served. The objective of classifying or clustering is to assign a set of data to a certain attribute, the class, or to identify any "hidden" relationship. To learn the classification, the algorithm is trained using a test data set (Harrington, 2012). In regards to customer data, the classification can be explained by many different attributes. For instance, customers can be classified according to their sportiness, based on their hobbies or their health. Overall, this method can be used for applications in which specific customer groups are to be determined. In contrast to classification, a Markov chain can model stochastic processes characterized by transitions between their different states (Waldmann and Stocker, 2013). Due to its temporal component in the form of transitions, this method is suitable for modeling time sequences. In the area of customer data, this could be used to model a customer's school education or professional career. Therefore, the focus is not on the structure of the entire data set, but rather on the development of individual customers over time. As third possible ML algorithm, Generative Adversarial Networks (GAN) could be used, which represents a special and relatively new application of artificial neural networks. This concept means several neural networks, which interact with each other, following the rule of supervised learning (Li et al., 2013; Goodfellow et al., 2014). The generating neural network adapts the generated data successively to the training data set based on the decisive network. Thus, GAN generates data that cannot be distinguished from real data. The method was mainly used to generate images that were as realistic as possible and later, to create plausible orders in the e-commerce environment. Possibly, this method can also be applied for the creation of customer data. Like classification and clustering, neural networks are only conditionally suitable for the representation of

[1] https://www.statista.com/

temporal processes. By adding new properties to existing customers, only a very simple form of temporal change can be represented. Furthermore, this method can tend to evoke the issues that newly generated customers do not differ enough from existing customers. In conclusion, the usefulness of these methods depends largely on the requirements of the specific use case.

In particular, attention must be paid to performance and volume, since data is to be generated in the Big Data context. The importance of high data quality has to be considered particularly with regard to the use of the data in CRM systems. In many areas, this data is processed using Big Data models and ML techniques. The quality of the results after application of the latter depends significantly on the quality of the used databases (Hazen et al., 2014). Hence, as one possible evaluation (framing CCTD in the figure), data quality dimensions could be considered to quantify data quality (Geuer, 2017). In this article, Geuer describes and defines eleven different data quality criteria, which are depicted in table 1.

Table 1: Criteria for quantifying data quality.

Accuracy	The data must be available in the required accuracy.
Actuality	All data sets must always correspond to the current state of the depicted reality. Each data set must be uniquely interpretable.
Clarity	Each data set must be uniquely interpretable.
Completeness	A data set must contain all necessary attributes. Attributes must contain all necessary data.
Consistency	A data set must not contain any contradictions within itself or to other data sets.
Correctness	The data corresponds to reality.
Comprehensi-bility	The data sets must correspond in their terminology and structure to the ideas of the information recipients.
No Redundancy	No duplicates may occur within the data sets.
Reliability	The origination of the data must be reliable.
Relevance	The information content of data sets must meet the respective information needs.
Uniformity	The information of a data set must have a uniform structure.

The degree of fulfilment of a criterion can be in the range of 0 to 100 percent, whereby a degree of fulfilment of 100 percent should be aimed for. Geuer (2017) points out that in practice not all criteria are usually applied, but only a subset of the requirements. In view of the motivation, consistency plays a promising role to meet the requirements of the determined use cases. In addition to data quality, information quality metrics could extend the evaluation (Rohweder et al., 2015). However, in both cases, it is important to define the dimensions or metrics for each use case individually.

4 FIRST USE CASE

First results were achieved in smaller practical projects. In fact, a static customer data set for a fictitious bicycle manufacturer could be prototypically generated in order to expand the CRM TLE of an education service provider. The aim of this use case is to create a marketing campaign based on the health of people using a generated data set that represents a non-specific German region. For this purpose, the procedure was based on the aforementioned concept.

Firstly, two experts (owners of the TLE) were separately interviewed. These experts are employees of the education service provider and they each deal with the topics and the tools of (digital) marketing in the education context for more than ten years. To determine the nature of data, mostly open questions were asked. For the further identification of special requirements of this use case, the experts had to name and rate different customer data attributes and their importance by answering questions and tasks like: *“Which attributes are necessary to use the model?”, “How are these attributes related to each other?”* and *“Please rate the attributes according to their importance.”*. As a central requirement, it was mentioned that the generator should be able to process discrete (as far as possible meaningful) distributions with any variables. Furthermore, the dependencies within the different attributes were stated as an important characteristic. Both should lead to statistical correctness. A further requirement was the effective generation of up to 500,000 data sets for this use case. These specifications extended the pool of generic requirements (e.g. data quality).

