Decision Support and Online Advertising
Development and Empirical Testing of a Data Landscape

Thomas Hansmann[1] and Florian Nottorf[2]
[1]*Institut für Elektronische Geschäftsprozesse, Leuphana University, Lüneburg, Germany*
[2]*Adference GmbH, Am Urnenfeld 5, Lüneburg, Germany*

Keywords: Data Landscape, Decision Support, Online Advertising, User Journey.

Abstract: The number of data sources available inside and outside companies and total data points increase, which makes the coordinated data selection in the forefront of decision making with respect to a specific economic goal becomes more and more relevant. To assess the available data and enhance decision support, we develop a framework including a process model that supports the identification of goal-oriented research questions and a data landscape that provides a structured overview of the available data inside and outside the company. We empirically tested the framework in the field of online advertising to enhance decision support in managing display advertising campaigns. The test reveals that the developed data landscape supports the identification and selection of decision-relevant data and that the subsequent analysis leads to economic valuable results.

1 INTRODUCTION

Since 2000, data generation by various sources, such as Internet usage, mobile devices and industrial sensors in manufacturing, has been growing enormously (Hilbert and López, 2011). As the number of data sources as well as the total number of data points available inside and outside of companies have increased, coordinated data selection in the forefront of decision making with respect to a specific economic goal has become more relevant (LaValle et al., 2011). The lack of a detailed and goal-oriented data selection process may lead to inefficient decision support (DS) because i) questions regarding which data sources are generally available for specific analytic purposes and ii) questions about which data sources and respective results should be integrated into the decision making process remain unanswered.

To identify relevant, available data, we propose that both a *process model* for identifying specific optimization problems and the development of a *data landscape* that provides a structured overview of the available data inside and outside the company as well as its characteristics are mandatory. To the best of our knowledge, neither such a process model nor a data landscape for DS currently exist (Author, 2013).

We test our model in the field of online advertising, as the process of data selection and data evaluation is particularly relevant for companies doing online advertising. The field of online advertising offers multiple possible data sources within and outside the advertising company in different levels of aggregation (e.g., specific user-level data vs. aggregated data) at different levels of temporal availability (e.g., frequently vs. sporadic). Online advertising has become increasingly important for companies in their attempts to increase consumer awareness of products, services, and brands. With a share of nearly 50% of total online advertising spending, paid search advertising has become the favored online advertising tool for companies. In addition to paid search advertising, companies can combine several forms of display advertising, such as banner or affiliate advertising, on multiple platforms (i.e., information sites, forums, or social network sites) to enhance consumer awareness (Braun and Moe, 2013). These increased opportunities to advertise online add complexity to managerial decisions about how to optimally allocate online advertising spending, as consumers are often exposed to numerous types of online advertising during their browsing routines or their search-to-buy processes (Rutz and Bucklin, 2011b).

The goal of this paper thus is twofold: i) the development of a process model for the generation of a data landscape and ii) its empirical application.

The paper is structured as follows: after describing the current state of science about data selection and its weaknesses in the field of DS applications,

Hansmann T. and Nottorf F..
Decision Support and Online Advertising - Development and Empirical Testing of a Data Landscape.
DOI: 10.5220/0005060401110122
In *Proceedings of the 11th International Conference on e-Business* (ICE-B-2014), pages 111-122
ISBN: 978-989-758-043-7
Copyright © 2014 SCITEPRESS (Science and Technology Publications, Lda.)

a process model for the development of a data landscape is developed. This section is followed by the testing of the proposed model in the field of online advertising. Finally, based on the identified data, we apply the model of (Nottorf and Funk, 2013) to enhance DS in the field of display advertising. After outlining our findings and discussing our results, we conclude this study by highlighting its limitations and providing suggestions for future research.

2 DATA LANDSCAPE AND DECISION SUPPORT

2.1 Current Research

An initial literature review revealed that no process models specific to the development of data landscapes have been published in the field of online advertising or decision support, although (Chaudhuri et al., 2001) claim that "what data to gather and how to conceptually model the data and manage its storage" is a fundamental issue.

The fields of data warehouse (DW) and information system (IS) development represent a preliminary stage in developing data landscapes in terms of information requirement analysis, which includes the identification of data and information necessary to support the decision maker (Byrd et al., 1992). (Winter and Strauch, 2003) distinguish between the two systems, citing the underlying IT-infrastructure, the number of interfaces and connections, the degree of specification, and the number of involved organizational units as distinguishing factors. The different characteristics lead to a disparity in the information requirement analysis because IS requirements target "necessary and desirable system properties from prospective users" whereas the required information for a data warehouse system can usually not be gathered correctly due to the "uniqueness of many decision/knowledge processes". Consequently, how extensively these models can be applied to data landscape development must be tested.

The existing identification approaches for DW can be categorized as data/supply-, requirement/goal/demand-, or process-driven (Winter and Strauch, 2003). Data-driven approaches focus on the available data, which can be found in the operational systems (e.g., ERP or CRM systems) (Moody and Kortink, 2000; Golfarelli et al., 1998). This approach can help identify the sum of the overall available data but fails to incorporate the users' respective decision-makers actual and future requirements. Requirement-driven approaches focus on the requirements of the system user, assuming that a user can best evaluate his information need, which is simultaneously a limiting factor because most users are not aware of the overall available data sources (Gardner, 1998). Furthermore, in an early study (Davis, 1982) explains human biasing behaviors, which have a negative influence on data selection in the initial phases of a data warehouse development. He describes strategies to determine the information requirements, including asking, deriving them from an existing information system, synthesizing them from characteristics of the utilizing system, and discovering them through experimentation with an evolving information system. He also emphasizes the relevance of data characteristics, claiming, "the format of the data is the window by which users of the data see things an events. Format is thus constrained by the structure."

