Towards Interactive IS Research Papers to Play With - A Process Mining Showcase

Completed Research Paper

Jörg Becker
University of Muenster – ERCIS
Leonardo Campus 3,
48149 Münster,
Germany
becker@ercis.uni-muenster.de

Philipp Bergener
University of Muenster – ERCIS
Leonardo Campus 3,
48149 Münster,
Germany
philipp.bergener@ercis.uni-muenster.de

Dominic Breuker
University of Muenster – ERCIS
Leonardo Campus 3,
48149 Münster,
Germany
dominic.breuker@ercis.uni-muenster.de

Tobias Heide
University of Muenster – ERCIS
Leonardo Campus 3,
48149 Münster,
Germany
tobias.heide@ercis.uni-muenster.de

Abstract

The paper explores how reading experience of scientific publications can be improved through interactive web content. Constructivist learning theories advertise active involvement of learners as a key requirement for successful teaching. Analogously, we argue that active involvement of readers can facilitate the process of conveying scientific research to other scholars. Today’s web technologies provide plenty of opportunities to generate interactive content. However, almost no attempts to explore these opportunities have been undertaken so far. Step by step, we discuss learning theory to demonstrate the importance of actively involving the reader. Then, we show how interactive web technology and visual media can be used to realize this involvement. By subsequently showcasing an original research result from the area of business process mining in an interactive fashion, we invite the reader to experience himself what an interactive research papers can look like. Furthermore, brief sketches of technical realization are included to encourage readers to engage in their own online publication projects.

Keywords: Learning Theories, Interactive Research Papers, Business Process Mining
Motivation

“Tell me, and I will forget. Show me, and I may remember. Involve me, and I will understand.” is a quote often ascribed to the ancient Chinese philosopher Confucius. What has been said 2500 years ago is now more relevant than ever. As compared to only passively consuming something, being actively involved is often believed to improve understanding and to facilitate long-term memorization.

This belief has led to an entire paradigm of learning theories: the constructivist learning theories (Jonassen et al. 1999). The general theme of these theories is that learning is not a process of knowledge transmission from teacher to learner. Instead, it is one in which the learner constructs knowledge himself. A teacher is merely a facilitator who establishes a suitable learning environment. As a consequence, active participation of the learner is a constituent element for learning theories belonging to this paradigm.

For instance, in their theory of andragogy (pedagogy for adults), Knowles et al. (2011) describe the self-concept of a learner as autonomous and self-directed. He actively exerts control over the process by deciding what to learn and how to learn. Kolbs experiential learning theory advocates active experimentation as an important form of learning experience (Kolb 1984). Being involved and making own experiences is an integral part of the learning process. These ideas are reflected in instructional methods as used in many classrooms today. Probably, one of the most popular examples is problem based learning, emphasizing that learners must be encouraged to take control and to develop and test their own ideas (Savery and Duffy 1995). Empirical studies found support for different forms of active learning (Prince 2004).

If active involvement is good for conveying learning matters to students, then why should it be bad for conveying research results to other scholars? In both cases, there is learning matter, there are those who are familiar with it, and those who want to know more about it. Bearing in mind this analogy, we argue for a more interactive design of research papers. Where hundreds of years ago written text was the most efficient means for dissemination of research, the interactive nature of the World Wide Web now provides entirely new, unprecedented opportunities. While the educational potential of interactive web content has long been acknowledged (Harasim et al. 1995), it is rarely seen in scientific publishing. Of course, there are research papers available online. Some may even come along with accompanying material, but the main part still is static content. It is our belief that the scientific community could benefit from a change.

Following the spirit of what we are about to advocate, we do not only present a plea in favor of online interactive research papers. Instead, we also want to let the reader experience himself what we mean with it. Moreover, we hope to encourage him to engage in activities similar to this himself. To this end, the remainder of this paper consists of two parts.

We use the first part of this paper to show what advice learning theory has to offer for the practice of scientific publishing. To that end, we first discuss learning theory in general, identify andragogy as an area particularly suitable to scientific publishing, and finally derive implications for publishing from design principles of andragogical teaching methods. In particular, we argue that web technology allows incorporating components into papers that facilitate active involvement of readers.

Then, in the second part of this paper, we present a novel research result in the way advertised in part one. What we present is a new algorithmic approach to business process mining, which is the problem of identifying the structure of business processes automatically by means of data collected from IT systems. The goal is not to lay down mathematical technicalities in their entirety. Rather, we want the reader to understand the basic architecture of the algorithm and to experience how web technology can help understanding complex topics. Our presentation does not only include a textual description of the approach, but also provides:

- (1) an animated example illustrating how petri nets work and how event-logs are created,

1 This paper is meant to be read online. We highly recommend opening up your internet browser and going to: http://prob-miner.uni-muenster.de/
Part I – Designing Interactive Research Papers

Learning

The first things many people think of when confronted with “learning” are activities such as sitting in a classroom listening to a teacher or working through a book to acquire some information. Research on learning is concerned with more general concepts though. It is a highly interdisciplinary topic and as such tackled from a multitude of different perspectives. For instance, Neuropsychology examines physiological processes in brains that explain how learning or memorization takes place. Computer scientists develop machine learning algorithms to make sense of data. Finally, psychologists investigate the abstract processes taking place in individuals while they learn (Meltzoff et al. 2009). It is this last area we want to dive into in this paper.