After identifying the requirements, a data model was created based on the named and rated attributes (cf. augmentation and information density). The final data set includes ten *simpleType* attributes, such as *name*, *date of birth* and *address*, as well as one significantly simplified *complexType*. The latter, *health,* is depending among others on age and hobby. Subsequently, existing data sources were searched for and examined trying to consider all the mentioned requirements, such as statistics on age distribution and popular recreational activities or databases with names and addresses. Some of this raw data was already available in CSV format and could be downloaded without any problems. The majority of these are websites of various (German) institutions. However, some information had to be converted into the correct format. All the data was checked and cleaned if necessary. This procedure should ensure a high data quality.

The prototype was implemented in *Python*. This programming language was chosen because the algorithms are easy to implement as well as to use and because extensive machine learning packages are already available for future developments (with regard to dynamization). In the first step, the necessary data sources were loaded and processed. For this purpose, different arrays containing the attribute values and probability distributions were generated for every *simpleType* attribute to build the source database. Based on this, the *simpleType* attributes of the respective customer records were generated step by step, whereas the *complexType* attributes were derived according to the data model. For simple data retention and further processing, the use of a database system for storing the data is not required. The generated test data sets were stored as CSV format because of its versatility. It is possible to save generated customer data directly from the prototypical implementation and to import them into different applications for further processing.

Following this process, a total of 500,000 customer data records were generated several times. In order to estimate whether it will be possible to generate data in an acceptable time with the developed model for customer data generation, the duration was measured. This evaluation was performed on virtual machines of the education service provider. To obtain the most meaningful values, each data record was generated ten times. Afterwards, the average of the respective total duration was calculated. In total, 500,000 data records could be generated in 32,293.1 ms, which means about 0.064 ms per data record. Additionally, the data sets were qualitatively evaluated according to parts of the quality criteria of Geuer (2017). Due to the methodical and structured approach of CCDT generation, completeness, reliability and uniformity are ensured. Furthermore, no redundancies could be found in the generated data sets. Besides, the data model implicitly leads to a certain clarity as well as consistency. Subsequently, the experts confirmed relevance and comprehensibility as sufficient for the use case. About correctness, accuracy and actuality, no detailed statements could be made as these depend highly on the used information. However, the statistical correctness of the generated data was compared with the distributions of the source data. A maximum deviation of 0.16 percent could be achieved for all attributes.

In summary, it can be said that the requirements of both data quality and experts are met. The concept developed within the scope of this position paper could be used to generate a static customer data set. However, comprehensive evaluations in real-world classrooms and different structured extensions of this specific case are still pending. Additionally, further use cases have to be defined in cooperation with the education service provider.

5 CONCLUSION AND FUTURE WORK

This study has shown that the identified generators mainly address database performance, but not solve the motivated problems in Big Data education. One promising approach in this field is the generation of CCTD. Nevertheless, this concept still raises some questions. For instance, how detailed complex real-world structures can be represented? This highly depends on the certain domain knowledge for the designed uses cases. If the data generation or especially the data model should be configured correctly, expert knowledge is mandatory.

The aforementioned ML algorithms could provide one possible solution. These approaches can analyse complex source structures and consequently generate data in large-scale accordingly. But, it is also necessary to examine how large the source data base has to be. Using ML, the aim of future works is the generation of other static and dynamic data sets for various use cases. Nevertheless, even ML algorithms are just calculations. Hence, unpredictable events or environmental influences cannot be reproduced if they are not included in the original data.

However, the concept presented in this paper is very general. In order to refine it, the transferability of methods from existing approaches will be investigated further. Based on the future formulated concept, specific data models will be designed, implemented and evaluated.

REFERENCES

Barbosa, Denilson; Mendelzon, Alberto; Keenleyside, John; Lyons, Kelly (2002): *ToXgene*. In David DeWitt, Michael Franklin, Bongki Moon (Eds.): Proceedings of the 2002 ACM SIGMOD international conference on Management of data - SIGMOD '02. the 2002 ACM SIGMOD international conference. Madison, Wisconsin, 03.06.2002 - 06.06.2002. New York, New York, USA: ACM Press, p. 616.