As a special form of the requirement-driven approach, the process-driven approach focuses on data from existing business processes and therefore avoids the subjectivity of the requirement-driven approach and the constraints of the data-driven approach (List et al., 2000). Depending on the coverage of business processes by IT systems, this approach can produce results that are similar to those of the data-driven approach; as more process steps are covered, the results from the two approaches are more comparable. One challenge for the use of the process-driven approach in landscape development can be the identification of the relevant decision process.

Using a method engineering approach, the information requirement analysis by (Winter and Strauch, 2004) introduces the information map that described "which source systems provide which data in which quality" but does not amplify the development of this data landscape. (Giorgini et al., 2005) present a mixed demand/supply-driven goal-oriented approach, incorporating the graphical representation of data sources and attributes depending on the particular analytic goal. The graphical representation contains aspects of a data landscape but does not contain a characterization/evaluation of the attributes and focuses on existing, internal data sources. (Mazón et al., 2007) also propose a goal-oriented approach, introducing a hierarchy among the strategic, decisional and informational goals. Based on the information goals, measures and dependencies among them are identified.

Less research has been published regarding information requirement analysis for IS/decision support systems. (Byrd et al., 1992) categorize existing approaches into observation techniques (prototyping), unstructured elicitation techniques (e.g., brainstorming and open interviews), mapping techniques (e.g.,

variance analysis), formal analytic techniques (repertory grid), and structured elicitation techniques (e.g., structured interviews and critical success factors), which can be used to identify requirements based on existing information systems. (Davis, 1982) presents four strategies for generic requirement identification on the organization or application-level: i) asking, ii) deriving it from an existing information system, iii) synthesizing it from characteristics of the utilizing system, and iv) discovering it from experimentation with an evolving information system. In their literature review, (Stroh et al., 2011) compare and evaluate methods for analyzing information requirements for analytical information systems based on the requirement engineering by (Kotonya and Sommerville, 1998). Their analysis reveals that most publications address elicitation, but the issue needs to be pursued further. The same applies to research about documentation of the information requirement, which lacks a "sufficient level of detail" that is coherent for both business and IT.

The presented models can not be utilized for the information requirement analysis in the context of decision support as the existing models focus on internal company data and hence do not consider possible valuable external data for DS purposes. Therefore, an external perspective has to be incorporated. Second, to cope with the multiple data sources, a structure must be provided that supports focusing only on decision-relevant data which can only be found in the work by (Giorgini et al., 2005) and (Mazón et al., 2007). Consequently, we propose a process model decision support that enhances the process of identifying and evaluating potential data sources.

2.2 Development of the Process Model for the Data Landscape

The proposed process model for data landscape development combines and extends the goal-oriented approaches by (Giorgini et al., 2005) and (Mazón et al., 2007) and the data model-oriented level-approach by (Inmon, 2005). The initial goal-oriented approach helps identify relevant analysis tasks, whose results support the overall decision making process.

The starting point can be the pursuit of a strategic goal or a specific analytic question. In the first case, the decision and information goals are derived based on the strategic goal, using a top-down approach. For example in the field of online advertising, a strategic goal can be the improvement of the overall company reputation or an increase in sales. These goals can focus on the department level or the company level.

In the next step, the strategic goal is itemized into decision goals, which, when completed, contribute to the achievement of the overall strategic goal.

In the third step, the decision goals are specified by developing information goals as the lowest hierarchical step. Information goals are concrete goals that contain distinctive analytic questions. These form the basis for the subsequent identification of relevant data sources in an information requirement analysis.

The goal hierarchy supports the identification of analytic questions, based on requirements, as a first step to frame the requirements based on the necessary decision support, incorporating the uniqueness of each decision making process (Winter and Strauch, 2003). Furthermore, it fosters the definition of analytic goals, independent of the perceived limitations regarding employees' knowledge of available data sources. Due to their granularity, information goals can be used to derive concrete hypotheses that can be tested. In case a concrete analytic goal exists, this technique can be used as a bottom-up approach to identify further informational goals. In this case, the related decisional and strategic goals are first defined. Based on the decision goal, further information goals are derived.

In the next step, the related business process is defined for each analytic goal. For example in the field of online advertising, for the possible information goal "analyze online customer conversions under the influence of online advertising" the related generic business process is established as a potential customer interacts with an advertisement (i.e., by being exposed to a banner advertisement or clicking on a paid search advertisement), visits the online shop, and purchases a product.

In the next step, the related data sources, e.g., ERP-/CRM-systems, and attributes for each process step are identified. To this point, this approach for a high- or mid-level data analysis is similar to the one proposed by (Inmon, 2005). We extend this approach to cope with the requirements of DS in the emerging Big Data context regarding the dimensions volume, variety, velocity and veracity. Considering the numerous data sources within and outside the company that can contain business process and decision-relevant data, we extend the approach by distinguishing internal and external data sources (Stonebraker and Robertson, 2013). For example, the data sources regarding a purchased product are not limited to product master data and sales data on the product level. They can be enriched by customer reviews from external product platforms regarding customer satisfaction or product weaknesses and can therefore foster the decision support, e.g., with regard to companies spending on product development, product quality manage-

