Analysis, Visualization and Exploration Scenarios: Formal Methods for Systematic Meta Studies of Big Data Applications

Klaus P. Jantke(✉) and Jun Fujima

ADISY Consulting GmbH & Co. KG,
Frauentorstraße 11, 99423 Weimar, Germany
{klaus.p.jantke,jun.fujima}@adisy.de

Abstract. There is not much doubt that the progress of information and communication technologies, the computerization of all areas of life, and the engagement of increasingly more human beings in the usage of computerized gadgets results in an enormous growth of data available. The data available bear potential for solving urgent problems such as, e.g., forecasting of the spreading of diseases and related prevention, the estimation of the impact of forthcoming disasters like sinkholes, earthquakes, and tsunamies and the preparation of adequate measures, or the development of more precise weather forecasts, to name just a few. Data need to be analyzed. There is a manifold of methodologies and tools to support human exploration. How to do this is treated as an art. But scenarios of data analysis, visualization, and exploration are not yet considered. The present work is intended to fill the gap and to contribute to a paradigmatic shift from the art to a science.

1 Introduction

Human-computer interaction for purposes of *data analysis, visualization, and exploration*–to shorten the expression, this will subsequently be abbreviated by DAVE, in many places–is overwhelmingly manifold and largely unforeseeable. Processes of discovery are, to some extent, cases of serendipity (see, for instance, Schubert (2013) and Jantke and Fujima (2015)).

On the one hand, there are increasingly many efforts world-wide to improve the analysis of big data and high expectations of the effects in literally unlimited fields of science, technology, and the society as a whole.

On the other hand, although studies of big data are intensified, there are no attempts at all to better understand the process of doing so. To the author's very best knowledge, there is not yet any systematic and theoretically well-founded investigation of *scenarios* of data analysis, visualization, and exploration (subsequently more shortly named 'scenarios of DAVE' or 'DAVE scenarios').

This paper is introducing the term and the terminology, is explaining the methodology, and aims at a demonstration of applications.

E. Grant et al. (Eds.): ISIP 2015, CCIS 622, pp. 107–127, 2016.
DOI: 10.1007/978-3-319-43862-7_6

1.1 Motivation

To set the stage for appropriate formalizations, for problem representation and process description, and for systematic reasoning focusing human (re)search and its results, we need a firm scientific basis.

As Norbert Wiener put it nicely, “Der Gedanke, daß Information in einer sich ändernden Welt ohne merkbare Minderung des Wertes gestapelt werden kann, ist falsch.” (Wiener (1958), p. 122) To translate Wiener’s message shortly, storying large amounts of information is bringing with it a severe loss of value.

So, what are human beings doing when searching big data and, in particular, what are they looking for ...? Do they hope to see something unforeseeable? Do they dig for golden nuggets of information or even knowledge?

To Nobel Prize winner Albert Szent-Györgyi are ascribed the appealing words that “discovery consists of *seeing what everybody has seen* and thinking what nobody has thought” (emphasis by the authors). However appealing, this seems to contradict the current practice of big data analysis, visualization, and exploration in which humans strive hard to look at data–heterogeneous data from largely varying sources, in particular–to see data differently. DAVE scenarios aim at *showing much more than anybody has ever seen before*.

The authors oppose as well the opinion that big data analysis is digging for golden nuggets of information (Veluswamy (2008), Zhang and Zhou (2004)). The saying that “visualization exploration is the process of extracting insight from data via interaction with visual depictions of that data” (Jankun-Kelly et al. (2007), p. 357) is a similar misconception. Instead of squeezing insights out of the data, it is a creative process of model formation based on incomplete information, very much like theory induction Popper (1934).

Seen from the perspective of serendipity, knowledge discovery based on big data is an art. In his 1974 Turing Award lecture, Donald Knuth said that “the science without the art is likely to be ineffective; the art without the science is certain to be inaccurate” (Knuth (1974), p. 37). Seen from this point of view, the present work is intended to be some contribution toward transforming the art into a science. Scenarios of DAVE are among this science’s principles of work.

With the above perspective in mind, what may be the goal of systematizing the involved creative work of big data analysis, visualization, and exploration? Even more fundamentally, is it really appropriate to aim at a formalization of (some of) the intellectual processes taking place when dealing with big data? This sounds like a question for what we nowadays call Artificial Intelligence.

To put a reliable cornerstone for our endeavor, Norbert Wiener is providing an interesting hint: “If I were to choose a patron saint for cybernetic ... I should have to choose Leibniz” (Wiener (1962), p. 12). What did apply to Cybernetics then, does apply to Artificial Intelligence, i.e. to automated reasoning, nowadays.

Leibniz describes the vision that philosophers–instead of arguing–write down their respective positions and find out who is right by calculation: “calculemus” (see Gerhardt (1849), vol. 7, p. 200). Similarly, this paper aims at representing DAVE scenarios to provide a foundation of automated reasoning about big data.

1.2 Related Work

As Jankun-Kelly et al. put it, *the human-computer interaction (HCI) community has long been concerned with the low-level mechanics of user interface interaction* (Jankun-Kelly et al. (2007), p. 359). They characterize their own work as being situated "between the low-level syntactic models and high-level semantic models of user interaction" (Jankun-Kelly et al. (2007), p. 359).

Jankun-Kelly et al. see visualization exploration as a process of parameter modification (Jankun-Kelly et al. (2007), Sect. 3, Fig. 2 on p. 360). Accordingly, the interaction processes under consideration are sequences of *parameter derivations* (for an illustrative example see Jankun-Kelly et al. (2007), p. 364, Fig. 5).

The present authors, however, go beyond the limits of such a perspective. The higher expressiveness of the present approach is based on features of meme media (for details, see below) that allow for the decomposition of visualizations.

In Amar et al. (2005), the authors contrast "representational primacy", a data-centric view of information visualization that relies on user skills to generate insight, to "analytic primacy" that puts the human user in focus.

Amar et al. believe that *in general, information visualization can benefit from understanding the tasks that users accomplish while doing actual analytic activity. Such understanding achieves two goals: first, it aids designers in creating novel presentations that amplify users' analytic abilities; second, it provides a common vocabulary for evaluating the abilities and affordances of information visualization systems with respect to user tasks* (Amar et al. (2005), p. 111).

Toward this goal of putting human activity in focus, they present a set of ten low-level analysis tasks that largely capture people's activities while employing information visualization tools for understanding data.

These tasks–there are, among others, "Find Extremum", "Determine Range" and "Find Anomalies" (Amar et al. (2005), p. 114)–are of a much more rough granularity than what is in focus in the present paper. Consider the task "Correlate" sketched vaguely as follows: "Given a set of data cases and two attributes, determine useful relationships between the values of those attributes" (Amar et al. (2005), p. 114).

The approach by Heer, Mackinlay et al. is characterized by these authors' interest in tools that facilitate iterative forms of interaction Heer et al. (2008). They focus on "the design of history mechanisms for information visualization" (Heer et al. (2008), p. 1189).

At a first glance, their basic concepts are very close to the present authors' concept of play states (see below). However, the motivation is completely different and, thus, leads to different investigations and results. Heer, Mackimlay et al. explicitly visualize interaction histories to extend the data visualization by an extra visualization of the user's interaction history (Heer et al. (2008), Fig. 2 on p. 1192).

There are some doubts that substantially extending visualizations makes exploratory analysis significantly easier. Therefore, the present authors study interaction histories, but refrain from revealing the history representations to the human users. Cognitive effort and cognitive load must be kept low.

In the present approach, interaction scenarios are intellectual tools on a meta-level that are subject to studies in their own right.