Bizer, Christian; Boncz, Peter; Brodie, Michael L.; Erling, Orri (2012): *The Meaningful Use of Big Data. Four Perspectives – Four Challenges*. In *SIGMOD Rec.* 40 (4), pp. 56–60.

Bruno, Nicolas; Chaudhuri, Surajit (2005): *Flexible database generators*. In Klemens Böhm (Ed.): Proceedings of the 31st International conference on very large data bases. Trondheim, Norway, 30.08.2005-02.09.2005. New York: ACM, pp. 1097–1107.

Bungartz, Hans-Joachim; Zimmer, Stefan; Buchholz, Martin; Pflüger, Dirk (2013): *Modellbildung und Simulation*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ceravolo, Paolo; Azzini, Antonia; Angelini, Marco; Catarci, Tiziana; Cudré-Mauroux, Philippe; Damiani, Ernesto et al. (2018): *Big Data Semantics*. In *J Data Semant* 7 (2), pp. 65–85.

Federal Ministry of Justice and Consumer Protection (2018): *Federal Data Protection Act. BDSG*. Available online at http://www.gesetze-im-internet.de/bdsg_2018/, checked on 1/28/2020.

Fernandes de Mello, Rodrigo; Antonelli Ponti, Moacir (2018): *Machine Learning. A Practical Approach on the Statistical Learning Theory*. Cham: Springer Nature; Springer.

Gao, Jerry; Xie, Chunli; Tao, Chuanqi (2016): *Big Data Validation and Quality Assurance -- Issuses, Challenges, and Needs*. In : 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE). 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE). Oxford, United Kingdom, 29.03.2016 - 02.04.2016: IEEE, pp. 433–441.

Geuer, Marco (2017): *Datenqualität messen: Mit 11 Kriterien Datenqualität quantifizieren*. Available online at https://www.business-information-excellence.de/datenqualitaet/86-datenqualitaet-messen-11-datenqualitaets-kriterien, updated on 10/29/2017, checked on 1/28/2020.

Goodfellow, Ian J.; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil et al. (2014): *Generative Adversarial Networks*. Available online at http://arxiv.org/pdf/1406.2661v1.

Gray, Jim; Sundaresan, Prakash; Englert, Susanne; Baclawski, Ken; Weinberger, Peter J. (1994): *Quickly generating billion-record synthetic databases*. In Richard T. Snodgrass, Marianne Winslett (Eds.): Proceedings of the 1994 ACM SIGMOD international conference on Management of data - SIGMOD '94. the 1994 ACM SIGMOD international conference. Minneapolis, Minnesota, United States, 24.05.1994 - 27.05.1994. New York, New York, USA: ACM Press, pp. 243–252.

Grover, Varun; Chiang, Roger H.L.; Liang, Ting-Peng; Zhang, Dongsong (2018): *Creating Strategic Business Value from Big Data Analytics: A Research Framework*. In *Journal of Management Information Systems* 35 (2), pp. 388–423.

Harrington, Peter (2012): *Machine learning in action*. Greenwich, Conn.: Manning (Safari Tech Books Online). Available online at http:// proquest.safaribooksonline.com/9781617290183.

Häusler, Robert; Bernhardt, Chris; Bosse, Sascha; Turowski, Klaus (2019): *A Review of the Literature on Teaching and Learning Environments*. In: 25th Americas Conference on Information Systems, AMCIS 2019, Cancun, Q.R, Mexico, August 15-17, 2019: Association for Information Systems.

Hazen, Benjamin T.; Boone, Christopher A.; Ezell, Jeremy D.; Jones-Farmer, L. Allison (2014): *Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications*. In *International Journal of Production Economics* 154, pp. 72–80.

Houkjær, Kenneth; Torp, Kristian; Wind, Rico (2006): *Simple and realistic data generation*. In Umeshwar Dayal (Ed.): Proceedings of the 32nd international conference on very large data bases. VLDB Endowment, pp. 1243–1246.

Janowicz, Krzysztof; van Harmelen, Frank; Hendler, James A.; Hitzler, Pascal (2015): *Why the Data Train Needs Semantic Rails*. In *AIMag* 36 (1), p. 5.