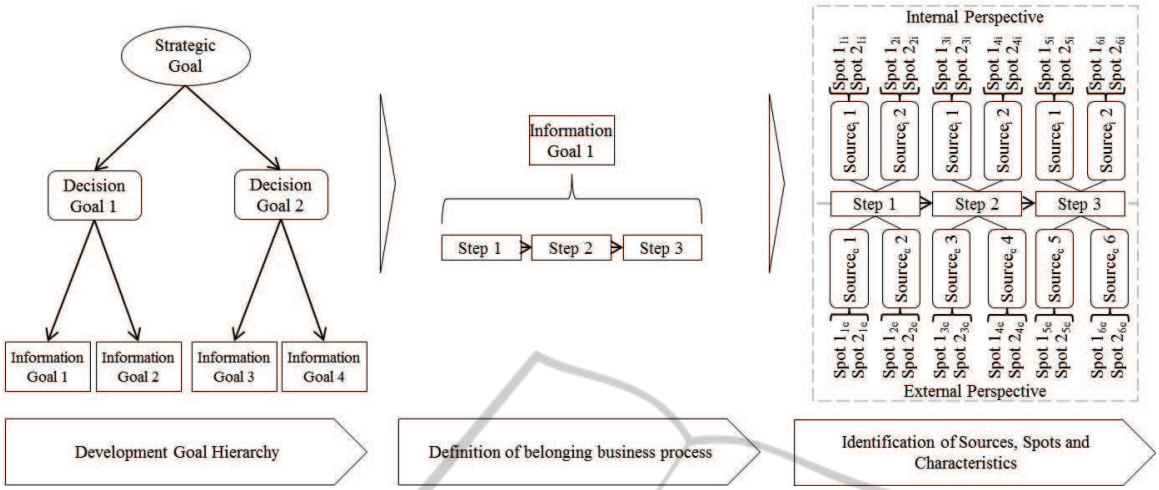

Figure 1: Process Steps.

ment, or reputation management.

The available data sources, spots and attributes in the field of online advertising and decision support are heterogeneous. We understand data spots to be the next lower level of data sources — the customer or product master data which contain again attributes, e.g., name, address. Consequently, for decisions in which data sources and spots should be integrated into the analysis, information about the potential information content and the amount of data processing work resulting from its characteristics is needed. For example, the characteristics in Table 1 must be defined for each attribute.

Therefore, the second extension is the introduction of a low-level attribute characterization that contains the determination of data characteristics for each attribute in addition to the type and source system of the data, which are already known from database development-related approaches. Furthermore, attributes that do not contain further insight independent of the decision in focus (e.g., customer telephone number) are eliminated in the following data cleaning step. This removal step aims to simplify the subsequent model building process. Previous approaches to information requirement analysis do not consider further data characteristics as the physical attributes like data type (e.g. varchar). With the increase in the number and points of origin of potential available data sources, a cost estimation in the early stages of heterogeneous source utilization is crucial.

The determination of characteristics fosters the evaluation of attributes regarding costs and effort for an integration into the DS. Using Twitter as an example, although data collection is simplified by using the available API, the process of data cleaning with regard to the noisy data is time consuming. Conversely, the (pre-)processing of clickstream data is less time consuming due to the higher degree of structure. To incorporate these characteristics, the degree of structure and distinction between machine- and human-generated data is introduced, assuming that unstructured data generated by humans, such as reviews or blog entries, are more likely to contain noisy data, which increase the time needed for data (pre-) processing due to typos (e.g., “gooood” instead of “good”) or linguistic features (e.g., irony, sarcasm). With regard to blog entries or tweets from different countries, the text language also influences the preprocessing time, although research has revealed that machine-based translation does not necessarily impair the results (Forcada et al., 2011). The effort for data preprocessing is related to the data quality, which is a major subject in the field of Big Data (Madnick et al., 2009). In addition the available volume influences the sample size and the coverage of the analysis. The velocity influences the time intervals in which the decision model can be updated based on new data. The costs per unit target purchased data, e.g. advertising data or market research data. The level indicates in how far decisions can be made on customer level. The historical availability defines the period, which can be incorporated in the analysis. This is of special interest regarding the changes in customer online behaviour. In case different internal and external data sources are supposed to be integrated in a decision support system, the data characteristics can support the technological decisions regarding database management software as well. The introduced aspects of external data integration and characterization incorporate the requirements from decision support into the Big Data context. Based on the developed data landscape, a model building process that is used to answer the ori-

Table 1: Data characteristics and possible features for each attribute.

Characteristics	Features	Implications
Data type	Integer, Small Integer etc.	
Degree of Structure	High/mid/low	Time for (pre-)processing
Volume	Actual available amount of data	Size of test sample
Velocity	Amount per time unit	Update cycle of decision model
Costs	Costs per unit	Cost estimation per decision
API available	y/n + data throughput	Effort for data gathering
Level	Individual/Aggregated	Explanatory power on individual level
Data origin	Machine-generated/Human-generated	Time for (pre-)processing
Historical availability	Time units of backwards availability	Period the decision model is based on
Language	Country code	Need for translation

gin question can be established.

In order to develop a sound theoretical foundation, the presented model and its development is evaluated in the next step based on the design-science research guidelines (Hevner et al., 2004). Although the presented paper is not solely linked with information system research, it exist extensive overlaps with the field of IT infrastructure especially data warehousing. With regard to the limited space of this paper, the guidelines are only shortly described and than cross checked with the presented model.

1) *Design as an artifact* demands the production of a viable artifact. This is fulfilled as an independent process model is developed, applicable as a basis for the respective information system development. The 2) *Problem Relevance* is given as until today companies are confronted with an 1.4-fold annual data growth (Manyika et al., 2011) based on numerous different company-internal and -external sources which results in the described insecurity about the data selection for decision support applications. The 3) guideline *Design Evaluation* demands for an evaluation of the utility, quality, and efficacy of the designed artifact. Therefore, in the next section, the model is applied in a two step approach in the field of online marketing, both qualitative and empirical. Guideline 4) targets the *Research Contribution*. As no comparable process model for the development of a data landscape exists so far, the presented model is a distinct contribution. This aspect is in conjunction with guideline 5), the *Research Rigor* in terms of the application of rigorous methods. This is given as the in the forefront of the model building, an extensive literature review has been carried out, which led to the selection of the two presented publications, which act as a basis for the developed model, complemented by an two step model evaluation as described. Guideline 6) contains the *Design as a Search Process*, demanding the utilization of available means. This transfer of this guideline can not be executed completely as the with regard to the novelty of this approach, the run through several test cycles in order to refine the means could not be carried out so far.