2 Toward the Introduction of Formal Concepts

When describing human behavior in formal terms, there is a need to formalize elementary activities. There is rarely an optimal level of abstraction (see Sect. 4 for a more detailed discussion). As a consequence, there are varying approaches. This leads to the necessity to discuss several variants and to explain choices. Issues under discussion may be of varying complexity. Illustrations might help. Therefore, the authors decided to base the subsequent main part of this paper on the second author's implementation of a prototypical tool for data analysis, visualization, and exploration within a certain context of business intelligence. Part of the conceptualization to come will be illustrated by means of screenshots taken from this implementation when running.

All elementary human activities to be introduced subsequently will be named. To keep the formalization short, single letters such as q to indicate querying and f to indicate filtering, for instance, are preferred. All names of elementary actions are collected in a set denoted by M. As usual, M^* denotes the set of all finite strings over M including even the empty string ε. To exclude the empty string, we set $M^+ = M^* \backslash \{\varepsilon\}$.

Strings $\pi \in M^+$ denote sequences of human activities. To make this explicit, we sometimes use notations like $\pi = \mu_1 \ldots\ \mu_n$ where every μ_i belongs to M.

Among the elements of M, there are actions such as *extraction* and *inspection* which may appear less intuitive than, e.g., *filtering*. Those human activities in the process of big data analysis, visualization, and exploration which possibly need some more detailed illustration will be introduced by means of exemplifying webble manipulations.

The webble technology according to Kuwahara and Tanaka (2010) has been chosen as an appropriate underlying knowledge media technology Tanaka (2003). The following Sect. 3 introduces webble technology in some depth.

Webbles are objects on the human-machine-interface which have a certain Model-View-Controller architecture. They are manipulated on the screen and some of the manipulations mean certain activities abstractly represented in M. Other typical activities are pushing buttons, e.g., and typing in terms specifying queries or filters.

Elementary activities of interest are abstractly represented by elements of M. Finite sequences $\mu_1 \ldots\ \mu_n \in M^+$ formally represent particular human behavior in the course of data analysis, visualization, and exploration.

In dependence on the available opportunities, human activities are of largely varying significance. Modal logic Blackburn et al. (2001) provides appropriate ways of reasoning about alternatives of behavior. Part of this reasoning–in full agreement with Leibniz' vision and program–may be computerized.

All conceptualization and terminology shall be seen in the light of reasoning about human behavior in dependence on certain contexts.

Before we can dive deeper into formal representations of human behavior and logical reasoning, we need to summarize webble technology in Sect. 3 and to complete the conceptualization in Sect. 5 for which Sect. 4 is intended to provide some intuitive approach.

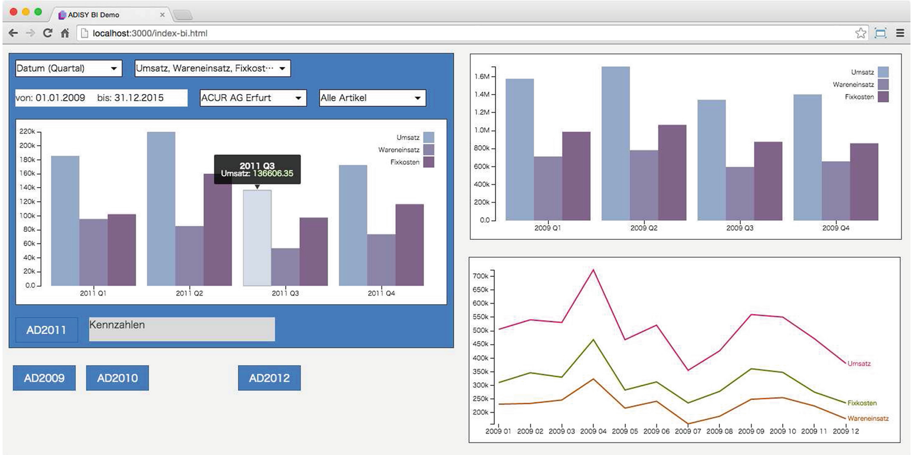
3 Webble Technology for Big Data Analysis

Webble technology Kuwahara and Tanaka (2010) is the latest implementation of the meme media architecture Tanaka (2003). It provides a web-based middleware platform where users can make use of published media objects.

In the webble platform, knowledge resources including texts, images or videos as well as application tools, databases, or services are represented as visual media objects called *webbles*. Users cannot only consume webbles as elementary media objects but also reuse them as components of more complex applications by combining them at runtime environment.

The feature of flexible customization or composition has a beneficial effect on data analysis, visualization, and exploration tasks. It is not trivial to select proper combinations of target data, statistical methods, or visualization techniques from uncountably many possibilities. It may depend on tasks as well as the domain of the tasks. Therefore, it is helpful to provide a flexible environment for publishing elementary functionality as components and combining those components to construct data analysis tools on demand.

To demonstrate the potential of the webble technology in data analysis, visualization, and exploration process, we have developed a prototypical application (Fig. 1) based on the webble platform implemented in Fujima (2013).



ADISY BI Demo

localhost:3000/index-bi.html

Datum (Quartal)

Umsatz, Wareneinsatz, Fixkost···

von: 01.01.2009 bis: 31.12.2015

ACUR AG Erfurt

Alle Artikel

2011 Q3
Umsatz: 136606.35

Kennzahlen

Fig. 1. ADISY Business Intelligence Demo based on the webble technology

Webbles are persistent objects and each webble has its Model-View-Controller structure internally. The view is implemented as a custom HTML element that works as a wrapper of a variety of types of computational resources. It provides a standard set of user manipulations such as *select*, *deselect*, *move*, *copy*, *paste*, *peel*, and *drag-and-drop*.

The view also exposes *slots* which work as input/output ports of communication between webbles. Slots hold data or property values of webbles. When a new value is submitted to a slot, the owner webble changes its behavior according to the submitted value.

By pasting one webble on another through a drag-and-drop operation, the pasted webble becomes a child of the other webble. With this manipulation, a user can combine webbles physically. Further, the user can define s slot connection between physically combined webbles to define a communication channel. Through the slot connection, two webbles communicate with each other and work in a coordinated manner by sending or receiving some values.

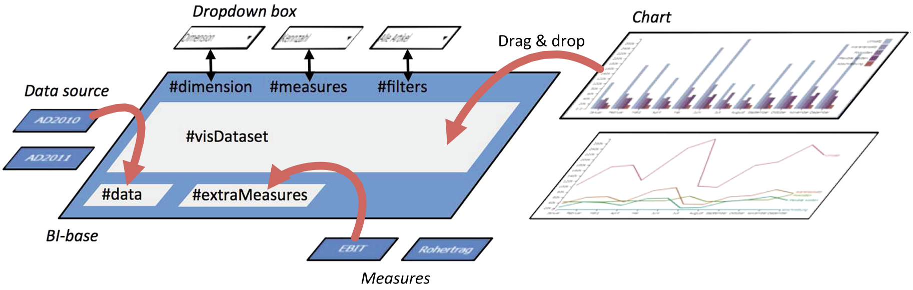


Fig. 2. The composition structure of the ADISY Business Intelligence Demo

The implemented application is for exploring sales data of a company. It mainly consists of BI-base webble, data source webbles, chart webbles, measure webbles, and some GUI components (Fig. 2).

The main functionality is implemented as the BI-base webble. It has a basic data manipulation functionality of multi-dimensional data to convert source data to the form that fits to the input of chart components. It has *#data* slot to receive the source data. When a certain data is submitted to the slot, BI-base analyzes them and automatically detects possible dimensions and measures of the original data. The detected dimensions and measures are held in the *#dimension* and *#measures* slots, respectively. The dropdown box webbles are connected to these slots as input interface, so users can easily select a dimension and multiple measures to aggregate the source data with a certain view point. As soon as a user changes these parameters, the BI-base makes data conversion and the converted result is set to *#visDataset*.

Users can connect data source webbles to specify the target data source and chart webbles to make a visualization of converted data. Drag-and-drop manipulation of webbles does all these connections, so users don't have to connect manually slots in the process of data exploration (a key feature briefly named *auto-connect*).