Jöns, Johanna (2016): *Daten als Handelsware*. Deutsches Institut für Vertrauen und Sicherheit im Internet. Available online at https://www.divsi.de/wp-content/uploads/2016/03/Daten-als-Handelsware.pdf, checked on 1/28/2020.

Lang, Andreas (2012): *Anonymisierung/ Pseudonymisierung von Daten für den Test*. In: D.A.CH Security Conference 2012. Konstanz.

Li, Wei; Gauci, Melvin; Gross, Roderich (2013): *A coevolutionary approach to learn animal behavior through controlled interaction*. In Enrique Alba, Christian Blum (Eds.): Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference - GECCO '13. Proceeding of the fifteenth annual conference. Amsterdam, The Netherlands, 06.07.2013 - 10.07.2013. New York, New York, USA: ACM Press, p. 223.

Maroufkhani, Parisa; Wagner, Ralf; Wan Ismail, Wan Khairuzzaman; Baroto, Mas Bambang; Nourani, Mohammad (2019): *Big Data Analytics and Firm Performance: A Systematic Review*. In *Information* 10 (7), p. 226.

Müller, Oliver; Fay, Maria; Vom Brocke, Jan (2018): *The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics*. In *Journal of Management Information Systems* 35 (2), pp. 488–509.

Namvar, Morteza; Gholamian, Mohammad Reza.; Khakabimamaghani, Sahand (2010): *A Two Phase Clustering Method for Intelligent Customer Segmentation*. In : 2010 International Conference on Intelligent Systems, Modelling and Simulation. 2010 International Conference on Intelligent Systems, Modelling and Simulation (ISMS). Liverpool, United Kingdom, 27.01.2010 - 29.01.2010: IEEE, pp. 215–219.

NIST Big Data Public Working Group (2018): *NIST Big Data Interoperability Framework: Volume 1, Definitions, Version 2*. Gaithersburg, MD: National Institute of Standards and Technology. Available online at https://bigdatawg.nist.gov/uploadfiles/NIST.SP.1500-1r1.pdf, checked on 1/28/2020.

Rohweder, Jan P.; Kasten, Gerhard; Malzahn, Dirk; Piro, Andrea; Schmid, Joachim (2015): *Informationsqualität – Definitionen, Dimensionen und Begriffe*. In Knut Hildebrand, Marcus Gebauer, Holger Hinrichs, Michael Mielke (Eds.): Daten- und Informationsqualität. Auf dem Weg zur Information Excellence. 3., erweiterte Auflage. Wiesbaden: Springer Vieweg, pp. 25–46.

Staegemann, Daniel; Hintsch, Johannes; Turowski, Klaus (2019a): *Testing in Big Data: An Architecture Pattern for a Development Environment for Innovative, Integrated and Robust Applications*. In : Proceedings of the WI2019, pp. 279–284.

Staegemann, Daniel; Volk, Matthias; Jamous, Naoum; Turowski, Klaus (2019b): *Understanding Issues in Big Data Applications – A Multidimensional Endeavor*. In : 25th Americas Conference on Information Systems, AMCIS 2019, Cancun, Q.R, Mexico, August 15-17, 2019: Association for Information Systems.

Tao, Chuanqi; Gao, Jerry (2016): *Quality Assurance for Big Data Applications– Issues, Challenges, and Needs*. In : The Twenty-Eighth International Conference on Software Engineering and Knowledge Engineering. California. Pittsburgh: KSI Research Inc. and Knowledge Systems Institute Graduate School, pp. 375–381.

Tsou, Hung-Tai; Huang, Yao-Wen (2018): *Empirical Study of the Affecting Statistical Education on Customer Relationship Management and Customer Value in Hi-tech Industry*. In *EURASIA J. Math., Sci Tech. Ed*.

Waldmann, Karl-Heinz; Stocker, Ulrike M. (2013): *Stochastische Modelle. Eine anwendungsorientierte Einführung*. 2., überarb. und erw. Aufl. Berlin: Springer (EMIL@A-stat Medienreihe zur angewandten Statistik).

Yin, Shen; Kaynak, Okyay (2015): *Big Data for Modern Industry: Challenges and Trends [Point of View]*. In *Proc. IEEE* 103 (2), pp. 143–146.