Psychologists ask the question how learning takes places in individuals (as opposed to e.g., organizations). The scope of their question is not restricted to classroom settings. Rather, typical definitions of learning have a much wider scope. For instance, Jarvis (2004) defines it as “a combination of processes whereby whole persons construct experiences of situations and transform them into knowledge, skills, attitudes, beliefs, values, emotions and then senses, and integrate the outcomes into their own biographies” (Jarvis, 2004, p. 111). Such a broad definition applies to a pupil in school in the same way as to a scientist at work. Illeris (2007) includes even non-human beings when he says learning is “any process that in living organisms leads to permanent capacity change and which is not solely due to biological maturation or ageing” (Illeris, 2007, p. 3). Knowles et al. (2011), after discussing numerous different definitions, put it simple by defining “learning as process of gaining knowledge and/or expertise” (Knowles, et al. 2011, p. 17), yet warn that theories of learning are too heterogeneous to actually understand the term without knowing the respective theory.

With such broad definitions of learning, it appears obvious that models of learning, though typically applied to education in school/university or to professional job training, could also explain how scholars in any scientific discipline make sense of the body of knowledge and how they develop their ideas. Consequently, research on how to create an environment that facilitates learning in those traditional settings may allow drawing analogies to a scientific learning context. But where to look for suitable theories of learning?

An important distinction made is that of pedagogy versus andragogy (Taylor and Kroth 2009). In pedagogy, one deals with learning in childhood, while in andragogy learning of adults is in the center of attention. The main aspect that sets those two areas apart is the degree to which a teacher is supposed to control the process of learning. In pedagogy, teachers are responsible for virtually any decision regarding what, how, and when to learn. The learner passively receives what he is supposed to learn. Andragogy constitutes an alternative in which the learner assumes the active role and the teachers role is changed from controller to facilitator of learning (Conner 2007).

The history of andragogy dates back to at least 1833, when this term was used to describe the teaching style of ancient Greek philosopher Plato (Knapp 1833). Although the concept was later used by several others (see Taylor and Kroth (2009) for a discussion of the early days), it gained momentum when Malcolm Knowles developed it in the 70’s into what is now called the theory of andragogy (Knowles et al. 2011). It can best be characterized as a set of six assumptions about how adults learn and, based on that, how teaching can be tailored to learners needs. His theory became a key result, inspired several
subsequent learning theories (Merriam and Caffarella 1999), and is increasingly applied to areas such as criminal justice (Birzer 2004) or medicine (Norrie and Dalby 2007). Yet it is not without its critics, mainly because empirical evidence of the degree to which an adult actually is as the theory describes is scarce (Rachal 2002).

Consequently, we spent the next section on discussing the six assumptions of andragogy in detail. In relating them to the practice of academic work, we analyze to which extent they apply to scholarly learning. We will argue that, while the assumptions may be problematic as a model of adult learning in general, they are rather natural assumptions when it comes to scholars.

**Andragogy and Scholarly Learning**

The assumptions of the andragogical model of learning (Knowles et al. 2011) have been put forth to distinguish adult from child learning. As such, they have to be read as trend statements (one end of a continuum), not as binary categories. Below, all of them are discussed:

- **Need to know:** This first assumption states that adults must be convinced of a need to know in order to be able to acquire knowledge. As a result, an adult teacher has to put considerable emphasis on motivating a topic if he wants to convey knowledge about it. One can safely assume that properly motivating one’s own research is a crucial requirement to convey knowledge about it to other scholars.

- **Learner's self-concept:** In contrast to children, adults have learned to be responsible for their own lives and the decisions they make. Being lectured by somebody else can easily create an internal conflict with this self-concept which leads to resistance. The implication is that adult teachers must be careful not to impose their own will on learners. Again, it is safe to assume scholars do not want to be treated paternalistically by others but rather want to be their own bosses.

- **Learner's experience:** As adults, learners bring a lot of background with them when they enter an educational setting, simply because they necessarily have some kind of experience. Tapping into these experiences can facilitate the learning process, but they can also be a source of presuppositions. A teacher has to be aware of these effects. Scholars working in academia typically (should) have substantial knowledge of in one or more areas related to those they want to learn about, which is why this assumption appears unproblematic as well.