3 EMPIRICAL APPLICATION IN THE FIELD OF ONLINE ADVERTISING

3.1 Testing the Process Model

We test the model using the example of a telecommunication service provider that sells its products and services both online and in brick-and-mortar outlets. We first define a strategic goal and then develop respective information goals. This is followed by the definition of the corresponding business process and the identification of related data sources, data spots and attributes.

For online advertising, a strategic goal may be optimizing the company's advertising spending, such as by reducing the cost per order (CPO). The CPO is the sum of the advertising costs divided by the total number of purchases. Therefore, two possible resulting decision goals are reducing the advertising spending while keeping sales constant and vice versa. Therefore, related information goals include measuring the effects of reduced advertising spending on sales or the targeted exposure of online advertising activities to potential consumers to reduce scattering losses. The latter information goal is the basis for the further analysis of related data sources, spots and their characteristics. Scattering losses can be analyzed and optimized for each active advertising channel, such as paid search advertising or social media advertising. In the following example, we will focus on display advertising activities.

Based on the information goal of "reducing scattering losses of display advertising activities", we identify the related business process, which contains the process of redirecting possible customers from

third-party websites to the company's online shop with the help of display advertisements. Because the company sells products with different technical specifications, the process begins with the customer's browsing routines or internet-based information search regarding a product or service. During the search, an advertisement for the company is displayed to the potential customer, who either clicks on the advertisement or visits the online shop directly. The visit to the shop leads to a purchasing decision, which terminates the analyzed process.

This business process given the information goal serves as the basis for the following identification of related data sources and spots as described in Section 2.2. The description of each data spot and its attributes and characteristics would be beyond the scope of this paper. Therefore, we analyze only a selection of data sources sufficient to demonstrate the functionality of the process model:

- The main *internal data source* (high level) in the information search process step is the company's website respective to the company's webserver. On the middle level, the contained data spots are primarily customer reviews and clickstream data (Bucklin and Sismeiro, 2003). On the low level, which contains the data characteristics, the reviews are poly-structured (i.e., text, evaluation scheme, time of creation, and user name) and written in the customers national language . They are written on a sporadic basis. Furthermore, because they are stored on own servers, the acquisition costs are low in the first step. However, due to the low structure of text and potential noisy data, the data preprocessing is time-consuming and therefore cost-intensive. Reviews are human-generated on an individual level and are available because the product is sold in the online shop. As the second main data spot, the redirection to the company's website after clicking on advertisements creates individual user journeys (clickstream data including information of which user clicked on what type of online advertisement at which point of time and finally bought a product). These data have a high degree of structure and can be accessed free of charge because the telecommunication company in focus has its own webserver. The data are machine-generated on individual level. Therefore, less time is required for data preprocessing than for the customer reviews.
- The data sources and spots identified so far inside the company are enriched in the next step by the *external data perspective*. On a high level, websites from other online shops selling a product or service, such as Amazon.com or product review websites from magazines and product-related fora, are additional data sources. The contained data spots include the review texts and ratings, the time stamp and the reviewer's profile (e.g., number of reviews written, products reviewed so far). Compared to the review data from the company's website, the data are poly-structured, available since the product has been sold in the respective online shop and generated at irregular intervals. The information value differs significantly across reviews and is based on the length of the review, the active vocabulary used and the reviewer's intention (Mudambi and Schuff, 2010). In addition, fora may contain phony reviews by reputation management agencies that are designed to influence product sales. Therefore, the data preprocessing effort is high. The difference between internal and external reviews is the absence of an API to access and store the data. Therefore, its acquisition costs are higher than are those for internal review data, and access is not always possible due to crawling limitations.
- A *next process step* is the contact of the potential customer with a displayed advertisement (such as individual "view"-touch point events of individual users with display advertisements). Because the company has outsourced its online advertising activities, the related data source is an external advertising server. The contained data include the cookie ID, type of advertisement displayed (e.g., banner, pop-up; here, a banner), timestamp, display duration, location (URL, position on-page, and size) and whether the advertisement has been viewed (y/n) and clicked (y/n). These data have a high degree of structure and contain low to no noisy data because they are machine-generated. On the downside, the data are cost-intensive because they must be purchased from the advertising agency.
- The data source for the *final process step*, the potential conversion, is again the company's web server, which contains the same data used in the first information-gathering step (internal clickstream data). Additional data spots include the conversion (y/n), products in the shopping cart and time of a potential cart abandonment

The structured process leads to numerous potential data sources with heterogeneous characteristics that analysis may generally be useful in reducing display advertising costs. However, each of the data sources has a different expected level of contribution to the information goal. For example, the internal data sources may include directly available information

about how display advertising affected consumers' decision and buying processes, which helps companies optimize display advertising activities (Braun and Moe, 2013; Nottorf and Funk, 2013; Rutz and Bucklin, 2011a), whereas the external available data sources, such as customer reviews, only have indirect effects on the effectiveness of display advertising activities and, therefore, will not directly contribute to the information goal.

Following the principal of first considering data that are easy to generate and analyze and that are expected to contribute to the information goal, we anticipate that the *internal clickstream data* offer deep insight into consumer online clicking and purchasing behavior. Based on this clickstream data, which contain highly detailed user-level information, we are able to analyze user clicking and purchasing behavior. The results are intended to contribute to the information goal of reducing display advertising costs given the same output or the same number of sales.

3.2 Analyzing Clickstream Data

The telecommunication company in question runs multiple advertising campaigns. As discussed above, the company generates highly detailed user-level data that contain time-specific touch points for individual users with multiple advertising channels. Analyzing the advertising-specific attribution to the overall advertising success (e.g., sales) is an ongoing problem that is the focus of recent scientific research because the options for online advertising have become increasingly complex, leading to the necessity of making sophisticated decisions (Nottorf, 2013). For example, because companies run multiple online advertising campaigns simultaneously, individual consumers are often exposed to more than one type of online advertising before they click or purchase. Standalone metrics, such as click-through rates, which are the ratio of clicks to impressions, or conversion rates, defined as the number of purchases in relation to the number of clicks, are not able to realistically assign these clicks and purchases to a specific type of online advertising. These metrics neither explain the development of consumer behavior over time (i.e., a consumer is first exposed to a display advertisement, later searches for the advertised product, and finally purchases it) nor account for the potential effects of interaction among multiple types of online advertising.