4 Scenarios of Playing Digital Games

In the preceding sections, we have set the stage for the conceptualization which is intended to be the main contribution of the present paper (see Sect. 5 below). The conceptualization's formalisms will allow for a computerization of a certain part of the reasoning process based on modal logic (Sect. 6).

Before going into all the details of the formalism, the authors feel the need to explain where the present approach comes from. It has been introduced for the purpose of analyzing and understanding human-computer interaction in an area where the interaction is particularly intense and the behavior of different human beings may be largely varying: *playing digital games* (see Jantke (2009)).

This application domain is motivating some of the notations. M contains all the elementary activities of interaction; the letter is intended to resemble the term *move*. For the same reason, elements of M are usually denoted by μ, possibly with indices for decoration. Finite sequences of those elements represent (parts of) game play and, hence, are denoted by π, as well with indices, if needed. In dependence on the game mechanics, some sequences of actions (moves) may occur, whereas others do not. For talking about, we denote any fixed game by G. All finite sequences within M^+ which represent admissible sequences of playing the game G from the beginning to the very end are collected in a set $\Pi(G) \subseteq M^+$. The letters π and Π are chosen to resemble the term *play*. Accordingly, the elements of $\Pi(G)$ are called *play states* of G.

Because every digital game–naturally–is a computer program, $\Pi(G)$ may be seen as a formal language Hopcroft et al. (2001). In some sense, the game serves as a grammar able to generate every string in $\Pi(G)$.

This point of view is particularly useful when pondering the varying levels of abstraction. What is reasonably seen as an action? And what, in contrast, is either too fine or too rough? When analyzing, visualizing, and exploring game playing behaviors, there are different *layered languages of ludology* Lenerz (2009). Between these language levels, there do exist mappings up and down. Actions on a higher layer have an interpretation by a sequence of actions on a lower layer. Vice versa, some sequences of actions on a lower layer establish some meaning on a higher layer. Similar questions are of great relevance to the present work.

The application area of playing digital games makes some key issue obvious: Many of the potential sequences of human game play in $\Pi(G)$ will never happen. There is the need for another concept representing what may really take place. $\Psi(G) \subseteq \Pi(G)$ denotes the set of all those sequences of game play which really occur when humans play the game G. Usually, there is a big difference between $\Psi(G)$ and $\Pi(G)$. For real digital games, $\Psi(G)$ can hardly be a formal language.

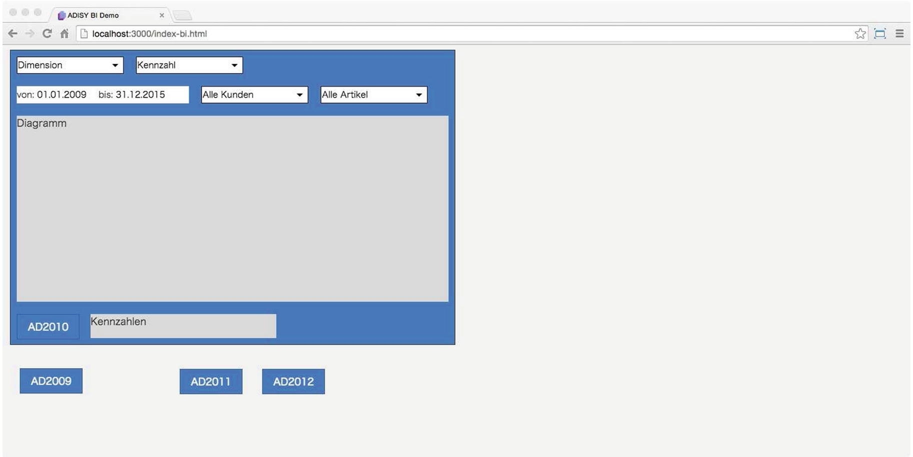
To illustrate the expressiveness and the reach of the present formalization when applied to games, we discuss some example. The difference between $\Psi(G)$ and $\Pi(G)$ allows for the precise characterization of challenges in game design.

In a play state π, some move μ is enforced, if $\forall \pi' \in \Pi(G)(\pi \preceq \pi' \rightarrow \pi\mu \preceq \pi')$ holds, where $\preceq$ indicates that the left string is an initial segment of right one. Now, contrast this condition to $\forall \pi' \in \Psi(G)(\pi \preceq \pi' \rightarrow \pi\mu \preceq \pi')$ and ponder the challenge of a design in which the second formula holds, but the first does not.

5 Formalisms of Analysis, Visualization, and Exploration

The preceding Sects. 1 and 4 provide a very first impression of the present approach which begins with a selection of what to speak about: elementary human-machine interactions. The set of these activities is named M.

For DAVE scenarios, we may assume a finite collection $\mathcal{D}$ of databases taken into account. To name these databases, we choose D_1,...,D_k. One may think of $\mathcal{D} = \{D_1, \ldots, D_k\}$. For simplicity, the action of selecting a certain database for access is simply denoted by the database's name. Thus, M contains all D_δ.



von: 01.01.2009 bis: 31.12.2015

Alle Kunden

Alle Artikel

Fig. 3. Access to a database "AD2010" by dragging and dropping the proxy webble

It depends on the functionality and on the implementation of tools for big data analysis, visualization, and exploration whether or not the selection of a particular database comes with a default visualization and/or a default query and/or a default filter. If all this does not hold, the access to a database does not directly result in some visualization (as on display in Fig. 3).

The ADISY Business Intelligence Demo implementation will be used for purposes of illustration subsequently. Figure 3 above shows the result after clicking one of the four database proxy webbles sitting in a row next to each other, then dragging the one selected over the business analytics tool and dropping it into the input place in the left lower corner. This elementary action is denoted by D_{AD2010}. There are four of them: $D_{\text{AD2009}}, D_{\text{AD2010}}, D_{\text{AD2011}}, D_{\text{AD2012}} \in M$.

After selecting a database, one may restrict the amount of records under consideration by a query.

In the present case study, we slightly suspend the focusing of investigation. Instead, we discuss the selection of some visualization as on display in Fig. 4. There are currently four types of visualization available which we shortly name Group, Line, Pie, and Table. In formal terms, $v_{\text{Group}}, v_{\text{Line}}, v_{\text{Pie}}, v_{\text{Table}} \in M$.

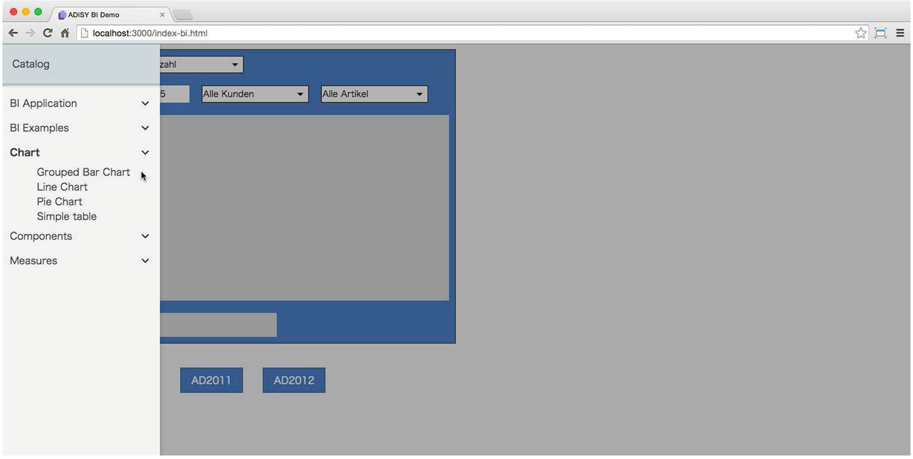


Fig. 4. Selection of the visualization "Grouped Bar Chart" formally named v_{Group}

Every visualization comes with, first, some default filtering and, second, some default rendering which determine *what* to show and *how* to show it. In the present application case, the default filtering shows just the number of records.