- **Readiness to learn:** Adults move through different phases of their lives and develop a need to know to cope with the challenges their encounter in these phases (e.g., a pregnant woman likely develops an interest in infant nutrition). Events responsible for this are called a source of readiness to learn. The lesson is that adult teaching is more successful if such a source is present. Analogously, scholars move through different phases of their academic lives, develop interests in new topics, and will likely feel a need to know about research they perceive as related to their current interests.

- **Orientation to learning:** Adults are concerned with their current situation in life and thus more inclined to learn how to solve problems they currently encounter. They are much less interested in acquiring skills that may be of use in some hypothetical future. Consequently, learning matter should be of similar level of abstraction with respect to what a learner is interested in. If it is too abstract, the relationship may be obfuscated, resulting in resistance. With respect to scholarly work, this assumption may seem problematic in the first place. There is no particular reason why learning matter has to be problem-oriented, as one may keep learning merely for the beauty of the intellectual journey itself. However, as scholars pursue academic careers, professional necessity and personal interest may (or should) go hand in hand. Thus, there is likely some job-relatedness in a scholar’s interest in a research results, i.e., his current work is related to his readings.

- **Learner's motivation:** There are two main sources of motivation. The first is external motivation, arising from things such as better job opportunities or salary. The second is internal motivation, arising from desire for job satisfaction or better quality of life. For adults, the latter source is believed to be the much stronger one and should thus be considered in designing teaching situations. As for academic work, it is easily conceivable that internal desire to extent the boundaries of human understanding is more likely to keep scholars going than the hope of getting rich with it.

To conclude, as assumptions about how scholars learn, the andragogical model does not appear to be
extraordinary questionable, even though it is debatable as a general model adult learning. As a consequence, design principles for teaching methods that have been created with respect to the andragogical model should give insights into how “teaching between scholars” (i.e., conveying knowledge from one scholar to another) should take place.

Andragogical Methods and the Balance of Control

In line with the six assumptions of andragogy and as outlines above, andragogical methods are characterized by high degrees of learner-orientation. Of course, there is no single andragogical method. Several quite heterogeneous learner-centered teaching methods have been developed in the past. Examples are Error Training (Gully et al. 2002), Enactive Exploration (Debowski et al. 2001), and Discovery Learning (McDaniel and Schlager 1990). In a meta-review of these and other techniques, Bell and Kozlowski (2008) consolidated them into a unified framework. They identified three core design elements at the heart of all these techniques. Therefore, we use the design elements as a starting point to derive implications for an andragogical way of publishing research:

- **Exploration**: This first area is all about finding the appropriate trade-off between control exerted by teachers and control exerted by learners. Clearly, all methods shift a substantial amount of control to learners, as learning actively and self-directed is believed to increase metacognitive activity (e.g., planning and monitoring own behavior), which is in turn crucial for learning. The importance of learning through trial and error is emphasized. But at the same time, too much freedom may result in learners being out of their depth. Thus, some guidance may be beneficial, but should be used with care.

- **Training frame**: The second area is about providing conditions in which learners develop a positive orientation towards learning. The main point is that learners learn best when they strive for mastery, i.e., their goal is to develop skills. If they instead strive for compliance with performance goals, either framed positively (“be good!”) or negatively (“avoid mistakes!”), motivation will suffer. Thus, any kind of performance evaluation should not be part of an adult teaching method, or should at least not be emphasized too much.

- **Emotion control**: This last area is about keeping negative feelings about learning at bay. Learners with a lot of freedom can easily get stuck, which may lead to frustration. Others may develop anxiety and stress if they feel they perform badly. Thus, techniques of controlling negative emotions are part of andragogical teaching methods.

With the main design principles at hand, what are the implications for a scholar learning about the body of knowledge through exploring scientific literature? Starting from the end of the list, techniques of emotional control clearly cannot be part of scientific publications. While it is easily conceivable that there exist scholars who may benefit from using such techniques, it cannot be the author’s responsibility to care about the reader’s feelings. Similarly, there is no conclusion to draw from the second point. Performance evaluations are no part of scientific publications, which is exactly what is considered optimal from a theoretical point of view.

This leaves only the first point. At first sight, scientific publishing also appears to fully comply with the requirements set up in this point. In general, scholars are architects of their own fortune, and no single author of any paper would ever be in the position to enforce certain behavior through writing. Apart from guidance from e.g., a PhD supervisor or other superiors, any scholar is basically free in his choice to read or not to read a particular paper. From this perspective, it appears to be the most self-directed process imaginable.

Yet when changing the perspective a little, from the point of deciding what to read to actually reading a paper, the situation changes drastically. Once started, the reader of a paper is almost entirely at the mercy of its author. During writing, the author made the decision what to present and how to present it. This is not necessarily a problem, as theory says guidance is to some extent desirable. A good presentation can be the guide that illuminates the way to understanding.