(Nottorf and Funk, 2013) have recently demonstrated how having and analyzing clickstream data can explain consumer online behavior and consequently optimize online advertising activities. Therefore, we follow (Nottorf and Funk, 2013) in modeling clickstream data and analyzing individual consumer purchasing behavior. That is, we interpret all interactions with advertisements as a repeated number of discrete choices (Bucklin and Sismeiro, 2003). For example, consumers can decide whether to buy a product after clicking on an online advertisement, which results in a conversion/non-conversion decision. Note that we model the consumer choice of buying or not buying (binary choice) by incorporating the effects of repeated interaction with multiple types of online advertising as explanatory variables. As already demonstrated by (Chatterjee et al., 2003), it is useful to consider short-term advertising effects on consumers' success probabilities by adding variables to the model specification that vary across time t with each advertisement interaction (X_{ist}) as well as their long-term effects by incorporating variables that only vary across sessions s (Y_{is}). To model the individual contribution of each advertising effort and its effect on the probability that a consumer i will purchase, we specify a binary logit choice model following the specification of (Nottorf and Funk, 2013). The probability that consumer i purchases a product at time t in session s is modeled as follows:

$$Conv_{ist} = \begin{cases} 1 & \text{if user } i \text{ purchases at time } t \text{ in session } s \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

with the probability

$$P(Conv_{ist} = 1) = \frac{exp(\alpha_i + X_{ist}\beta_i + Y_{is}\gamma_i + \varepsilon_{ist})}{1 + exp(\alpha_i + X_{ist}\beta_i + Y_{is}\gamma_i + \varepsilon_{ist})}, \tag{2}$$

where X_{ist} are variables varying within (t), across sessions (s), and across consumers (i); Y_{is} are variables varying across sessions (s) and consumers (i); and α_i, β_i, and γ_i are consumer-specific parameters to be estimated.

α_i accounts for the propensity of an individual consumer to purchase a product after clicking on a respective advertisement. For example, previous research indicates that consumer responses to banner advertisements are highly dependent on individual involvement (Cho, 2003; Danaher and Mullarkey, 2003) and exhibit strong heterogeneity (Chatterjee et al., 2003; Nottorf, 2013).

To account for the effects within a consumer's current session across multiple advertising types, we follow Nottorf and Funk (2013) and define the following variables incorporated by X_{ist}:

$$\begin{aligned} X_{ist} = \{ & x_{ist}^{\text{search}}, x_{ist}^{\text{social}}, x_{ist}^{\text{display}}, x_{ist}^{\text{affiliate}}, x_{ist}^{\text{newsletter}}, \\ & x_{ist}^{\text{other}}, x_{ist}^{\text{brand}}, x_{ist}^{\text{direct}}, x_{is(t-1)}^{\text{conv}}, \text{Conv}_{is(t-1)}, \text{TLConv}_{ist} \}. \end{aligned} \tag{3}$$

We expect the effect of repeated clicks on advertisements to vary depending on the type of online advertising that is being clicked on. Thus, x_{ist}^{search}, ..., x_{ist}^{other} refer to the cumulative number of clicks on the respective type of advertisement.[1] x_{ist}^{brand} accounts for the cumulative number of brand-related interactions (e.g., the search query of the consumer included the company's name). x_{ist}^{direct} refers to the cumulative number of direct visits of a consumer (e.g., via direct type-in or the use of bookmarks). $x_{is(t-1)}^{\text{conv}}$ is the cumulative number of conversions until the consumer's last touch point $(t-1)$ in the current session s. $\text{Conv}_{is(t-1)}$ is an indicator function that assumes the value 1 if a consumer has purchased in $t-1$. TLConv_{ist} refers to the logarithm of time since a consumer's last purchase. If a consumer has not yet purchased, the variable remains zero.

The variables Y_{is} are similar to those specified as X_{ist}, but now account for the long-term, inter-session effects of previous touch points of a consumer:

$$Y_{is} = \{y_{is}^{\text{search}}, y_{is}^{\text{social}}, y_{is}^{\text{display}}, y_{is}^{\text{affiliate}}, y_{is}^{\text{newsletter}}, y_{is}^{\text{other}}, y_{is}^{\text{brand}}, y_{is}^{\text{direct}}, y_{i(s-1)}^{\text{conv}}, \text{IST}_{is}, \text{Session}_{is}\}. \quad (4)$$

y_{is}^{search}, ..., y_{is}^{other} refer to the number of clicks on respective advertisements in previous sessions. y_{is}^{brand}, y_{is}^{direct}, and $y_{i(s-1)}^{\text{conv}}$ also account for the total number of respective interactions in previous sessions. IST_{is} is the logarithm of the intersession duration between session s and $s-1$ and remains zero if a consumer is active in only one session. Session_{is} refers to the number of sessions during which a consumer has been active.[2]

3.3 Empirical Data

The dataset analyzed consists of information on individual consumers and the point in time at which they clicked on different advertisements and made purchases. The internal clickstream data were collected within a one-month period in 2013 and consist of more than 500,000 unique users. Because no information on the number of consumer sessions and their duration is accessible, we follow (Chatterjee et al., 2003) and (Nottorf, 2013) and manually define a session as a sequence of advertising exposures with breaks that do not exceed 60 minutes. We report the descriptive statistics of our final set of variables in Table 2. To test the out-of-sample fit performances of the model, we split the data into a training sample (50,000 consumers) and a test group (470,906 consumers). The dataset has been sanitized, and we are unable to provide any further detailed information on the dataset for reasons of confidentiality.