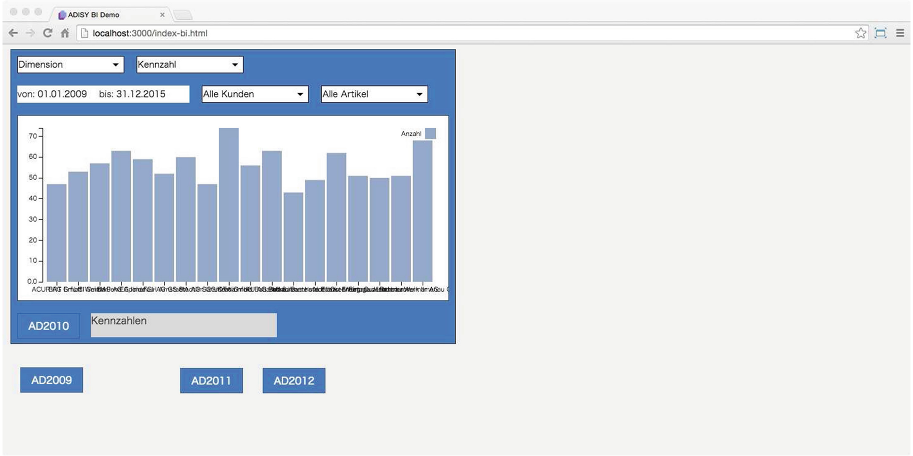
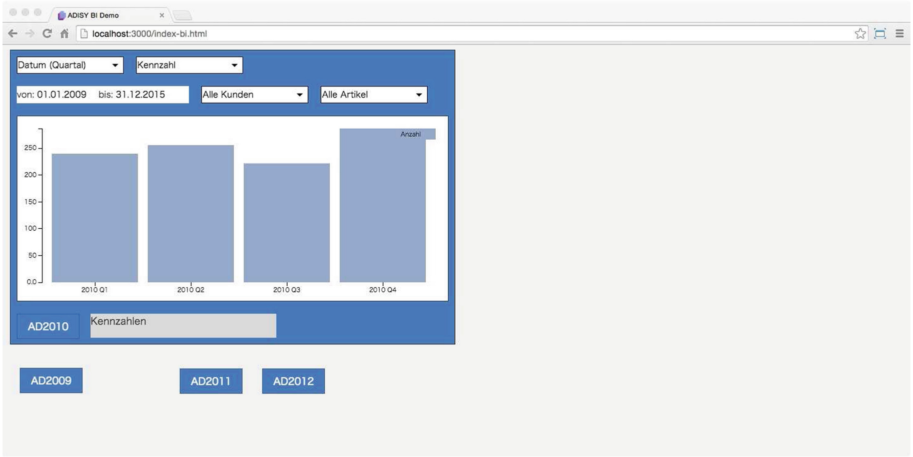


Fig. 5. Data visualization by a "Grouped Bar Chart" with its default rendering

The selection of a particular visualization shows the data in the default rendering as on display in Fig. 5. Frequently, users consider the initial rendering inappropriate and modify it. Because rendering is technically quite involved, we do not go into further details. The investigation of variants of renderings is worth some extra effort and should be accompanied by a sufficiently detailed practice.



ADISY BI Demo

localhost:3000/index-bi.html

Datum (Quartal)

Kennzahl

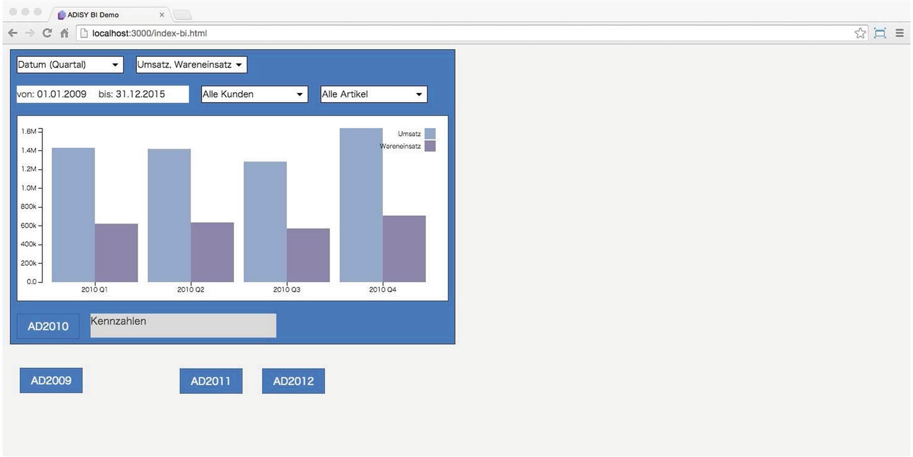
von: 01.01.2009 bis: 31.12.2015

Alle Kunden

Alle Artikel

Fig. 6. Filtering and aggregation by selection of the “Datum (Quartal)” option

Instead, we put some more emphasis on filtering as shown on the present page. The above screenshot in Fig. 6 shows an aggregation which is a particular form of filtering. The number of records remains the unchanged measure shown.



ADISY BI Demo

localhost:3000/index-bi.html

Datum (Quartal)

Umsatz, Wareneinsatz

von: 01.01.2009 bis: 31.12.2015

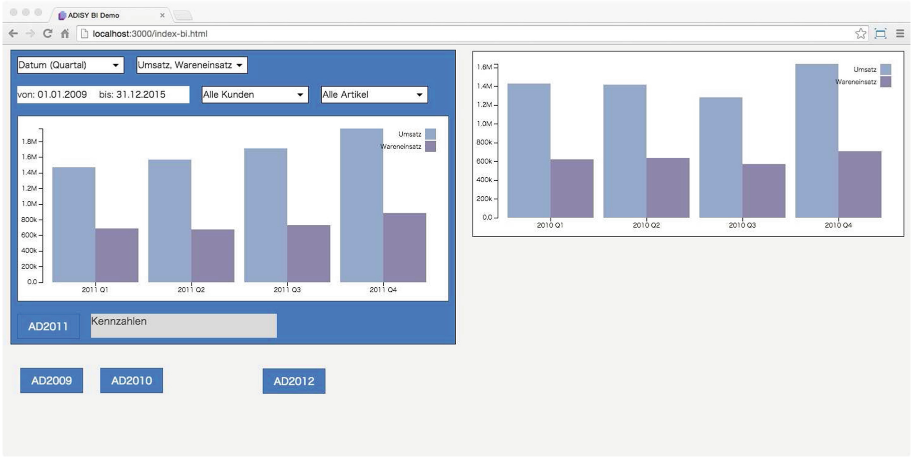
Alle Kunden

Alle Artikel

Fig. 7. Filtering by means of selecting the two measures “Umsatz” and “Wareneinsatz”

There are many intuitive ways of filtering. From Figs. 6 to 7, the user has selected the measures “Umsatz” and “Wareneinsatz”.