However, a static, written presentation can also be a wall between the reader and the actual research matter behind a paper. Imagine a situation in which results of an empirical study are being discussed. A reader is guided through the data and pointed towards certain aspects of it, yet he is unable to explore any aspect not discussed by the author. He is not able to develop and test his own hypothesis. Any kind of trial
and error exploration of the data is limited to what is being presented on a few pages of paper. There is virtually no active involvement of the reader.

This shortcoming is exactly where interactive web technology can provide a benefit to the practice of scientific publishing. Computer programs have long been seen as a particularly suitable tool in learner-oriented teaching (Bell and Kozlowski 2002; Brown and Ford 2002). With publications being moved from actual paper to the digital world, “computer programs” can be part of them. There is now space behind the text sufficiently large to contain vast amounts of supplementary content and there are tools to make this content accessible. Web technology can be used to stimulate self-direct cognitive processes within readers. In short: they enable authors to provide their results in a form such that readers can explore them actively, in their own way, and in ways the author never anticipated.

**How Web Technologies Can Be Used**

With ever changing technical developments and given the heterogeneity of research results, no one can give a comprehensive overview of how web technology can be used to facilitate more self-directed, active exploration. Thus, we provide a showcase for interactive research papers in the upcoming second part of the paper. The selection of the three types of components we use in it is exemplary rather than exhaustive.

First, we include an animated visualization. Such visualizations are capable of communicating dynamic content such as the behavior of processes over time. While watching an animation, the reader is invited to follow the logic and actively deduce knowledge from it. Especially for depicting temporal changes, animations are believed to be superior to static illustrations (Bétrancourt and Tversky 2000). Yet they have to be used carefully. Empirical studies found both positive (Höffler and Leutner 2007) and negative effects (Lowe 1999; Mayer et al. 2005) of animations on learning performance. A typical caveat is excessive animation speed (Meyer et al. 2010), which is why we designed a rather slow animation.

Second, we provide an applet implementing an algorithm to analyze data. Applets are a versatile tool to actively involve the reader. He is not only able to think about an algorithmic procedure and the result it would produce when applied to data, but is instead able to immediately produce the actual result. The applet allows the reader to experiment with different parameter settings, apply his own input data, and develop and test his own hypothesis about what is going to happen.

A third component is visual media content. For instance, videos can be used to demonstrate certain activities which would otherwise be hard to explain. This way, the reader (watcher) can be provided with the information necessary to engage in the activity himself. Consequently, a video can be just that small bit of guidance required to set off trial and error behavior with other components, such as an interactive applet. Apart from that, video are believed to be capable of producing stronger cognitive activity when being watched than text when being read. For this reason, they are an important tool in active learning methods and can foster learning processes (Mayer 2009).

**Part II – The Business Process Mining Showcase**

**Introducing Business Process Mining**

Process mining is a field of research dealing with the problem of identifying the structure of business processes from instance data, collected from an organization’s IT systems (van der Aalst et al. 2004). Consider the small animated example shown in figure 1. It contains a simple petri net with four transitions A, B, C, and D. Assume this petri net represents a business process, with transitions being events in that process, such as performing a particular activity like calling a customer. The animation shows how the process progresses as the tokens pass through the model. On the bottom, an event-log builds up step by step when a transition passes a token from one place to another, showing the last 5 runs of the processes. It is easy to see that the first event is always A, then B and C follow in any order, and D is always the last event. The problem considered in process mining is to reverse engineer the process given only the event-log.
The animated figure presented above shows how an exemplary petri net works. It visualizes how the process is executed and how an event-log is created during this execution. The learning goal of this animated petri net is to communicate a basic understanding of the functionality of petri-nets and their relation to event-logs. As a petri net describes a temporal sequence of events, illustration of static content is less suitable than using an animation (Bétrancourt and Tversky 2000; Höffler and Leutner 2007). The behavior of the petri net can be visualized without lengthy formal definitions of token containment, production, and consumption semantics of the petri net modeling elements. This way, technical details unnecessary for understanding can be omitted. By watching process and event-log changing in parallel, the reader can discover himself the rule by which an event-log is generated, which triggers cognitive processes.

Technical realization: The animated figure is implemented using Scalable Vector Graphics (SVG) for representing figures combined with JavaScript (JS) to manipulate the SVG figure. In particular, we used the jQuery SVG plugin. It allows easy manipulation of geometric elements (circles, rectangles, text, lines, etc.) and also facilitates definition of animation steps, either in interaction with the reader or driven by a timer.