3.4 Results and Discussion

Similar to Nottorf and Funk (2013), we use a Bayesian standard normal model approach to account for consumer heterogeneity and to determine the set of individual parameters. We apply a Markov Chain Monte Carlo (MCMC) algorithm including a hybrid Gibbs Sampler with a random walk Metropolis step for the coefficients for each consumer (Rossi et al., 2005). We perform 5,000 iterations and use every twentieth draw of the last 2,500 iterations to compute the conditional distributions.

The parameter estimates for X_{ist} and Y_{is} can be found in Table 2. The mean of the intercept α_i, which accounts for the initial "proneness to purchase" (following (Chatterjee et al., 2003)), is -5.85. This results in a very low initial conversion probability of 0.29%. In contrast to the prior findings of Nottorf and Funk (2013) who modeled click probabilities, only a few significant parameter estimates exist. For example, whereas each additional click on a social media x_{ist}^{social} or display x_{ist}^{display} advertisement significantly decreases conversion probabilities within consumers' current sessions (-6.08 and -1.14), consumers' clicks on the remaining channels do not significantly influence conversion probabilities. However, although the parameter estimates of the remaining channels are not significant, they are still influencing conversion probabilities differently. For example, x_{ist}^{search} is negative, with a value of -0.58, indicating that each additional click on a paid search advertisement within a consumer's current session decreases the conversion probability. Conversely, $x_{ist}^{\text{newsletter}} = 0.48$ is positive, so each additional click on newsletter-links slightly increases the probability of a purchase.

To demonstrate how the analysis of clickstream data can optimize the display advertising efficiency, we propose a method for short-term decision support in real-time bidding (RTB).[3] Therefore, we first high-

[1] "search" refers to clicks on paid search advertisements, "social" to clicks on advertisements on Facebook, "display" to clicks on generic banner advertisements, "affiliate" to clicks on banner advertisements of the affiliate networks, "newsletter" to clicks on emails sent to consumers, and "other" to further advertisement interactions that do not belong to one of the previous groups.

[2] For a more detailed description of preparing the clickstream data for the analysis, please see (Nottorf and Funk, 2013).

[3] In RTB, display advertising impressions are bought in an auction-based process and displayed in real time on the individual consumer level. In other words, the knowledge of a consumer's success probability (such as a click or a con-

Table 2: Descriptive statistics of the variables used in the final model specification.

X_{ist} variables	Min.	Max.	Mean	Sd.	Y_{is} variables	Min.	Max.	Mean	Sd.
x_{ist}^{search}	0	86.00	0.47	2.63	y_{is}^{search}	0	5.42	0.52	1.50
x_{ist}^{social}	0	4.00	0.00	0.11	y_{is}^{social}	0	2.57	0.01	0.13
x_{ist}^{display}	0	432.00	1.60	24.05	y_{is}^{display}	0	7.23	0.25	1.30
$x_{ist}^{\text{affiliate}}$	0	148.00	0.18	4.18	$y_{is}^{\text{affiliate}}$	0	5.30	0.14	0.73
$x_{ist}^{\text{newsletter}}$	0	10.00	0.00	0.16	$y_{is}^{\text{newsletter}}$	0	3.22	0.01	0.15
x_{is}^{other}	0	6.00	0.00	0.12	y_{is}^{other}	0	3.76	0.01	0.17
x_{ist}^{brand}	0	86.00	0.84	2.62	y_{is}^{brand}	0	5.29	1.26	2.23
x_{ist}^{direct}	0	41.00	0.53	1.08	y_{is}^{direct}	0	5.29	1.08	2.13
$x_{is(t-1)}^{\text{conv}}$	0	3.00	0.01	0.11	$y_{i(s-1)}^{\text{conv}}$	0	3.22	0.02	0.22
$\text{Conv}_{is(t-1)}$	0	1.00	0.00	0.10	IST_{is}	0	7.76	3.14	4.32
TLConv_{ist}	0	7.77	0.13	1.19	Session_{is}	1	198.00	12.61	34.49

Table 3: Parameter estimates of the proposed model. We report the mean and the 95% coverage interval; significant estimates are in boldface.

X_{ist} variables	Mean	(95% cov. interval)	Y_{is} variables	Mean	(95% cov. interval)
x_{ist}^{search}	-0.58	(-1.38, 0.22)	y_{is}^{search}	0.42	(-0.41, 1.26)
x_{ist}^{social}	**-6.08**	(-7.03, -5.13)	y_{is}^{social}	**-0.84**	(-1.57, -0.11)
x_{ist}^{display}	**-1.14**	(-2.18, -0.10)	y_{is}^{display}	0.22	(-0.64, 1.07)
$x_{ist}^{\text{affiliate}}$	0.06	-0.68, 0.80	$y_{is}^{\text{affiliate}}$	-0.03	(-0.81, 0.76)
$x_{ist}^{\text{newsletter}}$	0.48	(-0.30, 1.26)	$y_{is}^{\text{newsletter}}$	-0.56	(-1.21, 0.10)
x_{is}^{other}	-0.43	(-1.16, 0.29)	y_{is}^{other}	0.11	(-0.58, 0.80)
x_{ist}^{brand}	0.64	(-0.17, 1.45)	y_{is}^{brand}	0.00	(-0.90, 0.89)
x_{ist}^{direct}	-0.64	(-1.48, 0.20)	y_{is}^{direct}	0.19	(-0.76, 1.14)
$x_{is(t-1)}^{\text{conv}}$	-0.49	(-1.41, 0.43)	$y_{i(s-1)}^{\text{conv}}$	**2.01**	(0.53, 3.48)
$\text{Conv}_{is(t-1)}$	**2.05**	(1.10, 3.00)	IST_{is}	-0.14	(-0.70, 0.42)
TLConv_{ist}	0.11	(-0.63, 0.85)	Session_{is}	-0.31	(-0.75, 0.13)

light the out-of-sample fit performance of our proposed model by predicting the actual outcome for the last available touch point of each consumer from the test data set (conversion/no conversion) and comparing them with the actual, observed choices. Furthermore, we rank all of these consumers by their individual conversion probabilities at the last touch point, separate them into quartiles, and examine how many conversions each of the quartiles actually receives (Table 4).[4] For example, the quartile with the lowest 25% of conversion probabilities (0-25%) receives 12.9% of the total 2,017 conversions that were observed at the last available touch point for each consumer from the test data set, whereas 25% of the consumers with the highest conversion probability (75-100%) receive nearly 50% of the conversions. Directing the company's bidding behavior and advertising-spending toward this upper quartile bin may lead to improved short-term decision support and potential financial savings and, thus, contribute to the overall strategic goal of reducing the CPO.