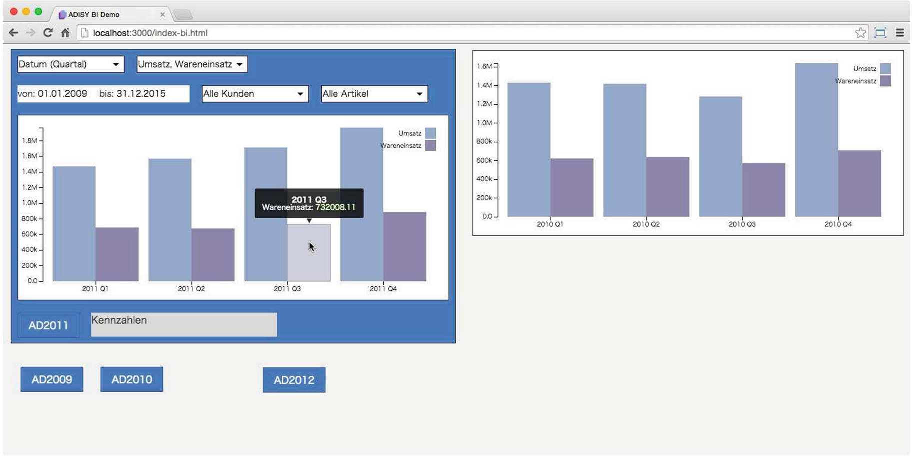
The data are worth some closer inspection. For this purpose, one may click the data visualization and drag a copy of the grouped bar chart off the blue frame of the tool. Copying is another type of elementary action (see Fig. 8).



von: 01.01.2009 bis: 31.12.2015

Fig. 8. Taking a copy of the 2010 data visualization and accessing the 2011 database

After taking the copy of the data of 2010, it makes sense to have a closer look for the same data from another year. This means another database access as on display in Fig. 8 (see lower left corner of the tool).



von: 01.01.2009 bis: 31.12.2015

Fig. 9. Inspection of slighter differences between the data from two different databases

Some differences are easy to spot. Others may need some closer inspection. Opening tooltips as shown in Fig. 9 is a method of inspection.

Within the framework of the present scenarios of data analysis, visualization, and exploration, *inspection* is another type of elementary actions. For a more detailed description of inspections, decomposable visualizations are advantageous.

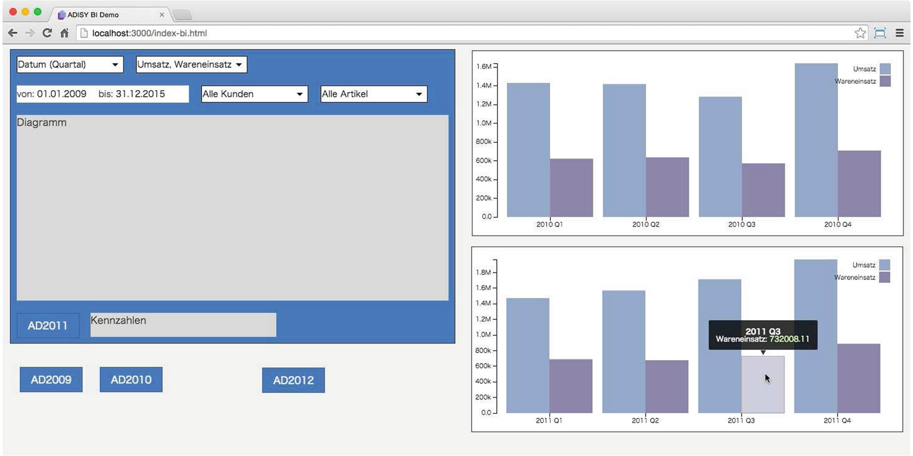


Fig. 10. Two copies of related visualizations put aside for an in-depth comparison

An essential step of data exploration is comparison of varying data presented in a similar form. To support this, the ADISY Business Intelligence Demo implementation allows for arbitrarily many copies of visualization webbles and their related arrangement on the screen (Fig. 10).

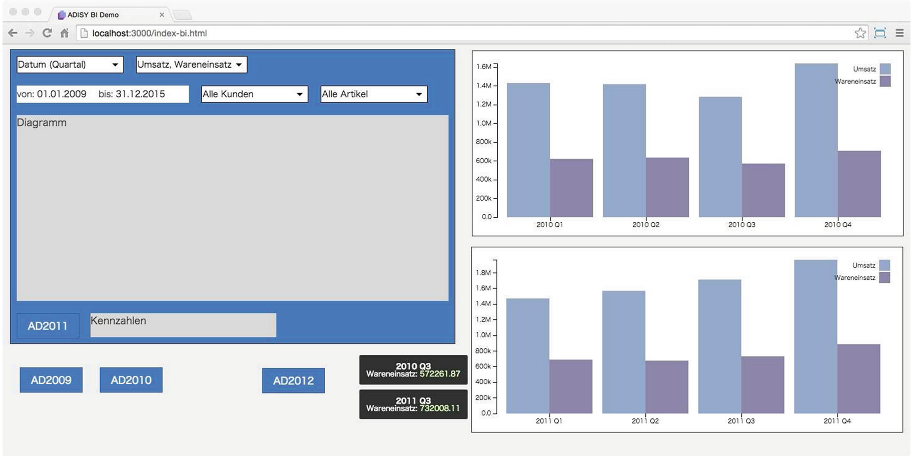


Fig. 11. Two tooltips extracted for post-processing at another place and time

Webble technology offers appropriate features to support the extraction of building blocks such as tooltips, e.g., which carry possibly valuable information. The extracted objects being webbles as well–like the two tooltip objects in the above Fig. 11–may be processed by other webble-based tools.

The screenshots on display in the series of Figs. 3, 4, 5, 6, 7, 8, 9, 10 and 11 exemplify a certain process of human-computer interaction aiming at data analysis, visualization, and exploration which may be abstractly described by a sequence of finite length built over M or, in other terms, by some string π of M^*.

According to the preceding explanation accompanying the above sequence of figures, this string is of the particular form

$$D_{\mathrm{AD2010}}\ v_{\mathrm{Group}}\ f^{0}_{\mathrm{Group}}\ r^{0}_{\mathrm{Group}}\ f^{\mathrm{Datum(Quartal)}}\ f^{\mathrm{Umsatz,Wareneinsatz}}\ \ldots$$

which will be continued after a short, but necessary supplementary discussion.

An action of accessing a database such as D_{AD2010} does not need any further specification. The selection of a particular visualization method such as v_{Group} may bring with it some default filtering and/or some default rendering.

In contrast, other elementary actions are ambiguous. When making a copy, it may be necessary to name the object which is duplicated. When inspecting a certain part of a media object, it may be necessary to name this part explicitly.

Consequently, it may be sometimes very difficulty to specify with sufficient precision what may possibly occur as an element of M.

This is the point where the choice of meme media technology, in general Tanaka (2003), and of contemporary webble technology based on HTML5, CSS and JavaScript, in particular Fujima (2013), turns out to be valuable.

When the digital object copied is a webble, this allows for a sufficiently clear syntactic representation. When the object which occurs in response to an inspection activity is a webble, this allows for a precise specification of action. Furthermore, this does allow for extraction as well.

To manipulate webbles (see Sect. 3), one selects a particular webble and, then, manipulates the selected object as desired, e.g., by peeling it off from the compound webble hosting it and moving it to another place (for shortness, we call this *extraction*), by drag and drop over another webble, by duplication, or by any other admissible activity. This does apply to all actions including D_{AD2010} and v_{Group} which occur in the string above.

When accessing a database by means of an action like D_{AD2010}, there is no need to mention the click before. Notation is simplified by dropping unnecessary details. However, there is no ideal level of granularity as we know from related studies of representing game play (see Sect. 4 and Lenerz (2009), especially).

In other cases, however, making the selection click explicit helps to avoid misunderstanding and to resolve conflicts.

Therefore, we introduce an action s representing the selection of an object, i.e., a webble. This action's parameter is the name (the identifier) of the webble selected. Consequently, M contains as many potential actions of the form $s(\ldots)$ as there are webbles in use. This allows for continuing the string shown above.

$$\ldots s(\mathrm{gbc}_1)\ ex\ D_{\mathrm{AD2011}}\ s(\mathrm{gbc}_2)\ s(\mathrm{gbc}_3)\ in\ s(\mathrm{gbc}_2)\ ex\ s(\mathrm{gbc}_2)\ s(\mathrm{gbc}_3)\ in\ \ldots$$

where names such as gbc_1 (for grouped bar chart) are identifiers of webbles. This represents the human actions leading to the situation on display in Fig. 10. A few more steps of selection and extraction bring us from Figs. 10 to 11.

To sum up intermediately, human behavior of data analysis, visualization, and exploration taking place in a possibly longer interaction with certain tools is represented by a string of symbols. Every symbol represents an action which is considered elementary. The set of symbols taken into account is denoted by M.

It depends on the available tools and their functionalities as well as on the focus of investigation what is considered relevant to be represented in the set M.

The deployment of webble technology for providing flexible environments tailored toward effective data analysis, visualization, and exploration processes brings with it some hints about what to represent: webble manipulations.

The following Table 1 summarizes a minimal set M of elementary actions.

Table 1. A set of elementary actions underlying the formalization of DAVE scenarios

Symbol	Meaning	Comment/Explanation
D_n	Database access	The index n names the database
q	Query	
f	Filter	
v	Visualization	Selecting a type of visualization
r	Render	Determining the look of a visualization
$s(n)$	Select	The parameter n names the object
in	Inspect	Searching by digital manipulation
ex	Extract	Peeling off a webble and putting it aside

Those readers who are experienced in the field of data analysis, visualization, and exploration as well as those readers who are familiar with webble technology might easily come up with further elementary actions missing in the table above. It seems highly desirable to see all standard webble manipulations (see Sect. 3) as elementary actions.

However, for the introduction of DAVE scenarios and for an investigation of this approach's reach, the actions listed above are sufficient.

The symbols named in the table form the set M. User behavior is abstractly described by sequences $\pi \in M^+$ of finite length. Very similar to the area of game play (see Sect. 4), for every environment serving the purpose of data analysis, visualization, and exploration–like the ADISY Business Intelligence Demo–there are sequences which may occur and others which are technically impossible. Those strings which are possible form a set $\Pi \subseteq M^+$. Many of the interactions which are possible never happen. The strings which occur form Ψ.

The set Ψ of strings over M which contains abstract descriptions of what humans really do in the course of data analysis, visualization, and exploration is the field of study. DAVE scenarios are intended to understand what is in Ψ. Scenarios are initial segments of strings in Ψ. The set of scenarios is named Σ and formally defined as $\Sigma = \{\, \sigma \mid \exists \pi \, (\pi \in \Psi \wedge \sigma \preceq \pi) \,\}$.

6 Reasoning About Search and Research Behavior

Underlying every DAVE scenario, there is static knowledge about the domain and about the tools at the user's fingertips. The user's behavior is represented by some string π of Ψ. This may be seen as the relevant dynamic knowledge.

Prior work on, so to speak, scenarios of game play (see Sect. 4 above) has revealed the potential of the approach. An analysis of strings π representing game playing behavior lead to a characterization of the players' mastery of crucial game features and, thus, of learning effects induced by game play Jantke (2012).

In the present section, the authors confine themselves to a survey of the essentials of the logical reasoning approach.

When investigating human behavior and studying insights which may be deduced from human behavior, there is always the above mentioned static background knowledge behind. For a short formal treatment, this basic knowledge is denoted by BK. Assume a particular statement expressed in logical terms by a formula φ. Assume furthermore some recently observed human behavior represented by a string $\pi \in M^+$. The question of interest is whether or not the statement φ can be deduced from π. In logical terms, the expression is $BK \cup \{\pi\} \models \varphi$.

Because all reasoning takes place in a fixed context in which the background knowledge can be assumed to be fixed, one may simplify the terminology by dropping BK. The problem in the simplified notation is the question for $\pi \models \varphi$.

To ease the readers access, the present formal introduction is interrupted by a short illustration. The intention is to show how to deduce statements from observed behavior. A few particularly simple cases are sketched.

First, imagine a string π in which a very long subsequence of rendering actions occur, one rendering followed by the other. This may be interpreted as the human user starring at the same data and step by step looking at the data differently. For illustration, one may look at data visualizations such as in Tanaka and Sugibuchi (2001), Fig. 1, Ito et al. (2006), Fig. 5, Ito et al. (2011), Fig. 1, and others. It is very easy to imagine that humans look at the data representation turning it backwards and forwards, to the left and to the right, doing so repeatedly. Long sequences of subsequent renderings are an indicator of humans being lost in the data, so to speak. In combination with the actions following the rendering sequences, one may draw conclusions about success or failure.

As a second example, imagine a string in which substructures of the form $D_{\ldots}\ s(\ldots)\ s(\ldots)\ in$ occur immediately one after the other, where the database changes from one substring to the next. This leaves the impression of a stringent inspection. But there is no way to say anything about the success. Assume, instead, that the repeatedly occurring substrings are all of the extended form $D_{\ldots}\ s(\ldots)\ s(\ldots)\ in\ s(\ldots)\ ex$. Every low level step of analysis, visualization, and exploration ends with the extraction of some materialized piece of information.

Apparently, we are talking about some type of patterns or instances of patterns, resp. (see, e.g., Angluin 1980), Jantke (2012), Jantke and Arnold (2014)). A more systematic study is beyond the limits of this introductory contribution.

To continue the more formal investigation of abstractly represented human behavior, recall that every string $\pi \in M^+$ may be explicitly written as a finite sequence of elements of M, i.e., $\pi = \mu_1 \ldots \mu_n$.

This leads to the fundamental question of how to interpret a human user's action μ_{n+1} after $\mu_1 \ldots \mu_n$ has been observed so far.

In case the user's action was enforced without any opportunity left, one can not draw any conclusion from the execution of action μ_{n+1} after $\mu_1 \ldots \mu_n$. Consequently, logical reasoning intended to understand and, possibly, to evaluate a human user's activities needs to consider alternative behaviors.

Clearly, the preferable formal apparatus to deal with *possibility* vs. *necessity* is modal logic Blackburn et al. (2001).

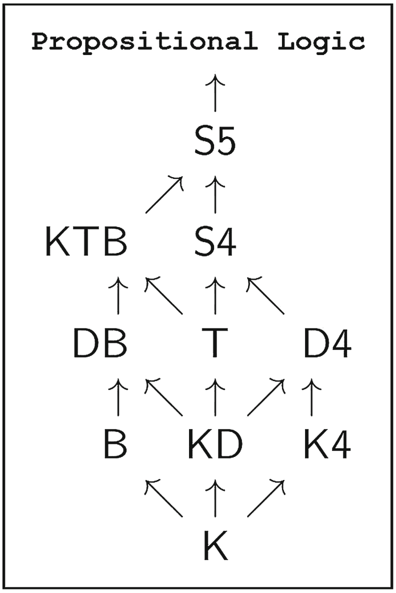


Fig. 12. A hierarchy of modal logics

There is a variety of modal logics characterized by modal operators and certain constraints of relations among them (Fig. 12).

However different, the core is built by the two operators $\diamond$ and $\Box$ meaning possibility and necessity, respectively.

It is custom to assume the standard relationship $\Box\varphi \iff \neg\Diamond\neg\varphi$ for all propositional formulas φ.

Here is no space to fully lay out modal logics for the purpose of reasoning about DAVE scenarios. We confine ourselves to the essentials.

In modal logics, the validity of propositional formulas is defined as known from conventional logics. The validity of possibility and, thus, of necessity (according to the standard relation mentioned above) is determined by means of a relation between potential models. Ψ is the set of models of interest. Therefore, one defines this basic relation R over $\Pi \times \Pi$. In terms of game play (see Sect. 4), the relation R declares which future play states π' can be anticipated when being in a play state π. This assumes $\pi \preceq \pi'$ i.e., π is an initial segment of π'.

When carrying over this approach to logical reasoning about DAVE scenarios, R specifies the expected foresight of human researchers when being engaged in analyzing, visualizing, and exploring big data.

Within this framework, reasoning about observed human behavior can be computerized, due to completeness results in modal logics Blackburn et al. (2001).

Just one interesting case shall be illustrated. For any play state π and any action μ, the formula ε^{π}_{μ} denotes that μ_{n+1} is an enforced action in the state π (see Sect. 4). This may be checked by trying to deduce $\pi \models \Box\varepsilon^{\pi}_{\mu}$. As long as this does not succeed, $\pi \models \Diamond\neg\varepsilon^{\pi}_{\mu}$ is hypothetically assumed. Then, μ_{n+1} may be considered a conscious human choice, thus, being worth an in-depth evaluation.

7 Abductive Learning as a Prerequisite for Discovery

Though being quite short, the preceding sections provide a sufficiently formal and comprehensive approach to in-depth investigations of human-computer interactions aiming at analysis, visualization, and exploration of big data toward novel insights or, at least, new hypotheses. This section is intended to sketch an application case based on some recent workshop presentation Yoshioka (2015).

As Yoshioka points out, the ability to literally see clusters in visualized data may depend on parameters of the underlying visualizations. When data records are shown in a 2D space or, possibly, in a virtual 3D space, the selection of attributes assigned to the axis is decisive. When searching for clusters, one may experiment with varying metrics (see Fig. 13 for a rough illustration).

Changing weights and stretching axis are elementary approaches to the stepwise transformation of the visual appearance of data. Those actions change the rendering. To say it the other way around, playing with renderings may lead to visual appearances of data which are easier to interpret than others.

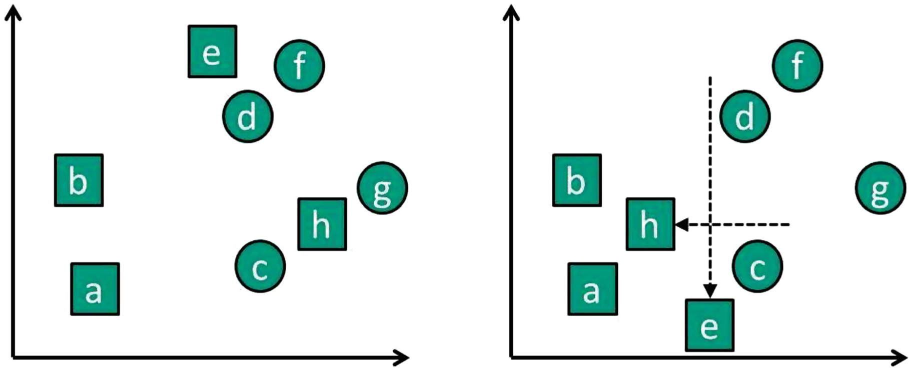


Fig. 13. Metrics variation toward intuitively perceivable clusters in visualized data; the form of the data record visualization indicates the cluster to which a record belongs

Assume there are two clusters of data records. If there a exist two disjoint convex areas in the plane such that the one contains all record visualizations of the first cluster and the other one all of the second cluster, then the clusters become visually perceivable. This property is summarized by a certain formula φ.

There are in-depth investigations into the modification of renderings by changing metrics of the 2D space Yoshioka (2015).

Assume a DAVE scenario π of the structure $\pi = \pi_1\pi_2$ with $\pi_2 \in \{r, in\}^+$, $\pi_1 \not\models \varphi$ and $\pi_1\pi_2 \models \varphi$. Sequence π_2 represents efforts to make clusters visible.

Formal language learning is a special case of exploratory clustering. It is known that formal language learning, i.e. clustering, may require the acquisition of appropriate metrics or similarity measures. π_2 represents the process–which may be computerized Sakakibara et al. (1994)–of learning those constituents.

8 Summary and Outlook

Within the present paper, the authors' contribution is focusing big data analysis, visualization, and exploration. There has been briefly introduced an appropriate technology based on which the second author has designed and implemented a demo tool, the so-called ADISY Business Intelligence Demo. All experimentation and illustrations presented within the figures of this paper have been made by means of this tool. However practically useful and illustrative, the ultimate focus of the paper is not on the ADISY Business Intelligence Demo, but on so-called DAVE scenarios, i.e. on meta-level investigations.

The authors' favored approach is formal, i.e. it relies on formal syntax and allows for processing with formal methods. Logical reasoning is of particular interest and the automation of this reasoning is an ultimate goal. Seen in this light, one may call it an Artificial Intelligence approach Grabowski et al. (1989).

8.1 The Reach of the Present Approach to DAVE Scenarios

As described above (see Sect. 4), the present approach originates from digital media research, especially from investigations of the impact of playing games. It has been demonstrated to be very useful to characterize mastery of game play Jantke (2012).

Furthermore, the approach turns out to be appropriate to the formalization of *pattern* concepts in game play Jantke and Arnold (2014). Very roughly speaking, patterns are logical formulas possibly valid in some play state. The two formulas $\forall\pi' \in \mathit{\Pi}(G)(\pi \preceq \pi' \rightarrow \pi\mu \preceq \pi')$ and $\forall\pi' \in \mathit{\Psi}(G)(\pi \preceq \pi' \rightarrow \pi\mu \preceq \pi')$ mentioned by the end of Sect. 4 are examples of patterns. In this particular case, one pattern is more general than the other one, as $\forall\pi' \in \mathit{\Pi}(G)(\pi \preceq \pi' \rightarrow \pi\mu \preceq \pi')$ implies $\forall\pi' \in \mathit{\Psi}(G)(\pi \preceq \pi' \rightarrow \pi\mu \preceq \pi')$ (but not vice versa).

The pattern concept may be easily carried over from play states to scenarios. Syntactically, this makes no difference.

When a particular scenario σ represents some human behavior in the course of an analysis, visualization, and exploration process, one may look for patterns valid in the scenario σ. Because this representation is thoroughly formalized, the search for patterns can be fully automated. Computer programs may monitor the emergence of scenarios over time–similarly to monitoring human game play–and may draw conclusions accordingly.

Patterns or instances of patterns[1] that occur in scenarios may characterize human behavior in manifold ways. Formally describable and, thus, automatically recognizable properties of strings exhibit human preferences and may reveal misconceptions and misunderstandings (see Vosniadou (2013) for valuable details).

[1] The distinction of patterns from their instances is blurred in the logical approach. In Dana Angluin's approach to patterns common to sets of strings Angluin (1980), the distinction is clear. Patterns are strings which may contain variables. In contrast, instances are ground. A string is an instance of a pattern, if it may result from a substitution of variables. The logical approach a bit more expressive. If two different formulas φ and ψ hold in some scenario and φ implies ψ, then φ is an instance of ψ.

8.2 Limitations of the Present Approach to DAVE Scenarios

There is no doubt that the present approach is having its limitations, most of them being due to its immature state of development. The present paper represents the very first publication of the authors' idea of DAVE scenarios. Subsequent papers will deal with the one or the other issue in some more depth.

Furthermore, there are some aspects which require a more comprehensive investment of scientific background, for instance, bridging the gap from knowledge about media perception and psychology to formal methods. Figure 14 below is intended to illustrate just one example. Both screenshots show three webbles extracted from the ADISY Business Intelligence Demo tool. In the lower left case, the webbles are cluttered over the screen, whereas they are well-arranged on the upper right screen. By analyzing view parameters of the corresponding webbles, this may be detected automatically. Positioning of webbles with respect to each other is another elementary action worth to be taken into account.

8.3 Outlook

Foremost, there is an obvious need to validate the present approach in practice. The authors are in close contact to a larger group of historians who are interested in investigating their own work in using big data under the perspective of DAVE scenarios, partially for a better understanding of serendipity Schubert (2013).

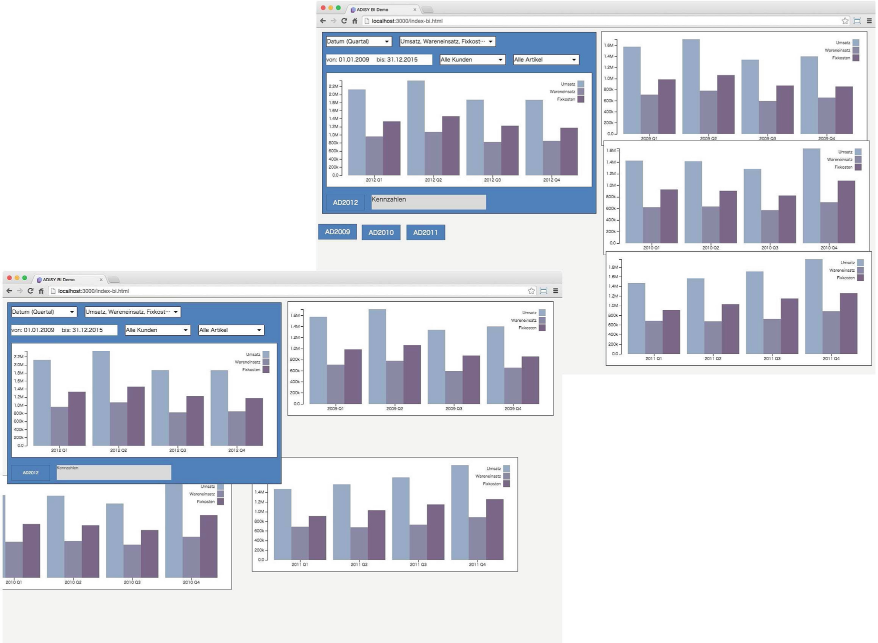


Fig. 14. Different ways of arranging extracted visualization objects on the screen

Acknowledgement. The authors are grateful to Yuzuru Tanaka for his seminal work and for his great enthusiasm to share his vision of meme media evolution of externalized knowledge. To both of them, he has provided very good working conditions on several occasions. Furthermore, he recently draw the authors' interest to the topic of the present contribution.

Part of this work has been supported by the German Federal Ministry for Education and Research (BMBF) within the joint project Webble TAG under the respective sub-project grants no. 03WKP41B (Webble ABI) and no. 03WKP41D (Webble TaT).

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

WK · POTENZIAL
UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative
Neue Länder

References

Amar, R., Eagan, J., Stasko, J.: Low-level components of analytic activity in information visualization. In: IEEE Symposium on Information Visualization, Minneapolis, MN, USA, pp. 23–25, October 2005

Angluin, D.: Finding patterns common to a set of strings. J. Comput. Syst. Sci. **21**, 46–62 (1980)

Arnold, O., Spickermann, W., Spyratos, N., Tanaka, Y. (eds.): Webble Technology. CCIS, vol. 372. Springer, Heidelberg (2013)

Blackburn, P., De Rijke, M., Venema, Y.: Modal Logic. Cambridge Texts in Theoretical Computer Science, vol. 53. Cambridge University Press, Cambridge (2011)

Fujima, J.: Building a meme media platform with a JavaScript MVC framework and HTML5. In: Arnold et al. (2013), pp. 79–89 (2013)

Gerhardt, C.I. (ed.): Die philosophischen Schriften von G. W. Leibniz, Berlin/Halle (1849)

Grabowski, J., Jantke, K.P., Thiele, H. (eds.): Grundlagen der Künstlichen Intelligenz. Akademie-Verlag, Berlin (1989)

Heer, J., Mackinlay, J.D., Stolte, C., Agrawada, M.: Graphical histories for visualization: supporting analysis, communication, and evaluation. IEEE Trans. Vis. Comput. Graph. **14**(6), 1189–1196 (2008)

Hopcroft, J.E., Motwani, R., Ullman, J.D.: Introduction to Automata Theory, Languages and Computation. Addison-Wesley, Boston (2001)

Ito, K., Igarashi, M., Takada, A.: Data mining in amino acid sequences of H3N2 influenza viruses isolated during 1968 to 2006. In: Jantke, K.P., Kreuzberger, G. (eds.) Knowledge Media Technologies, First International Core-to-Core Workshop, TU Ilmenau, Germany, Diskussionsbeiträge, vol. 21, pp. 154–158, July 2006

Ito, K., Igarashi, M., Miyazaki, Y., Murakami, T., Iida, S., Kida, H., Takada, A.: Gnarled-trunk evolutionary model of influenza A virus hemagglutinin. PLoS ONE **6**(10), 1–9 (2011)

Jankun-Kelly, T., Ma, K.-L., Gertz, M.: A model and framework for visualization exploration. IEEE Trans. Vis. Comput. Graph. **13**(2), 357–369 (2007)

Jantke, K.P.: AI planning of conflicts in non-linear spaces of time. In: IEEE Symposium on Computational Intelligence and Games, Milano, Italy, September 7–10 2009, pp. 88–95. IEEE Press (2009)

Jantke, K.P.: Patterns of game playing behavior as indicators of mastery. In: Ifenthaler, D., Eseryel, D., Ge, X. (eds.) Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives, pp. 85–103. Springer, New York (2012)

Jantke, K.P., Arnold, O.: Patterns - the key to game amusement studies. In: 3rd Global Conference on Consumer Electronics (GCCE 2014), Makuhari Messe, Tokyo, Japan, 7–10 October 2014, pp. 478–482. IEEE Consumer Electronics Society (2014)

Jantke, K.P., Fujima, J.: Toward far-reaching and effective participation in an e-society. In: Kommers, P., Isaias, P. (eds.) 13th International Conference on e-Society 2015, Madeira, Portugal, 14–16 March 2015, pp. 71–78. IADIS (2015)

Knuth, D.E.: Computer programming as an art. In: Turing Award Lectures, pp. 33–46. ACM Press, New York (1974)

Kuwahara, M.N., Tanaka, Y.: Programmable and customizable meme media objects in a knowledge federation framework einvironment on the web. In: Karabeg, D., Park, J. (eds.) Second International Workshop on Knowledge Federation, Dubrovnik, Croatia, 3–6 October 2010

Lenerz, C.: Layered Languages of Ludology - Eine Fallstudie. In: Beyer, A., Kreuzberger, G. (eds.) Digitale Spiele - Herausforderung und Chance, Game Studies, pp. 39–52. Boizenburg vwh (2009)

Popper, K.R.: Logik der Forschung. Tübingen (1934)

Sakakibara, Y., Jantke, K.P., Lange, S.: Learning languages by collecting cases and tuning parameters. In: Arikawa, S., Jantke, K. (eds.) Algorithmic Learning Theory. LNCS(LNAI), vol. 872, pp. 533–547. Springer, Heidelberg (1994)

Schubert, C.: Digital Humanities zwischen Informatik und Geisteswissenschaften? In: 20 Jahre Arbeitsgemeinschaft Geschichte und EDV. Abhandlungen der Arbeitsgemeinschaft Geschichte und EDV (AAGE), Band 2,pp. 167–186. Computus Druck Satz & Verlag, Gutenberg (2013)

Tanaka, Y.: Meme Media and Meme Market Architectures: Knowledge Media for Editing, Distributing and Managing Intellectual Resources. IEEE Press and Wiley-Interscience, New York (2003)

Tanaka, Y., Sugibuchi, T.: Component-based framework for virtual information materialization. In: Jantke, K.P., Shinohara, A. (eds.) DS 2001. LNCS (LNAI), vol. 2226, pp. 458–463. Springer, Heidelberg (2001)

Veluswamy, R.: Clinical quality data mining in acute care. In: The Physician Executive, pp. 48–53 (2008)

Vosniadou, S. (ed.): International Handbook of Research on Conceptual Change, 2nd edn. Milton Park: Routledge, New York (2013)

Wiener, N.: Mensch und Menschmaschine. Ullstein, Frankfurt am Main, Berlin (1958)

Wiener, N.: Cybernetics. The MIT Press, Cambridge (1962)

Yoshioka, M.: Interactive operation of MDS visualization results with distance metric learning. In: International Workshop on Information Search, Integration and Personalization, ISIP 2015, Grand Forks, North Dakota, USA, 1–2 October 2015 (2015, unpublished)

Zhang, D., Zhou, L.: Discovering golden nuggets: data mining in financial application. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. **34**(4), 513–522 (2004)