The reason why such event-logs are interesting is that process mining can provide a view on an organization’s business processes as they are implemented, not as they are perceived by an analyst. Traditionally, collecting information on business processes is done by conducting interviews with those involved in them. But this may produce an idealized version of the processes. Today, IT systems in organizations (e.g., Enterprise Resource Planning (ERP) or Customer Relationship Management (CRM) systems) collect vast amounts of data. Why not using them to get insights into what is actually going on? Process mining intends to discover the reality of an organization’s business processes, i.e., what they look like at the moment (van der Aalst et al. 2003).

An obvious application scenario is comparing actual with intended behavior to check if a process is performed as it should be. This is called conformance checking (Rozinat and van der Aalst 2008). Yet there are other purposes beyond merely identifying the structure of business processes as they are implemented (van der Aalst 2011). For instance, the availability of an event-log allows making statements regarding the frequency with which certain parts of the model are being visited. Other ideas are to assess capacity utilization or to simulate processes (Rozinat et al. 2009). Moreover, it can be applied in combination with techniques from areas such as social network analysis (van der Aalst et al. 2007).

The roots of process mining go back to work on discovering software processes from event data (Wen et al. 2010) that has been done by Cook and Wolf more than a decade ago (Cook and Wolf 1995, 1998). They experimented with techniques inspired by machine learning, among them neural networks and markov chains. But, unlike in our introductory example, their underlying process representation was based on finite state machines (FSM). While their idea has been transferred to a business process context,
representational models of early approaches in this area rest on a higher level of abstraction (Agrawal et al. 1998; Herbst and Karagiannis 2000). Up until today, numerous algorithms for discovering process models have been developed. They rely almost exclusively on petri nets and related process formalisms (Tiwari et al. 2008). With the process mining toolkit, a free process mining software tool is available. But also several popular vendors of business process management suites have incorporated such techniques into their products (Turner et al. 2012).

The reason why so many approaches to process mining rely on petri nets is that these are particularly suited for representing concurrency in processes. A process’ state space grows very fast with the number of concurrent events. Representing states explicitly in an FSM quickly becomes impractical (Murata 1989). Clearly, activities in business processes are often performed in parallel. With concurrency being more the rule than the exception, there is reason to abandon FSM and use petri nets instead. While petri nets constructed by early algorithms, such as the Algorithm (van der Aalst et al. 2004), belong to a restricted class, the expressiveness has been increased gradually to allow for wider classes. For instance, consider invisible activities changing the state of a petri net but not showing up in an event-log (Wen et al. 2010).

Apart from aspects regarding the expressiveness of discoverable petri nets, there were other problems to be dealt with. In particular, algorithms have to deal with outliers and noisy data in general. To accomplish this, the HeuristicsMiner has been developed (Weijters et al. 2006). But noise is not the only source of uncertainty in the data. As inductive approaches, process mining algorithms can only be fed with a finite sample of events, yet they aim to discover the ground truth from which these are generated. Avoiding overfitting to a finite sample is still perceived to be an important challenge (van der Aalst et al. 2011).

In presence of uncertainty, probabilistic models of learning are popular tools. However, as a consequence of using petri nets as the main representation of processes, such formalisms are not directly available. This could be the reason that current process mining algorithms are not rooted in probability theory and apply heuristic techniques to avoid overfitting (van der Aalst et al. 2009). It is our aim to fill this gap by providing a probabilistic model of business processes, accompanied with an algorithm to estimate them given an observed event-log. We perceive this as a first step towards a probabilistic framework of process mining which lays the ground for applying strong theory from the machine learning community.

In Quest of a Probabilistic Description of Business Processes

To develop a probabilistic description of business processes, it is natural to first look at what kind of data we observe. For process mining, the data is the event-log mentioned in the previous section. Remember that it contains sequences of events, where each sequences belongs to one particular instance of the process (a case). By grouping cases with respect to sequences and counting the number of cases in each group, we could transform the event-log into a histogram of sequences. This suggests interpreting a process as a distribution over sequences of events, where each sequence has a certain probability with which it is drawn. For the example of figure 1, this means describing the event-log in form of two sequences ABCD and ACBD, having for instance frequencies 2/5 and 3/5 respectively.

Yet the ultimate goal is not to learn such a simple distribution over sequences, but to provide insights into the structure of the underlying process generating them (e.g., by providing a petri net representation). Therefore, we need to investigate how the sequences come into being. What is needed is a probabilistic model generating sequences of events. At this point, we enter well-researched territory. Automata theory provides plenty of models for specifying how to generate words of a language, or sequences of events from a process equivalently (Hopcroft et al. 2001). Extending them with probabilities is straightforward (Rabin 1963). Inductively learning such automata (or languages) is an intensively researched problem (Carrasco and Oncina 1999).

Working with automata means being back at the level of states. Moving from one state to another along given transitions, associated with events, is how the sequence of events comes into being. Yet this level of abstraction has been abandoned long time ago for good reasons. The number of states, and with it the size of the state machine, grows quickly in case of concurrency. A petri net provides a much more appealing
visual picture by scattering state representation across its places. Fortunately, we can have both a petri net and a state machine. The theory of regions provides means to transform a state machine into a petri net describing the same process as the state machine does (Badouel and Darondeau 1998). A readily applicable implementation is available (Cortadella et al. 1997). It has been applied recently to process mining, but not in a probabilistic context (van der Aalst et al. 2009). With this tool at hand, we can safely reduce our problem to learning a probabilistic state machine.

Having defined the kind of representation we intend to use, we now provide a formalization of the probabilistic model of a process. For illustration, we provide a probabilistic graphical model (Koller and Friedman 2009) of it in figure 2. The main constituents of the model are discrete random variables, both for the (unknown) states and the (known) events. For each discrete point in time \( t \in \{1, \ldots, T\} \), two discrete random variables \( Z_t \) over states \( \{1, \ldots, K\} \) and \( X_t \) over events \( \{1, \ldots, E\} \) denote the state of the process as well as the event observed at time \( t \). Random variables are illustrated in form of circles, shaded grey if they are observed.

The next step is to define a distribution representing state machines. Basically, two ingredients are needed to accomplish this. First, we need a random variable representing the different states it might be in. It will be called \( Z \). Second, we need a random variable representing the different events that are observed. This variable will be called \( X \). As both the state and observed event change over time, variables must be replicated for each time step. This delivers chains of variables \( Z_0, Z_1, Z_2, \ldots \) and \( X_0, X_1, X_2, \ldots \). All these variables are represented in figure 2 as circles. Circles for variables \( X_0, X_1, X_2, \ldots \) are shaded grey as they are observed in the event-log, while circles for variables \( Z_0, Z_1, Z_2, \ldots \) are not, as these are latent constructs explaining used merely to explain the events observed.

![Figure 2. Proposed probabilistic graphical model of a business process](image)

Having defined the random variables, assumptions about distributions connecting them are required. They are a prerequisite to apply a learning algorithm. As we conceptualize business processes as discrete processes in time, discrete distributions are a natural choice. When a process begins, it must start off at one of its states. The discrete distribution \( P(Z_0|\pi) \) specifies for each state the probability \( \pi_k \) of it being the initial one. Depending on the current state, we define another discrete distribution \( P(X_t|Z_t, \mu) \) specifying the event being observed. Parameters \( \mu_{k,e} \) specify the probability of observing event \( e \) if the process is in
state $k$. A third distribution $P(Z_t|Z_{t-1}, X_{t-1}, A)$ determines probabilities for state updates. Each individual parameter $A_{j,g,k}$ determines the probability of moving to state $k$ when currently being in state $j$ and observing event $e$.

The resulting joint distribution over random variables $X$ and $Z$ is the product of all individual parts:

$$P(X, Z|\theta) = P(Z_0|\pi) P(X_0|Z_0, \mu) \prod_{t>0} P(X_t|Z_t, \mu) P(Z_t|Z_{t-1}, X_{t-1} - 1, A)$$

Technical Realization: To create the formula above we have applied the JS Library MathJax. It allows specifying equations either as LaTeX expressions or with the Mathematical Markup Language (MathML). Equations are rendered directly in the browser allowing for scaling an equation’s size without quality loss and providing a better reading experience as images. Furthermore, the LaTeX or MathML source code can be extracted by the user simply by right-clicking onto a formula.

Note that this parameterization requires specification of both the number of events as well as states. While it is easy to determine the number of events from the event-log, the number of states is unknown. We would rather like to estimate it than specifying it a priori. For now, we will just assume the value to be given.

Given an event-log, it is now possible to learn a process representation as described in figure 2. The learning algorithm we chose is called the Expectation Maximization (EM) algorithm (Dempster et al. 1977). Its result is a maximum likelihood estimation of the process that generated the event-log. Once values for all the distribution’s parameters have been estimated, we know, for each state, how likely it is to observe a particular event, and how likely it is to move to another state with respect to the event observed. Consequently, we can construct a state machine with probabilities on transitions. To retain only sufficiently likely behavior, those transitions having a probability smaller than a threshold value $p$ can be dropped. Subsequently, we apply the theory of regions to transform the resulting automation into a petri net, which finishes the mining procedure.

Using the Mining Algorithm

From the user perspective, the process mining tool we present works to a large extent automated. Yet in the process of using it, some decisions have to be made. To clarify these issues, we now discuss the process of doing a process mining analysis with this tool. The reader is encouraged to experiment with the tool using applet 1. Along with the process, we also provide information on how to use the applet. If the reader has opened the online version of this paper, he may also view video 1, which explains the process and how applet 1 can be used to go through it.

1. First, data has to be gathered and must be provided to the tool in form of an event-log. In a real world application, this typically means analyzing a company’s databases and collecting all relevant information.

   In applet 1, we provide three exemplary petri nets instead, from which data can be generated randomly (upper left part). Alternatively, the reader may specify his own petri net. A textual representation of the chosen petri net is displayed in the text box (upper middle part). The reader can make arbitrary changes to any petri net using this text field. To specify how many random cases will be included in the event-log, he has to set the corresponding parameter “Enter Cases:” (upper right part, second text field).

2. Having collected data, the learning algorithm can be applied to estimate a maximum likelihood state machine. It requires two user inputs. First, the reader should know that the EM algorithm is initialized with random values and will iteratively converge against a local optimum. Thus, it is recommended to perform it multiple times, specified by number of tries, and choose the best of all results. Second, the number of states (dimensionality of random variables $Z$) has to be specified before it can be run. Clearly, it cannot be determined when only the event-log is known, which is why it has to be guessed. However, there is guidance for choosing it. When choosing two different numbers of states and applying the algorithm to both of them, one can calculate a standard quality metric called the Bayesian Information Criterion (BIC) (Schwarz 1978) for each
of the results. It represents a trade-off between model size and complexity (large, high-dimensional state spaces can explain more complex processes, yet choosing a state space too large may result in overfitting). The decision rule is to always prefer the state space size that generated the lowest BIC value.

In applet 1, the number of states as well as the number of times the EM algorithm is performed can be specified using the two text fields “Enter States:” and “Enter Tries:” (upper right part, first and third text field).

3. Third, the maximum likelihood state machine produced by the EM algorithm must be pruned. It will include numerous highly unlikely transitions, but the result shall only represent the likely behavior of the process. Thus, one has to define a threshold value \( p \). Any transition with probability lower than this threshold is removed.

In applet 1, this value can be set using the text field “Enter Threshold:” (upper right part, fourth text field).

4. Finally, the pruned state machine is transformed into a petri net, which can be inspected to derive information on the underlying process which generated the event-log. No further information is required to perform this transformation.

In applet 1, this result is shown in the lower right part, beneath the petri net from which data has been generated. In addition, some statistics on running times as well as the BIC value are shown.

To summarize our approach, we first estimate a maximum likelihood state machine given an event-log, try different numbers of states, and subsequently choose an adequate model based on BIC. Then, after retaining only probable behavior, we generate a petri net to provide a human readable version of the process. The reader is invited to experiment with the algorithm provided in applet 1.

---

The tutorial video “Using the ProbabilisticMiner” demonstrates the usage of the interactive applet below. Besides providing a general description of usage, all parameters adjustable by the user are explained. Furthermore, the set of predefined

---

3 Available at: http://www.youtube.com/watch?v=UePLOnAMorA
petri nets in the tool is introduced. Video is much more effective in conveying information on how to use the applet, as explanations are provided verbally while the user can watch the author using the tool at the same time. This not only activates several senses at once but also connects explanation and location of buttons, textfields, etc. directly, thereby avoided lengthy explanations.

Technical realization: All videos are directly embedded in the website, using code supplied by the video hosting service. This allows watching the videos directly without having to follow a link.

The interactive applet of the ProbabilisticMiner is the centerpiece of this showcase. It enables the reader to execute the algorithm described in this paper. By providing example input data the reader can easily make his first steps in using the applet. If it catches his interest, he can go further and experiment with changing the setting to see how that changes the result. To provide even more degrees of freedom, he can input his own textual description of a model that is then used for data generation. By placing the model generating data right beside the mining result, success or failure of the algorithm is immediately visible.

The flexibility of this applet allows the reader to explore the algorithm's capabilities...
himself. He could directly verify any claim made in the paper. He could also ask questions not even discussed in the paper at all. By uploading his own model and specifying his own parameters, he can set up an experiment with only minimal effort, thus inviting him to get involved. Active involvement of this kind would otherwise require downloading and installing software tools in the best case or re-implementing the entire algorithm in the worst case.

Technical realization: There are several aspects making the implementation of the applet a challenge. First, the underlying algorithm was originally implemented in Java and relies on external binaries (petrify to create the Petri Net and GraphViz to render the graph). Second, the algorithm may require substantial running time depending on the input and parameters. This might exceed the client's computational resources, especially for mobile clients. Hence, parts of the algorithm have to be run on a server.

However, using modern web development toolkits, it is surprisingly easy to implement the applet. We have used the Google Web Toolkit (GWT), which has recently been advertised as particularly useful for developing scientific and educational gadgets in a web environment (Zhao et al. 2011). GWT is a framework that is able to compile Java code to JS and to establish client-server communication through web services. Thus, we programmed the entire applet in ordinary Java, compiled most of it to JS, and only extracted the algorithmic core and external binaries which are run server-side.

As for any kind of computational procedure, performance is a critical aspect. We present an initial evaluation, showing runtimes for learning the three example processes of applet 1, using 10 states and an event-log with 100 cases. Results are shown in figure 3. Any run of applet 1 extends these results (as indicated by “Run 1” in this PDF). This way, the reader can compare his own runs with the examples we already provide.

![Figure 3. Runtimes](image)

Figure 3 shows some example runtimes for the predefined input data and parameter setting. Furthermore, the chart is updated for each run of the algorithm triggered by the reader. Thus, it documents the effects of parameter changes on the runtime. This should help the reader to understand and estimate the tradeoff between accuracy of the algorithm and its runtime not only with respect to the data we provide, but also with respect to the parameter setting he chose in his own experiments. He can also compare his own results with those we provide by default.

Technical realization: The chart in Figure 3 is implemented using the D3.js library. It provides modules for displaying a variety of data-driven content in the form of typical diagrams such as bar-charts. Data from the applet is transferred using JS, which also
triggers the update. All content is rendered using HTML, CSS, and SVG.

In video 2, the reader can see several examples of what can be learned about the algorithm when interacting with it.

Video 2 contains an in-depth presentation of the how the applet could by a reader who chose to explore the capabilities of it himself. Apart from the positive effect of video explanations of tool usage discussed below video 1, it also demonstrates the potential of equipping the reader with a tool for experimentation. Within the paper, we as authors did not perform any experiments with respect to how the quality of the algorithm’s result is influenced by the parameter specifying the number of states. An interested reader can act in the way shown in the video to see how quality deteriorates with decreasing state number. Without the applet, there would be no other way to find out than implementing the algorithm. From the perspective of andragogy, it is exactly this behavior that should be stimulated.

As mentioned before, future research will focus on modifying the probabilistic model such that it accounts for the expected sparsity. This may not only speed up the algorithm, but will also provide a more rigorous theoretical way to avoid overfitting to noisy or incomplete data. Subsequently, a thorough analysis of both runtimes and result quality is the next logical step.

Conclusion, Limitations, and Outlook

Inspired by constructivist learning theories, we have argued for using web technologies to enrich scientific

\[\text{Available at: http://www.youtube.com/watch?v=AYf-cKpbSXo}\]
publications with interactive content. A discussion of learning theory and its implications on the practice of scientific publishing has been presented in part one. The key finding was web technologies can provide content in a more cognitively stimulating way, which is more suitable for scholarly learning. In particular, interactive web applets can enable the reader to explore research results beyond what the author believed he could be interested in.

Original research about business process mining was then presented in a second part. It was supplemented by interactive and visual components to illustrate how the principles discussed in the first part can manifest. An explanation how the components facilitate readers learning process has been provided along with each component. With providing this showcase, the reader is not forced to use his imagination but can actually experience it himself. It is our conviction that interactivity improves reading experience. Given the fairly substantial mathematical background required to follow a formal derivation of the algorithm, it could be laborious to analyze our process mining approach to an extent necessary to grasp a feeling for the algorithm’s capabilities. Yet by enabling a reader to try it out while reading the paper, we allow him to quickly explore himself what it does. Instead of diving into pages of formulas, he could set up his own ad-hoc experiments. Given the high degree of flexibility, his experiments could easily go beyond anything we had presented in a full evaluation.

At selected points in part two, we provided brief descriptions of the skills and tools necessary to create an interactive web publication. Thereby, we hope to reduce the adoption barrier and to encourage readers to work on their own interactive web publication projects. Although basic knowledge in web development is already sufficient to generate convincing results, we do not believe widespread adoption could be achieved without further simplifications, as barriers are too high for non-technical scholars. Besides programming, there is also the problem of hosting the online version of the paper. It is thus an open question how to provide an easy-to-use toolbox that is applicable without technical expertise and still takes account of the diversity of research. Similar to modern content management systems, it is conceivable to provide a framework with graphical user interface that allows insertion of text, formulas, figures, and different kinds of standard animations and applets. How such a framework could look like is left to future research.

With this work we do not attempt to provide a comprehensive picture of the opportunities arising from web technologies applied to scientific publishing. How suitable animations, videos, or applets are for research results other than the algorithm of our showcase cannot be said. Clearly, the appropriateness of using such components has to be reevaluated for each case specifically. Andragogical learning theory can provide guidance in choosing appropriate components, but an exhaustive discussion of such guidelines is beyond the scope of this paper.

To conclude, it is our firm conviction that scientific publishing can benefit enormously from web technologies and that interactive content will be a decisive factor in increasing research productivity in the future. Exploration of technological opportunities and their potential impact is just about to get underway.

References


Kolb, D. A. 1984. Experiential learning: Experience as the source of learning and development, Journal of