Table 4: Quartiles are grouped by predicted conversion probabilities for n = 470,906 consumers. In Scenario 1 (2), a CPC of € 0.50 (€ 0.30) is assumed to calculate the CPO.

			CPO	
Quartiles	Conv.	CVR	Scenario 1	Scenario 2
0-25%	260	0.22%	227.28 €	136.36 €
25-50%	353	0.30%	166.67 €	100.00 €
50-75%	442	0.38%	131.58 €	78.95 €
75-100%	962	0.82%	60.98 €	36.59 €
Total	2,017	0.43%	116.28 €	69.77 €

version) at any given time is vital for accurately evaluating each advertising type and appropriately adjusting financial resources.

[4]We do so following Nottorf and Funk (2013) and (Chatterjee et al., 2003) with respect to (Morrison, 1969), who suggested ranking observations in decreasing order of predicted probabilities and classifying the first x as clicks (where x is the total number of clicks observed in the hold-out sample) because the behavior to be predicted is relatively rare and the base probability of the outcome is very low. As Chatterjee et al. also emphasize, with a large number of nonevents (no conversions) and very few events (conversions), logistic regression models can sharply underestimate the probability of the occurrence of events.

Based on the forecast for each consumer-conversion probability-quartile, we can calculate the expected quartile-specific conversion rate (CVR). Let us now assume that the company in question actually engages in a RTB setting. Depending on the individual setting (i.e., the contribution margin of the advertised product), companies usually determine a specific maximum amount of money that they are willing to spend to acquire new customers (which is the maximum CPO the company is able to spend). In the following example, we consider two scenarios, each of which has a different cost per click (CPC), which results in different CPOs depending on the expected CVRs (the right side of Table 4). To be clear, let us consider an example and assume a maximum CPO of €75.00. Given that maximum,[5] we see that in Scenario 1, only the consumers within the quartile bin 75-100% should be exposed to display advertisements because the CPC of the other consumers is expected to be higher than €75.00. A company that does not have information on the clickstream data would not have exposed any consumers to display advertisements in the first scenario because the company would not have categorized consumers along their individual conversion probabilities; with €116.28, the total expected CPO is higher than the maximum CPO. In the second scenario with a decreased CPC, the company would expose all consumers to display advertisements, although only the consumers with the highest expected CVR have a CPO that is lower than the maximum CPO (€36.59).

The procedure outlined above leads to additional profit ($profit_{add}$), in contrast to a company that does not analyze clickstream data and consequently does not optimize display advertising activities. To illustrate this result for Scenario 1, we must consider the opportunity cost of a "lost" conversion ($cost_{opp}$) of a consumer whom we do not expose to display advertisements because we focus on the consumers who have the topmost conversion probabilities multiplied by the number of lost conversions ($conv_{lost}$). Simultaneously, we save on the consumers ($user_{lost}$) whom we do not expose to display advertising due to an expected CPO that is too high:

$$profit_{add} = user_{lost} * \text{CPC} - conv_{lost} * cost_{opp} \quad (5)$$

We assume that the cost of a lost conversion is equal to the maximum CPO (€75.00). Given that assumption, the expected profit is €26,828.85 for Scenario 2.[6] In the first scenario, a company that does not use the information derived from clickstreams would lose €13.286.75 because it misses 25% of the consumers with the highest predicted probabilities.[7] Please note that this profit/loss is a sample calculation and may not hold true for every hour/day iteration. Nonetheless, this example demonstrates how analyzing clickstream data contributes not only to the information goal of reducing display advertising costs but also to the overall strategic goal of reducing the global CPO.

4 CONCLUSION

The increasing amount of available data with heterogeneous characteristics regarding structure, velocity and volume hinders the selection of data for decision support purposes. The existing models primarily target the information requirement analysis for data warehouse development but do not support the data evaluation process in the early stages of data analysis for decision support.

We developed a data landscape that enhances both the data selection and the decision support process. The proposed framework incorporates the derivation of specific goals whose fulfillment enhance the decision support and the identification of related business processes as well as the selection of relevant data for each process step.

We tested the framework to enhance decision support in online advertising, partly by using approaches for information requirement analysis from the data warehouse and information system literature. Based on the derived information goal of optimizing display advertising spending, we have found that the internally available clickstream data offer deep insights into consumer online clicking and purchasing behavior. Applying the model of (Nottorf and Funk, 2013), we successfully analyzed and predicted consumers' individual purchasing behavior to optimize display advertising spending.

The utility of the process model for the development of a data landscape can be demonstrated because

[5] In a real setting, these expected CPOs should be calculated repeatedly because the parameter estimates may change over time and it is necessary to analyze the probabilities of new consumers.

[6] Note that there are additional costs (i.e., costs for data storage or for analyzing consumer-level data) that should also have been considered in the calculation above. For demonstration purposes, these costs are negligible. For example, the size of the initial dataset of 500,000 consumers is approximately 150MB, and the data storage prices for 1 GB of data are less than €0.10 at Amazon web services. While estimating the model is computationally expensive, determining the conversion probabilities is not. Therefore, we can neglect the costs for the computation of the expected conversion probability for an individual advertising exposure.

[7] $loss_{exp} = 470.906 * 0.25 * 0.50 - 962 * 75$

the model helps identity, classify, characterize and evaluate data in ways that can contribute to decision making. The characterization of data spots related to the business process fosters understanding about the data and their attributes for decision support purposes. The absence of such model results can lead to an incomplete basis for decision making. The limitation of the presented model results from the nature of processes, which have a static character and do not completely account for customer behavior, e.g., multiple runs through the process of information gathering.

The presented process model suggests different opportunities for further research. The proposed model was applied in the field of online advertising. It should also be tested in different scenarios to determine the degree of possible generalization and application-specific needs, particularly with regard to the identification of the related business process. Furthermore, the development of a graphical representation could foster the decision making process.

SCIENCE AND TECHNOLOGY PUBLICATIONS

REFERENCES

Author (2013). Big Data - Characterizing an Emerging Research Field using Topic Models (under review). *International Journal of Technology and Management*.

Braun, M. and Moe, W. W. (2013). Online display advertising: Modeling the effects of multiple creatives and individual impression histories. *Marketing Science*, 32(5):753–767.

Bucklin, R. E. and Sismeiro, C. (2003). A model of web site browsing behavior estimated on clickstream data. *Journal of Marketing Research*, 40(3):249–267.

Byrd, T., Cossick, K., and Zmud, R. (1992). A synthesis of research on requirements analysis and knowledge acquisition techniques. *Mis Quarterly*, 16(1):117–138.

Chatterjee, P., Hoffman, D. L., and Novak, T. P. (2003). Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, 22(4):520–541.

Chaudhuri, S., Dayal, U., and Ganti, V. (2001). Database technology for decision support systems. *Computer*, 34(12):48–55.

Cho, C.-H. (2003). Factors influencing clicking of banner ads on the www. *CyberPsychology & Behavior*, 6(2):201–215.

Danaher, P. J. and Mullarkey, G. (2003). Factors affecting online advertising recall: A study of students. *Journal of Advertising Research*, 43(3):252–267.

Davis, G. B. (1982). Strategies for information requirements determination. *IBM Systems Journal*, 21(1):4–30.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Gardner, S. R. (1998). Building the data warehouse. *Communications of the ACM*, 41(9):52–60.

Giorgini, P., Rizzi, S., and Garzetti, M. (2005). Goal-oriented requirement analysis for data warehouse design. *Proceedings of the 8th ACM international workshop on Data warehousing and OLAP - DOLAP*, page 47.

Golfarelli, M., Maio, D., and Rizzi, S. (1998). The dimensional fact model: A conceptual Model for Data Warehouses. *International Journal of Cooperative Information Systems*, 7(2-3):215–247.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information system research. *MIS Quarterly*, 28(1):75–105.

Hilbert, M. and López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science (New York, N.Y.)*, 332(60):60–65.

Inmon, W. H. (2005). *Building the Data Warehouse*. John Wiley & Sons, Indianapolis, 4th edition.

Kotonya, G. and Sommerville, I. (1998). *Requirements Engineering: Processes and Techniques*. John Wiley & Sons.

LaValle, S., Lesser, E., and Shockley, R. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, 52(2):21–31.

List, B., Schiefer, J., and Tjoa, A. (2000). Process-oriented requirement analysis supporting the data warehouse design process a use case driven approach. In *DEXA '00 Proceedings of the 11th International Conference on Database and Expert Systems Applications*, pages 593–603.

Madnick, S. E., Wang, R. Y., Lee, Y. W., and Zhu, H. (2009). Overview and Framework for Data and Information Quality Research. *ACM Journal of Data and Information Quality*, 1(1):1–22.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). Big data : The next frontier for innovation , competition , and productivity. Technical Report June, McKinsey Global Institute.

Mazón, J., Pardillo, J., and Trujillo, J. (2007). A model-driven goal-oriented requirement engineering approach for data warehouses. In *Advances in Conceptual Modeling – Foundations and Applications*, pages 255–264. Springer, Berlin Heidelberg.

Moody, D. L. and Kortink, M. A. R. (2000). From Enterprise Models to Dimensional Models : A Methodology for Data Warehouse and Data Mart Design Objectives of Dimensional Modelling. In *2nd DMWD*, volume 2000.

Morrison, D. G. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research*, 6(2):156–163.

Mudambi, S. and Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS quarterly*, 34(1):185–200.

Nottorf, F. (2013). Modeling the clickstream across multiple online advertising channels using a binary logit

with bayesian mixture of normals. *Electronic Commerce Research and Applications*, (Article in Advance).

Nottorf, F. and Funk, B. (2013). The economic value of clickstream data from an advertiser's perspective.

Rossi, P. E., Allenby, G. M., and McCulloch, R. E. (2005). *Bayesian statistics and marketing*. Wiley, Hoboken, NJ.

Rutz, O. J. and Bucklin, R. E. (2011a). Does banner advertising affect browsing for brands? clickstream choice model says yes, for some. *Quantitative Marketing and Economics*, pages 1–27.

Rutz, O. J. and Bucklin, R. E. (2011b). From generic to branded: A model of spillover in paid search advertising. *Journal of Marketing Research*, 48(1):87–102.

Stonebraker, M. and Robertson, J. (2013). Big data is 'buzzword du jour;' CS academics 'have the best job'. *Communications of the ACM*, 56(9):10.

Stroh, F., Winter, R., and Wortmann, F. (2011). Method Support of Information Requirements Analysis for Analytical Information Systems. *Business & Information Systems Engineering*, 3(1):33–43.

Winter, R. and Strauch, B. (2003). A method for demand-driven information requirements analysis in data warehousing projects. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, number section 2.

Winter, R. and Strauch, B. (2004). Information requirements engineering for data warehouse systems. In *Proceedings of the 2004 ACM symposium on Applied computing - SAC '04*, New York, New York, USA. ACM Press.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS