DOI: 10.20535/1810-0546.2018.5.146178
UDC 004.42+519.85+004.9

A.Ye. Batyuk, V.V. Voityshyn*
Lviv Polytechnic National University, Lviv, Ukraine

PROCESS MINING: APPLIED DISCIPLINE AND SOFTWARE IMPLEMENTATIONS

Background. A precise picture of how business processes (in the interpretation by Andrea Burattin) are performed in real-life is vitally important for an organization because it shows actual situation revealing gaps and bottlenecks. Process mining is a discipline with the purpose to research processes using as the input so-called event data (or event logs) which in essence is a digital footprint left in IT systems as the result of business processes execution.
Objective. The goal of the study is to overview current state of process mining and find actual scientific and practice tasks in this field as well as justify and formalize requirements to the information technologies with the purpose to implement the found set of process mining applied tasks.
Methods. The method used by the authors to prepare current overview consisted of the following steps: (a) analysis of information sources; (b) finding and formalization of actual scientific as well as practice tasks; (c) description of the requirements to the information technologies with the purpose to implement the found set of actual tasks.
Results. It has been found out that process mining as an applied discipline has been actively developed for 20 years; significant contribution to creating the scientific basis of process mining has been done in Eindhoven University of Technology (The Netherlands) under direction of professor Wil M.P. van der Aalst. It also has been found actual scientific and practice tasks of process mining: event data preparation, dealing with concept drift, operational support, event data streams processing, handling big event data, improving process mining tools usability for the end users. It has been formalized requirements and specified quality attributes for the information technologies with the purpose to implement the found actual tasks. Architecture of the information technologies has been proposed by the authors.
Conclusions. Currently the theoretical core of process mining has mainly been developed and quite structured. However, despite of the fact that mathematical methods and software tools have been successfully used in practice for a few years, the request for the intellectual business process analysis has not been fulfilled yet. The authors have found out that relevant information technologies should supply such functions as handling big event logs, dealing with event data streams as well as operational support of business processes which are at the execution stage.

Keywords: process mining; information technology; business process management; BPM; event data; event logs; XES, ProM; Disco; Celonis.

Introduction

Events that happen in the surrounding world are not by themselves but belong to various kind of processes. For instance, when clients order products in an e-commerce system they usually perform sequences of related actions before submitting orders or leaving a web site without buying anything. From business standpoint it is crucial to know: what are the actual users' journeys in a system; what are the weaknesses and bottlenecks of the flow; and finally, how to evolve the service so that it becomes more comfortable and attractive for the end users. The discipline which deals with that kind of tasks is named "Process Mining". Process mining can be considered as an independent branch of data science which applies data mining techniques to datasets generated by processes executions (so-called event data or event logs). In fact, this discipline acts as a link between such widely known fields as data mining, business intelligence (BI), and business process management (BPM).

Companies usually use documentation and conduct manual investigations to get overview of their business processes (in the context of current study the "business process" term is understood in the interpretation by Andrea Burattin [1, pp. 11–13]). These traditional approaches can provide information that something does not work appropriately; however, the obtained information is not precise enough since it does not show all the facts that reveal what exactly does not work well and under what circumstances. Another disadvantage is that the manual analysis is time and resource consuming which in the agile world means that the results of analysis can quickly lose their value due to constant changes of the business environment. This is where process mining helps. The purpose of the process mining toolset is to provide valuable insights about actual situation using the digital footprints left by business processes in IT systems. With a few clicks the end user can have a look at how business processes are executed and drill down to identify the root cause of a problem analyzing

* corresponding author: voytyshyn@gmail.com

patterns in the underlying event data. In case of process mining application to key business processes of an organization the obtained insights can be used not only to improve the monitored processes but are valuable inputs for digital transformation of the entire company.

Despite of the fact that process mining is a relatively young market niche there are already a few mature its software implementations. One of the oldest is the ProM framework which is popular within the audience of scientific researches and process mining professionals. Disco is another process mining tool distributed under a commercial license. This tool is much simpler in comparison with ProM and targets users who are experts in a business domain but not necessarily have processes mining academic background. Celonis is a fast-growing German startup. This product is a powerful dashboard-based solution with visualization and analytics features which allow to take a look at the researched processes from different perspectives. The target audience of Celonis is medium and large enterprises. Of course, the list of process mining software is not limited by these three the most famous products and includes tools for data scientists (for example, the bupaR extension for R and RapidProM for RapidMiner) as well as commercial solutions (like Minit, myInvenio, QPR Process Analyzer, ARIS Process Performance Manager, Fujitsu Automated Process Discovery Service etc.).

The method used by the authors to prepare current study consisted of the following steps: (a) analysis of information sources; (b) finding and formalization of actual scientific as well as practice tasks; (c) description of the requirements to the information technologies with the purpose to implement the found set of actual tasks. The analyzed information sources included various kind of publications (books, academic articles, scientific conference proceedings as well as posts in technical blogs, reports, articles in periodicals) and official technical documentation provided by vendors of process mining tools. Significant part of the knowledge was received by testing of process mining software; of course, if a free/trial version was available (for example, ProM, RapidProM, Disco) or a vendor provided an academic program for researches (for example, Celonis). In case of impossibility to get direct access to a process mining product public technical documentation and web resources were the only way to obtain necessary information.

The paper is divided into two main sections. The first section represents an introduction into process mining as an applied discipline, outlines characteristics of datasets used in this area, and emphasizes found by the authors actual tasks for process mining. The second section is devoted to process mining software. It includes an overview of previously done studies on the process mining tools examination, description of the three notable software products, requirements and architecture for the information technologies with the purpose to implement found by authors actual process mining tasks.

Problem Statement

The goal of the study is to overview current state of process mining and find actual scientific and practice tasks in this field as well as justify and formalize requirements to the information technologies with the purpose to implement the found set of process mining applied tasks.

Process Mining as an Applied Discipline

What is Process Mining?

Process mining is a discipline considered as a bridge that connects a few research and technological fields such as data mining, machine learning, Business Intelligence (BI), and Business Process Management (BPM) [2, p. 30]. Process mining deals with time series data generated by processes execution. That kind of datasets is usually named "event logs" or "event data". The characteristic that differentiates process mining from the other related disciplines is the focus on the "processes" nature of analyzed data. From the BPM standpoint process mining can be seen as an intelligent instrument that allows to deal with real-life processes which do not necessarily have predefined control-flow models. In turns, for data science professionals process mining techniques are useful approaches to visualize and discover time series data.

Some preliminary works, which formed the theoretical basis of the modern process mining techniques, were carried out since 1950's [3–7]. Some of the very first papers devoted to the problem of process model discovery from event logs were published on 1990's [8–10]. Process mining as a research discipline known today has been developed in the Eindhoven University of Technology (The Netherlands) under direction of professor Wil van der Aalst, who has made significant

contribution into creating theoretical core of process mining and promoting its industrial applications. In particular, Wil van der Aalst and Ton Weijters are the authors of process mining research agenda [11] that has defined the direction of scientific efforts in this area since the middle for 2000's. Nowadays IEEE CIS Task Force on Process Mining is one of the widest communities that unities many organizations and professionals in this field around the world.

Process Mining Tasks

Process Mining Manifesto [12] defines the following three major categories of tasks: (a) discovery, (b) conformance checking, and (c) enhancement. The purpose of the tasks from the 1st category is to build model of a real-life process using event data. In most cases the generated process model is expressed with a visual notation like Petri Net, BPMN, UML Activity Diagram etc. It should be noted that "Automated business process discovery" (ABPD) [13] is another term that covers the tasks from the 1st category. The conformance checking tasks find the difference between the real-life process, which is generated from event data, and the "ideal" predefined model specified, for example, as a BPMN diagram. In turns, enhancement techniques compare event logs with the predefined process model and provide suggestions how to improve the process itself and its model.

Further level of details for the three categories of processes mining tasks are provided in [14] introducing a set of 19 use cases validated by expert interviews and survey (the initial purpose of the set of the use cases was to evaluate process mining tools available on the market that time). The identified use cases are grouped according to the three classical categories of tasks listed above. Outcomes of the interviews and survey shown that the most important use case was "Structure of the process" that belongs to the discovery task category. The enhancement group (which is the most complex from algorithmic perspective and has huge potential from the business application perspective) contains 12 use cases which is significantly greater in comparison with the discovery (4 use cases) and conformance checking (3 use cases). It should also be noted that the enhancement group includes use cases related to so-called social mining tasks which focus on the people involved in a process (for example, building a map of relationships among process participants and identifying central employees in the collaboration).

The Gartner's report [15] defines five business use cases that are used for segmentation of the process mining market. The goal of the 1st use case, which is called "Improving business processes by algorithmic process discovery and analysis", is to automatically create maintainable and validated process models applying set of special algorithms to real-life events. First of all, the generated models visualize discovered processes how they look like in reality and also provide insights related inefficiencies and bottlenecks. Such kind of models are a powerful instrument of documenting "as is" state which is a base line for further processes optimizations and redesigning. The next use case is named "Improving auditing and compliance by algorithmic process comparison, analysis and validation". The main task of this use case is to identify the difference between a real-life process model obtained by applying process discovery algorithms (use case 1 from the Gartner's report) and predefined instructions. Another variation of current use case can be comparison of how the same business process is executed in different departments of an organization. The name of the 3rd use case is "Improving process automation by discovering and validating automation opportunities". It is about finding opportunities for automation. In contradiction to traditional process analysis techniques mostly based on interviewing stakeholders, the process mining approaches allow to receive much more accurate picture how operations are performed in real-life and as the result help to identify automation of what of the performed tasks are going to bring the biggest value. As it can be seen the first three use cases from the Gartner's report correlate with the three categories of tasks defined in Process Mining Manifesto [12]. The 4th use case (named "Support digital transformation by linking strategy to operations") is a little bit different from the traditional ones. It states that insights obtained by means of process mining techniques can be used by organizations' strategic initiatives related to the trend called "digital transformation". The 5th use case ("Improving IT operations resource optimization by algorithmic IT process discovery and analysis") is also a relatively new one. It connects process mining to the usual tasks performed by IT organizations (e.g. software development, testing, maintenance, support etc.). For example, in this case process mining can bring benefits in supporting an organization's IT department providing instruments to analyze logs (usually stored in text files) generated by various kind of applications starting, for instance, from an enterprise resource

planning (ERP) system to a middleware like an enterprise service bus (ESB).

As a scientific discipline process mining deals with the three categories of tasks defined in [12]. The process discovery task is the most popular one with well worked out algorithmic basis. Practical applications of process mining techniques are usually started from that kind of tasks. However, according the Gartner's report [15] adoption of the techniques from the second and third categories is supposed to be increased in the next two years whilst usage of discovery algorithms will be a little bit decreased. In the authors' opinion, this forecast highlights the fact that process mining as a scientific discipline has a strong enough theoretical base which can be applied to more complex (in comparison with discovery) practice tasks.

Data Formats

Event data (or event logs) is a special type of data sources which process mining techniques require. Guiding principle 1 declared in Process Mining Manifesto [12] states: "event data should be treated as first-class citizens". Such kind of dataset is a collection of cases each of them is a sequence of events. A single event has to include the following attributes: (a) case (identifier of a process instance), (b) activity (what is done), (c) timestamp (the time when the event happens). More precise definitions of "event log", "case" and "event" are provided in [2, pp. 128–137]. In some cases, the "timestamp" attribute is represented by two fields defining start and finish time which make possible more precise measurement of timing metrics. In practice apart from the fields listed above an event usually contains other parameters, for example: information about the executor, product identifiers, related documents etc. In particular, having attributes with information about people (so-called the originator or resource attribute) involved into process execution makes possible to deal with social mining tasks that represent relationships among the individuals, teams, departments involved into processes execution.

Modern software systems generate a lot of event logs. In the most common scenario logs are written to text files so that support engineers analyze them in order to troubleshoot runtime errors. In practice to automate analysis of event data stored in text files log aggregators are used. Architecture concept of such kind of systems is represented in [16]. The other way to store event data is using either SQL or NoSQL databases. And finally, the most advanced approach (which often takes place in the Big Data and Internet of Things scenarios) is to read event logs from data streams in near real-time mode.

Like classical data science approaches process mining techniques require data preparation as an initial step. Additionally, in most cases it is necessary to convert a data set into a special format understandable for a particular process mining software. The simplest and most commonly used data format is CSV (comma separated value). The only requirements to a data set stored in CSV is that it has to contain at least the three columns mentioned above (case, activity, and timestamp). MXML (Mining eXtensible Markup Language) is an XML-based format created in 2003 [17]. The purpose of the format was to standardize event logs data structure. MXML used to be primary file format for the initial versions of the ProM software package. However, after some time it turned out that MXML was not flexible enough because it puts restrictions on what information can be contained in an event log. XES (eXtensible Event Stream) became a successor of MXML. It was officially defined as IEEE Standard in November 2016 [18]. XES is an XML-based standard with the purpose to allow software systems to exchange event data using a commonly understandable format. The XES format is supported by most process mining software, in particular ProM supported it starting from version 6. The reference Java-based implementation of XES is the OpenXES library distributed under the GNU Lesser GPL (LGPL) license [19]. There are also relatively new implementations of OpenXES on other programming languages: OpenXes.Net (.NET), OpyenXes (Python). Additionally, it should be noted that there is an alternative implementation of XES standard called XESLite with the purpose to deal with large event logs consuming less resources in comparison with OpenXES.

Challenging Tasks of Process Mining

One of the earliest reviews of the process mining challenging tasks was published in 2004 within the scope of the process mining research agenda [11]. The stated challenges were mostly focused on algorithmic perspective of process discovery (e.g. "mining hidden tasks", "mining duplicate tasks", "mining loops" etc.). The list also included challenges which have not lost their relevance even today (like "mining from different perspectives" and "gathering data from heterogeneous sources").

Once process mining methods were applied to real-life event data the researches faced with the problem that real processes usually had unstructured nature and contained quite big amount of details (so-called "spaghetti" processes). That circumstance made the models generated by early discovery techniques like the α-algorithm [20] hardly readable (or even useless) for the end users. One of the algorithms with the purpose to cope with such a problem was published in 2007 under the name "Fuzzy Miner" [21]. That technique introduced the concept of "road-map" when a process flow was represented on a certain level of abstraction hiding less important details. A little bit later Fuzzy Miner was successfully adopted by Disco [22].

Next review of challenging tasks was represented in Process Mining Manifesto [12] published in 2011. In contradiction to the research agenda [11] the challenges declared in the mani-

***Table* 1.** Process mining challenges: current state and actual tasks

#	Process mining challenge	Current state	Actual tasks
1	Event data preparation	– standardization of event data formats [17–19]; – maturity levels for event logs [12]; – a quantitative model to measure quality of event logs [27]	– event logs preparation for data streams mining (obtaining, cleaning, integration, selection, transformation, and definition of quality metrics)
2	Concept drift of business process model	– a method to detect changes in event data and identify the regions of change in a process [28]; – the online Heuristic Miner for evolving event data streams [23]; – a method to localize sudden concept drift based on applying the statistical hypothesis testing method [29]; – an online technique to deal with concept drift of a process model [30]	– handling noise in event data so that it is not interpreted as changes of a process model; – discovering types of concept drift (sudden, gradual, recurring, incremental); – identification of the points where a process model becomes different
3	Operational support	– a process mining framework that allows to deal with business processes at the execution stage (so-called "pre-mortem" event data) [2]; – completion time prediction for executing process instances [31]	– taking into consideration recurring patterns (e.g. weekly, monthly, annually etc.) and contextual factors of a business process when making predictions and recommendations
4	Event data streams	– modifications of the Heuristic Miner with the purpose to deal with event data streams [23]; – a generic method that allows to adapt the existing process discovery algorithms to event data streams [32]	– distributed computations to process streaming data; – conformance checking for event data streams; – handling concept drift of a business process model
5	Big event data	– process discovery and conformance checking frameworks with purpose to deal with large event data sets [26]; – implementation of the α -algorithm and Flexible Heuristic Miner using Map-Reduce [33]	– partitioning of event data sets and distributed computations; – adaptation of process mining techniques to the Big Data patterns (lambda [34] and kappa [35] architectures)
6	Improving usability and understandability for the end users	– the road-map concept for discovered process models [21]	– increasing interpretability of the results obtained by applying process mining methods (an overview of the existing techniques is provided in [36])

festo were more high-level and focused rather on practical aspects than theoretical basis. The first task stated by the manifesto was "finding, merging, and cleaning event data" which is very close to the "gathering data from heterogenous sources" challenge defined by the research agenda. In the authors' opinion this means that the task of collecting and preparing event logs has not been completely resolved yet and even has become more important and difficult nowadays.

It also should be emphasized that the manifesto introduced such new problems as "dealing with concept drift" (when a process model changes over time) and "providing operational support" (which aims to deal with incomplete process instances). These two tasks are related to one of the "hottest" modern challenges of process mining – handling event data streams (so-called "streaming process discovery" or SPD) [1]. One of the earliest stream-aware discovery algorithms was a modification of Heuristic Miner [23]. It should be noted that this algorithm has been used by the authors in the real-time business process monitoring information technology [24, 25].

Another ongoing challenge has been triggered by the Big Data trend. In the process mining field scaling of event data is considered in two different perspectives: (a) number of events and (b) number of activities [26]. The first case is rather relevant to size of an event data set, whilst the second one is about complexity of a process model.

The described above scientific and practice challenges of process mining are summarized in Table 1. It is worth to note that the summary does not pretend to be a comprehensive overview. Its purpose is to outline the most relevant (in the author's opinion) applied tasks of process mining.

Taken together, it can be summarized that there is a strong correlation between found actual process mining tasks and the Big Data trend. Consequently, the current requirements to process mining methods are the ability to deal with big event logs and infinite event data streams. This, in particular, means that it is necessary to handle running business process instances which is closely related to the operational support task.

Process Mining Software

Publications on Process Mining Software Examination

Examining of process mining software implementations has been done by different authors during last a few years. Each work differs from the others by an applied research methodology and a list of analyzed tools. As process mining is a rapidly growing field it is worth to pay attention to early overviews and also have a look at recent publications.

A comparative analysis of process mining tools available in the market as of 2011 is provided in [37]. The analysis is based on a process mining use cases framework [38]. The framework consists of 19 use cases validated by interviewing of process mining practitioners. The use cases are grouped by the three categories named with accordance to the classical process mining tasks: (a) discovery (4 use cases); (b) conformance checking (3 use cases); (c) enhancement (12 use cases).

In [39] it is represented an overview of business process mining software (not tools for scientists) available in the United Kingdom market as of 2012. The research represents comparison of products against the criteria: process mining task (discovery, conformance checking, enhancement), process model (BPMN, Petri Net, etc.), process mining problem (mining loops, dealing with noise, etc.), algorithm. Additionally, the study provides a concise but quite informative overview of existing at that time process mining techniques.

The research from [40] is based on an online survey executed from Nov 2013 till Feb 2014. The goal of the survey was to identify criteria of choosing process mining tools by the end users. As the result usability was chosen as the most important criterion. It is remarkable that such characteristics as integration capabilities and scalability (ability to handle constantly increasing volume of event data) were selected as important ones too.

An evaluation framework with the purpose to assist users in selecting the right tool is proposed in [41]. The framework maps business problems (e.g. inspecting and cleaning event data, understanding a business process model etc.) to process mining operations (e.g. filtering event data, process discovery, social network mining, decision rules mining etc.). The problems were linked to the operations by analyzing a set of case studies. As an example, ProM, Disco, and Celonis were evaluated by means of the proposed framework.

The classification introduced in [2] is based on the questions: (a) "How often is the same analysis repeated?" and (b) "Can the end user freely determine the analysis to be conducted?". According to these questions the three types of use cases are defined. Type 1 requires full flexibility to perform a certain analysis just one time. Type 2 in-

cludes questions asked repeatedly but relatively infrequently; the analysis approach is predefined but not completely fixed. And type 3 covers routine questions; narrow range of customizations is possible. As stated in [2], most ProM plugins belong to type 1 whilst RapidProM [42] aims type 2. In turns, tools of type 3 provide predefined dashboards that target users who are not process mining professionals. Moreover, in [2] it is provided an extended overview of such well-known products as ProM, Disco, Celonis and also briefly described more than 10 other free and commercial tools.

One of the most recent reviews of process mining software market has been done by Gartner [15]. The report provides a list of 15 representative vendors with descriptions of their products (or services). It is remarkable that this is the first case when process mining is identified by Gartner as a dedicated market niche.

The listed above publications highlight that during less than 10 years process mining has grown from a scientific discipline to a promising and rapidly developed market area.

Notable Process Mining Products

ProM. ProM is a mature process mining framework and today it is a benchmark in the scientific world. The primary contributor of ProM is Eindhoven University of Technology (The Netherlands). From the very beginning ProM has been an open source Java-based pluggable platform. The initial version of ProM was released in 2004, it included 29 plugins only [43]. ProM 5.2 was released in 2009 with 286 plugins. Whereas the versions from 1.1 to 5.2 were based on the same architecture ProM 6, which was released in 2010, was completely redesigned [44]. In 2018 ProM 6 was named by Gartner a leading process mining research platform [15].

In contradiction to the earlier versions ProM 6 introduced the concept of contexts which ensures clear separation of algorithms from GUI and also makes possible distributed calculations (i.e. a mining process is executed on a few instances of ProM located in different machines). Another significant architecture change made in ProM 6 was explicit specification of inputs and outcomes for a plugin. This has created the conditions for macro plug-ins development. However, this makes impossible of using plug-ins from the earlier versions of ProM on the new platform. Another powerful improvement introduced by ProM 6 is dynamic plug-ins installation by means of the ProM Package Manager. This tool allows to add/remove packages (each package is a collection of plugins). Consequently, ProM is installed with core functionality and the packages necessary for the end user can be added later. Due to such kind of flexibility ProM Core is distributed under GNU Public License (GPL) whereas plugins typically use Lesser GNU Public License (L-GPL).

Starting from version 6 ProM supports XES event data format [18]. However, the MXML data format (a predecessor of XES) is supported as well. ProM 6 includes a special tool called XESame with purpose to extract event data and convert it to XES format (e.g. CSV file can be easily converted to XES by means of XESame without coding). In 13 Oct 2017 ProM 6.7 was certified by XES Working Group (XES WG) against the A, B, C, D, and X levels for importing and exporting.

At the time of writing current paper the latest version of ProM is 6.8. This version includes about 1500 plug-ins. Simultaneously with ProM 6.8 ProM Lite 1.2 is available. The difference between these software packages is that ProM Lite provides the most commonly used features so that it is much easier to use than the full version. Another important difference is that in contradiction to ProM Lite the results received by means of the full version can be officially referred in scientific publications.

ProM is a mature and extremely powerful framework with wide variety of eatures. Detailed description GUI and functionality of this software product is out of scope of current paper. Only some of the most useful functions of ProM are briefly described below. ProM GUI has 3 different views: (a) workspace, (b) actions, and (c) view. The workspace deals with all resources either imported (e.g. event log files) or obtained by executing actions of other resources. The actions (or plug-ins) applicable to the resources are accessible through the actions view and grouped by the categories: (a) discovery, (b) conformance checking, (c) enhancement, (d) filtering, and (e) analytics. The 3rd view is accountable for representing outcomes of action executions. Once an event log is imported from the workspace it can be discovered by switching to the view. The Log Visualizer feature provides informative general overview of an event log including number of processes, cases and events, frequency of events, resources summary, detailed list view of the events. So-called helicopter view on an event log is supplied by another visualization called Dotted Chart: the events are repre-

sented on the 2-dimencional coordinate system with timestamp on one axis and case on the other.

As mentioned above one the basic process mining tasks is process discovery. ProM provides wide variety of plug-ins to deal with this task. The Alpha Miner plug-in is among the most widely used. The algorithms from the α-family take even log and build a Petri net. The plug-in implements the classical α-algorithm [20] and its extensions: α+ [45], α++ [46], α# [47]. The α+-algorithm deals with special cases of loops (short- and self-loops). More complex patterns such as non-free choice constructs can be discovered with the α++-algorithm. The α#-algorithm accounts for discovering so-called invisible tasks which are not observable but affect the process flow. Fuzzy Miner [21] is another useful plugin with the purpose to discover real-life processes. That kind of processes are usually executed in the environment with low level of constraints (unlike processes defined with BPMN model) so that the α-algorithms extract hardly readable models (so-called "spaghetti" processes). The Fuzzy Miner plug-in deals with this problem combining similar activities into clusters and hiding less significant activities. In contradiction to the α-algorithms, that build Petri nets, Fuzzy Miner generates so-called fuzzy models. Consequently, the models generated by the Fuzzy Miner plug-in emphasize the most important aspects of the real-life process so that the diagrams are much better readable for the end users. Due to its ability to build well readable models for real-life processes Fuzzy Miner was adopted by Disco miner [22].

Additionally, ProM provides a platform for plug-in developers. The ProM source code SVN repository is publicly available in read-only mode. To get commit rights and register own packages a developer should contact the administrator. It is recommended to use Eclipse IDE and Java 1.6. The Framework project should be referenced by developer's plugins since it contains the ProM core. The GettingStarted package can be used as a template for a new custom package.

It is appropriate to mention here another part of ProM ecosystem – RapidProM which is an extension to RapidMiner. RapidMiner is a unified data science platform that provides a complex solution for the enterprise level data science task. Primary software products of the platform are: (a) RapidMiner Studio, (b) RapidMiner Server, (c) RapidMiner Radoop. Additionally, extensions can be installed through the RapidMiner marketplace. It should be emphasized that the platform provides solutions for Big Data proposing machine learning on Hadoop and Spark. The mentioned above the three parts of RapidMiner products family are distributed under community and enterprise licenses. In 2018 Gartner named RapidMiner a leader for ability to execute and completeness of vision. RapidProM ensures combination of wide variety of process mining techniques supported by ProM with the data science workflows supplied by the RapidMiner platform [42]. RapidProM development was stated in 2014 in Eindhoven University of Technology. As both RapidMiner and ProM are Java-based software, porting ProM plug-ins is mainly the task of integration with the RapidMiner environment. In particular, ProM plug-ins are mapped to RapidMiner operators and plug-in visualizers are transformed to renderers. RapidProm 4.0.001 supports 58 operators including discovery (Alpha Miner, Heuristic Miner, Inductive Miner, Social Network Miner, Fuzzy Miner etc.), conformance checking, analysis algorithms, model conversion (e.g. Petri Net to BPMN etc.). It is worth to notice that event streams procession operators are supported as well, for example: Stream Alpha Miner, Stream Inductive Miner etc.

The most significant disadvantage of ProM is its complexity caused from one hand by wide variety of plug-ins and slightly unusual for desktop applications style of GUI from the other. This fact is not critical for an experienced ProM user; however, can make some difficulties for a newcomer. The good thing is that researchers can officially cite the results obtained by means of the full version of ProM in their scientific publications.

D i s c o. Disco is a commercial software developed by the Fluxicon company [22]. Fluxicon was started in 2009 by process mining researchers. The 1st software launched by Fluxicon was the Nitro product with the purpose to get event logs of real-life process and transform them into format convenient for process mining. Support of the Nitro tool was stopped as long as Disco provides the similar features. At the moment of writing current paper the latest version of Disco is 2.2.1 released in 28 Aug 2018. Apart from commercial version there is an academic license supported within Fluxicon Academic Initiative for Process Mining Research and Education. Disco is a Java-based software and supports the MacOS and Windows platforms. In 2018 Gartner named Disco the most popular stand-alone process mining tool in the market [15].

Disco allows importing event logs in classical process mining formats (CSV, MXML, XES), XLS and XLSX are supported as well. Furthermore,

Disco defines its native event data format called FXL which is more efficient on large data sets than the classical formats. Disco miner is a process mining algorithm used by this software product. Disco miner is an adaptation of Fuzzy Miner [21]. Owing to Fuzzy Miner's ability to generate "user friendly" process models the models mined by Disco are well-understood to the domain experts who has had no prior experience in process mining.

GUI of Disco has three primary views: (a) map, (b) statistics, (c) cases. The map view provides visual representation of the process control flow including visualization of frequency and performance metrics. The map view filtering is based on frequency of activities and paths, timeframe, performance, endpoints etc. Informative general overview of an event log is supplied by the statistics view. The cases view gives a convenient representation of event log raw data. Cases on this view are grouped by so-called variants (a variant is one "run" through the process from its start to end). As most cases of an event log usually follow only a few different variants, it is important to know what the most frequent variants are.

According to Gartner [15] Fluxicon focuses more (than its competitors) on customer interactions, social and organizational mining, data preparation and cleansing. In contradiction to ProM, that mostly suites researchers and process mining professionals, Disco targets business domain experts making possible to apply process mining techniques to real-life processes providing simple and convenient graphic user interface.

C e l o n i s. Celonis is positioned as a commercial technology with the purpose to visualize and analyze real-life processes in a company using digital footprints which every process leaves in IT systems. Celonis targets customers in various business domains (e.g. banking, logistics, telecommunication etc.) and covers wide range of business process types. In 2018 Celonis was named by Gartner [15] a process mining market leader.

Celonis was founded in 2011 in Munich. A year later the company participated in the German Silicon Valley Accelerator, open an international office in US and became a member of the SAP HANA Startup Focus Program. In 2015 SAP started promoting and reselling Celonis under the name "SAP Process Mining by Celonis" [48] around the world. In the same year Celonis won Deloitte Fast50 Award as the fastest growing tech company in Germany (this was achieved thanks to revenue growth of 4000% over 4 years). Next year two of the largest venture capitalists in the world Accel and 83North invested $27.5 million in Celonis (it was the 1st round of funding after starting the company).

Starting from December 2016 Celonis PI (Proactive Insights) Engine [49] is available for the end users. The PI Engine combines process mining with machine learning to achieve intelligent process analysis. Celonis PI consists of 4 parts: (a) conformance, (b) machine learning, (c) social, (d) companion. The conformance feature compares the actual process built using event logs with the predefined documented, detect weaknesses and provides intelligent suggestions for fixing the found issues. The second part is about integration of machine learning algorithms into Celonis, in particular, the R language scripting and R libraries are fully supported and available for the end users. The social part allows to discover how employees, teams, organizations interact with one another during execution of a process so that from one hand key players and top performers are identified and from the other hand inefficiencies in the organizational structure can be uncovered. The fourth part of Celonis PI integrates with IT systems (e.g. SAP Business Client) providing intelligent analytics while processes are executed. Celonis can work with various range of event data sources including relational databases and software solutions of other vendors (e.g. there are available connectors to SAP, SalesForce, Microsoft Dynamics NAV, Microsoft Dynamics AX etc.). In 2017 Celonis (version 4.2) passed XES-certification for the A1, B1, C1, D1 levels on data importing.

Celonis is distributed under two categories of licenses: (a) academic, (b) commercial. The academic license provides 180 days free of charge access to Celoins Ademic Cloud for researchers. There are commercial licenses for a single user, enterprise on-premises installation, and enterprise SaaS (System as a Service). Additionally, training materials are provided by Celonis Training Cloud.

Being a successful startup Celonis has demonstrated a good example of how a scientific concept such as process mining can be turned into a rapidly developed business that has managed to grow up to over 250 employees and gain more than 350 clients around the world for so short period of time.

Requirements to the Information Technologies

As stated above, the ability to deal with big event logs and infinite event data streams as well as providing operational support are the "hottest"

challenges of process mining. Current section is devoted to the information technologies with the purpose to implement the found process mining tasks (Table 1). Architecture significant requirements to the information technologies are provided in terms of quality attributes [50]. Then the architecture, that matches the specified quality attributes, is proposed.

Quality Attributes. Quality attributes (subset from [50]) of the information technologies are listed in Table 2. The specified requirements describe the most important properties which have significant influence on the architecture. Mapping of the quality attributes to the found by the authors actual tasks of process mining is also provided in Table 2 (the "Pr. min. tasks" column of Table 2 refers to the "#" column of Table 1).

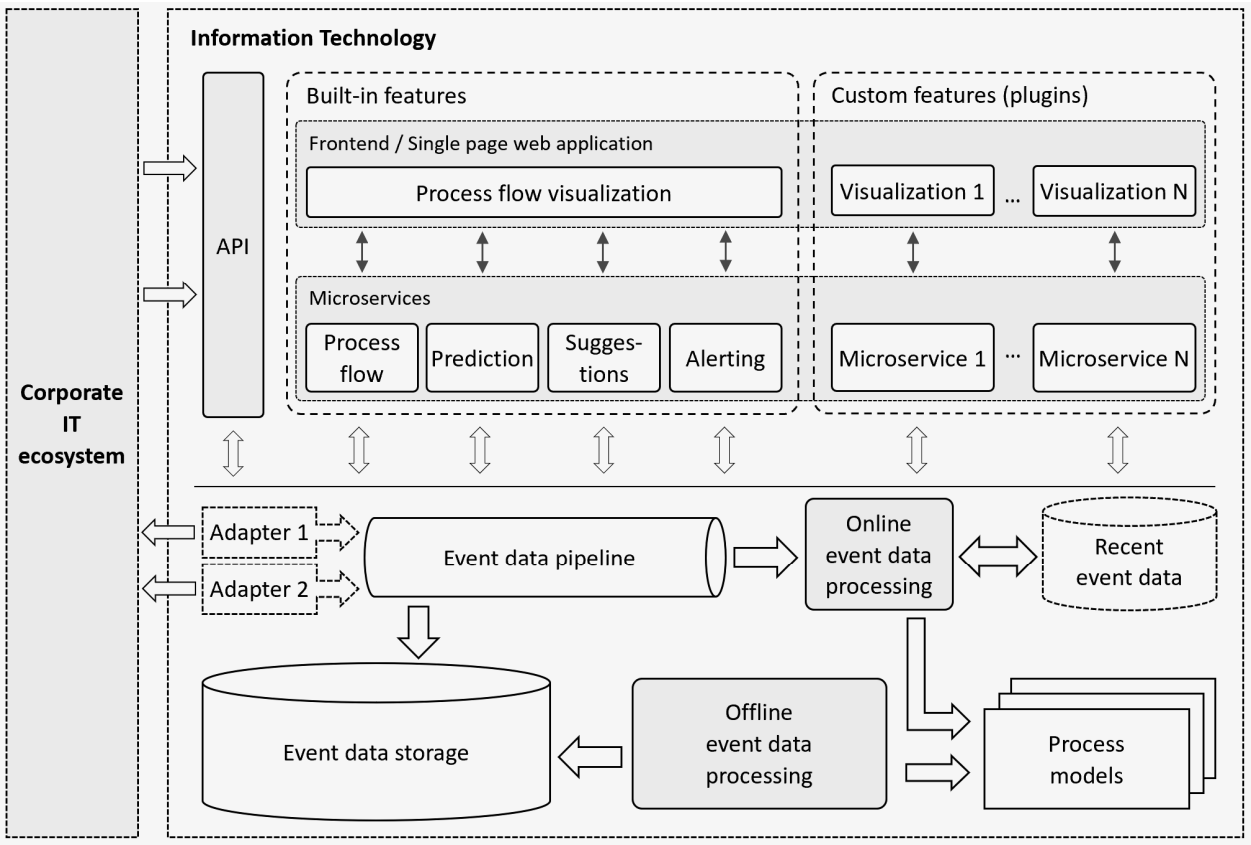
It should be noted that qualities from Table 2 covers so-called non-functional requirements (which is enough to design the architecture) whilst functional requirements to the information technologies are currently out of scope.

Architecture. The proposed architecture is depicted on the Figure. From the high-level standpoint the system consists of three parts: (a) presentation layer, (b) integration API, (c) data processing layer. The presentation layer is a single page web application (SPA) with microservices on the server side. The integration API is exposed so that other software tools can consume it (e.g. to read process models generated by the system). The data processing layer is a core of the system. It is based on the lambda architecture [34] pattern. One the advantages of such a pattern is that it ensures stream (online) and batch (offline) processing of event data.

The represented architecture satisfies the specified quality attributes (Table 2) and can be extended/customized according to the requirements of a

***Table* 2.** Quality attributes of the information technologies

#	Quality attribute	Sub characteristic	Pr. min. tasks	Requirements description
1	Compatibility	Interoperability	1	The system should be integrated with corporate software. Two types of integrations to be ensured: (a) input, (b) output. The input integrations are responsible for obtaining event data from various kind of sources and converting it into XES format. In turns, the purpose of the output integrations is to provide an API (i.e. application programming interface) that can be consumed by other software tools
2	Maintainability	Modifiability	–	Any component of the system can be replaced with the similar one. This allows to avoid a so-called "vendor lock" issue
3		Modularity	–	The system should be an extensible platform which supply only basic functionality. New features can be added a in plug-in manner
4		Testability	–	Implementations of event data processing algorithms should be covered by automated tests
5	Performance efficiency	Capacity	4, 5	It is necessary to variate capacity (scale) of the system depending on amount of processing data. If amount of data is relatively small the system functions in the minimal configuration. But it should be possible to scale the system (without architecture changes) so that it is able to handle big event data
6		Time behavior	3, 4	Event data streams should be processed in real-time mode. In the context of current paper "real-time" means the period during which the response of the system is relevant (in some cases it is seconds but in other it can be even minutes)
7	Reliability	Fault tolerance	3, 4	The system should not contain any single point of failure
8	Usability	Accessibility	6	The graphic user interface should be easily accessible for the end users (e.g. as a web application)
9		Learnability		The system should be easy to use for the end users who are not experts in process mining
10		Operability		



Architecture of the information technologies

concrete practice tasks. It also should be noted that the proposed architecture targets information systems for medium and large organizations.

Conclusions

Process mining as a scientific discipline originated in late 1990's in Eindhoven University of Technology (The Netherlands). Professor Wil van der Aalst has been one of the contributors under whose direction process mining has become as such as it is today. And now process mining has a strong community that unites researchers and developers all over the world. The described above notable software products represent three the most important target groups of the end users: ProM for researches, Disco for small and medium size organizations, and Celonis for enterprises. Despite of the fact that theoretical basis of process mining has already been created, applied researches are being carried out dynamically reflecting advanced industry trends.

The authors have concluded that nowadays the most challenging process mining tasks are the following: (a) event data preparation, (b) dealing with concept drift, (c) operational support, (d) event data streams processing, (e) handling big event data, (f) improving process mining tools usability for the end users. As can be seen, researches in directions (d) and (e) are driven by the practical necessity to deal with constantly increasing amount of event data (which is compliant with so-called "5Vs" of Big Data [51]). The authors have found out that relevant information technologies should supply such functions as handling big event logs, dealing with event data streams as well as operational support of business processes which are at the execution stage. It has been formalized requirements and specified quality attributes for the information technologies with the purpose to implement the found actual tasks. The architecture proposed by the authors is based on the lambda architecture pattern [34] which provides strong technical basis for implementation of the listed above challenging tasks.

In the authors' opinion process mining is an excellent example of how a theoretical concept can be transformed into successful industrial applications which, in turns, have grown up to a self-sufficient market niche. In 2018 Gartner [15] de-

fined process mining as a separate market area linking it with the most recent technological trends such as big data, digital transformation (DX), internet of things (IoT), artificial intelligence (AE), robotic process automation (RPA) etc. As of 2017 Gartner's estimate of process mining market for new product licenses and maintenance revenue was about 120 million USD (the estimate does not include following consulting and service revenue). Gartner's experts expect that this market to grow 3 or event 4 times in the next two years. Bastian Nominacher (co-founder and co-CEO of Celonis) is even more optimistic. He estimated the process mining market in 2017 at about 150 million EUR and forecasted its growing to 15 billion EUR by 2025 [52].

One of the future steps of the study can be to customize the proposed architecture for the needs of the information systems which targets event logs generated by enterprise middleware (e.g. an enterprise service bus). Industry 4.0 has also prepared the background for further process mining applications. One of the most promising areas is robotic process automation (RPA). According to Gartner [15] process mining can be used to identify automation opportunities which is the initial step to robotics.

Acknowledgements

We would like to thank Celonis for provided access to their academic program which made possible for the authors to try out one of the most advanced process mining software products currently available in the market.

References

[1] A. Burattin, *Process Mining Techniques in Business Environments*. Cham, Switzerland: Springer, 2015. doi: 10.1007/978-3-319-17482-2

[2] W.M.P. van der Aalst, *Process Mining: Data Science in Action*, 2nd ed. Berlin, Germany: Springer, 2016. doi: 10.1007/978-3-662-49851-4

[3] A. Nerode, "Linear automaton transformations", *Proc. Am. Math. Soc.*, vol. 9, no. 4, pp. 541–544, Aug. 1958. doi: 10.2307/2033204

[4] E.M. Gold, "Language identification in the limit", *Information and Control*, vol. 10, no. 5, pp. 447–474, May 1967. doi: 10.1016/S0019-9958(67)91165-5

[5] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", *IEEE Trans. Inform. Theory*, vol. 13, no. 2, pp. 260–269, April 1967. doi: 10.1109/TIT.1967.1054010

[6] L.E. Baum *et al.*, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, Feb. 1970. doi: 10.1214/aoms/1177697196

[7] A.W. Biermann and J.A. Feldman, "On the synthesis of finite-state machines from samples of their behavior", *IEEE Trans. Comp.*, vol. C-21, no. 6, pp. 592–597, June 1972. doi: 10.1109/TC.1972.5009015

[8] J.E. Cook and A.L. Wolf, "Discovering models of software processes from event-based data; CU-CS-819-96", University of Colorado, Department of Computer Science, Boulder, CO, USA, Nov 1996.

[9] R. Agrawal *et al.*, "Mining Process Models from Workflow Logs", in *Advances in Database Technology – EDBT'98 (EDBT 1998). Lecture Notes in Computer Science*, vol. 1377, H.J. Schek *et al.*, eds. Berlin, Heidelberg, Germany: Springer, 1998, pp. 467–483. doi: 10.1007/BFb0101003

[10] A. Datta, "Automating the discovery of AS-IS business process models: Probabilistic and algorithmic approaches", *Inform. Syst. Res.*, vol. 9, no. 3, pp. 275–301, Sep. 1998. doi: 10.1287/isre.9.3.275

[11] W.M.P. van der Aalst and A.J.M.M. Weijters, "Process mining: A research agenda", *Computers in Industry*, vol. 53, no. 3, pp. 231–244, June 2004. doi: 10.1016/j.compind.2003.10.001

[12] W.M.P. van der Aalst *et al.*, "Process mining manifesto", in *Business Process Management Workshops. BPM 2011 International Workshops. Lecture Notes in Business Information Processing*, vol. 99, F. Daniel *et al.*, eds. Berlin, Heidelberg, Germany: Springer, 2012, pp. 169–194. doi: 10.1007/978-3-642-28108-2_19

[13] *Gartner IT Glossary: Automated Business Process Discovery (ABPD)* [Online]. Available: https://www.gartner.com/it-glossary/automated-business-process-discovery-abpd

[14] I. Ailenei *et al.*, "Definition and validation of process mining use cases", in *Business Process Management Workshops. BPM 2011 International Workshops. Lecture Notes in Business Information Processing*, vol. 99, F. Daniel *et al.*, eds. Berlin, Heidelberg, Germany: Springer, 2012, pp. 75–86. doi: 10.1007/978-3-642-28108-2_7

[15] M. Kerremans. (2018). *Market Guide for Process Mining* [Online]. Available: https://www.gartner.com/doc/3870291/market-guide-process-mining

[16] A. Batyuk and V. Voityshyn, "Business processes monitoring by means of real-time visual dashboards", in *Proc. 6th Int. Academic Conf. Information, Communication, Society 2017 (ICS 2017)*, Lviv, Ukraine, 2017, pp. 204–205.

[17] B.F. van Dongen and W.M.P. van der Aalst, "A meta model for process mining data", in *Proc. Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability, Co-located with CAiSE'05 Conference*, Porto, Portugal, 13–14 June 2005. Available: http://ceur-ws.org/Vol-160/paper11.pdf

[18] *IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams*, IEEE Standard 1849-2016, 2016.

[19] *OpenXES* [Online]. Available: http://www.xes-standard.org/openxes/start

[20] W.M.P. van der Aalst *et al.*, "Workflow mining: discovering process models from event logs", *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1128–1142, 2004. doi: 10.1109/TKDE.2004.47

[21] Ch.W. Günther and W.M.P. van der Aalst, "Fuzzy mining – Adaptive process simplification based on multi-perspective metrics", in *Proceedings of the 5th International Conference on Business Process Management. BPM 2007. Lecture Notes in Computer Science*, vol. 4714, G. Alonso *et al.*, eds. Berlin, Heidelberg, Germany: Springer, 2007, pp. 328–343. doi: 10.1007/978-3-540-75183-0_24

[22] Ch.W. Günther and A. Rozinat, "Disco: discover your processes", in *Proc. Demonstration Track of the 10th International Conference on Business Process Management (BPM 2012)*, Tallinn, Estonia, 2012, vol. 940, pp. 40–44.

[23] A. Burattin *et al.* (2012). *Heuristics Miners for Streaming Event Data* [Online]. Available: https://arxiv.org/abs/1212.6383

[24] A. Batyuk *et al.*, "Software architecture design of the real-time processes monitoring platform", in *Proc. 2018 IEEE 2nd Int. Conf. Data Stream Mining & Processing (DSMP'2018)*, Lviv, Ukraine, 2018, pp. 98–101. doi: 10.1109/DSMP.2018.8478589

[25] A. Batyuk and V. Voityshyn, "Streaming process discovery for lambda architecture-based process monitoring platform", in *2018 IEEE 13th Int. Sci. Tech. Conf. Computer Science and Information Technologies (CSIT'2018)*, Lviv, Ukraine, 2018, pp. 298–301.

[26] S.J.J. Leemans *et al.*, "Scalable process discovery and conformance checking", *Software & Systems Modeling*, vol. 17, no. 2, pp. 599–631, 2018. doi: 10.1007/s10270-016-0545-x

[27] M.O. Kherbouche *et al.*, "Towards a better assessment of event logs quality", in *Proc. 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, Athens, 2016, pp. 1–8. doi: 10.1109/SSCI.2016.7849946

[28] R.P. Jagadeesh Chandra Bose *et al.*, "Handling concept drift in process mining", in *Advanced Information Systems Engineering. CAiSE 2011. Lecture Notes in Computer Science*, vol. 6741. London, UK, 2011, pp. 391–405. doi: 10.1007/978-3-642-21640-4_30

[29] M.V.M. Kuma *et al.*, "Capturing the sudden concept drift in process mining", in *BPM Workshops*, vol. 1371, pp. 132–143, 2015.

[30] J. Carmona and R. Gavaldà, "Online techniques for dealing with concept drift in process mining", in *Advances in Intelligent Data Analysis XI. IDA 2012. Lecture Notes in Computer Science*, vol. 7619, J. Hollmén *et al.*, eds. Berlin, Heidelberg, Germany: Springer, 2012, pp. 90–102. doi: 10.1007/978-3-642-34156-4_10

[31] W.M.P. van der Aalst *et al.*, "Time prediction based on process mining", *Inform. Systems*, vol. 36, no. 2, pp. 450–475, 2011. doi: 10.1016/j.is.2010.09.001

[32] S.J. van Zelst *et al.*, "Event stream-based process discovery using abstract representations", *Knowl. Inform. Syst.*, vol. 54, no. 2, pp. 407–535, 2018. doi: 10.1007/s10115-017-1060-2

[33] J. Evermann, "Scalable process discovery using map-reduce", *IEEE Trans. Services Comp.*, vol. 9, no. 3, pp. 469–481, 2016. doi: 10.1109/TSC.2014.2367525

[34] *Lambda Architecture* [Online]. Available: http://lambda-architecture.net

[35] *Kappa Architecture* [Online]. Available: http://milinda.pathirage.org/kappa-architecture.com

[36] M. Du *et al.* (2018). *Techniques for Interpretable Machine Learning* [Online]. Available: https://arxiv.org/abs/1808.00033

[37] I.M. Ailenei, "Process mining tools: A comparative analysis", M.S. thesis, Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands. Available: http://alexandria.tue.nl/extra1/afstversl/wsk-i/ailenei2011.pdf

[38] I.M. Ailenei *et al.*, "Towards an evaluation framework for process mining systems", BPM Center Report BPM-11-13, BPMcenter.org, 2011. Available: http://bpmcenter.org/wp-content/uploads/reports/2011/BPM-11-13.pdf

[39] C.J. Turner *et al.*, "Business process mining: From theory to practice", *Business Process Management J.*, vol. 18, no. 3, pp. 493–512, June 2012. doi: 10.1108/14637151211232669

[40] D. Verstraete, "Process mining in practice: Comparative study of process mining software", M.S. thesis, Faculty of Economics and Business Administration, Ghent University, Ghent, Belgium. Available: https://lib.ugent.be/fulltxt/RUG01/002/165/042/RUG01-002165042_2014_0001_AC.pdf

[41] M. Kebede, "Comparative evaluation of process mining tools", M.S. thesis, Faculty of Mathematics and Computer Science, Institute of Computer Science, University of Tartu, Tartu, Estonia. Available: https://comserv.cs.ut.ee/home/files/gizaw_MSc.+in+Software+Engineering_2015.pdf?study=ATILoputoo&reference=BB4063305540E49644F08DD06F6C50F5D0266630

[42] W.M.P. van der Aalst *et al.* (2017). *RapidProM: mine your processes and not just your data* [Online]. Available: https://arxiv.org/abs/1703.03740

[43] B.F. van Dongen *et al.*, "The ProM framework: A new era in process mining tool support", in *Applications and Theory of Petri Nets 2005. ICATPN 2005. Lecture Notes in Computer Science*, vol. 3536, G. Ciardo and P. Darondeau, eds. Berlin, Heidelberg, Germany: Springer, 2005, pp. 444–454. doi: 10.1007/11494744_25

[44] H.M.W. Verbeek *et al.*, "ProM 6: the process mining toolkit", in *Proc. Business Process Management 2010 Demonstration Track*, vol. 615, M. La Rosa, Ed. CEUR-WS.org, 2010, pp. 34–39.

[45] A.K.A. de Medeiros *et al.*, "Process mining for ubiquitous mobile systems: an overview and a concrete algorithm", in *Ubiquitous Mobile Information and Collaboration Systems. UMICS 2004. Lecture Notes in Computer Science*, vol. 3272, L. Baresi *et al.*, eds. Berlin, Heidelberg, Germany: Springer, 2004, pp. 151–165. doi: 10.1007/978-3-540-30188-2_12

[46] L. Wen *et al.*, "Mining process models with non-free-choice", *Data Mining & Knowledge Discovery*, vol. 15, no. 2, pp. 145–180, 2007. doi: 10.1007/s10618-007-0065-y

[47] L. Wen *et al.*, "Mining process models with prime invisible", *Data & Knowledge Eng.*, vol. 69, no. 10, pp. 999–1021, 2010. doi: 10.1016/j.datak.2010.06.001

[48] *Showcase: SAP Process Mining by Celonis* [Online]. Available: https://www.sap.com/developer/showcases/process-mining-by-celonis.html

[49] F. Veit *et al.*, "The proactive insights engine: process mining meets machine learning and artificial intelligence", in *15th Int. Conf. Business Process Management (BPM'2017). BPM Demo Track and BPM Dissertation Award*, vol. 1920, Barcelona, Spain, 2017.

[50] *System and Software Quality Models*, ISO/IEC 25010, 2011.

[51] W. Fan and A. Bifet, "Mining big data: Current status, and forecast to the future", *SIGKDD Explorations*, vol. 14, no. 2, pp. 1–5, 2012. doi: 10.1145/2481244.2481246

[52] P. McGee. (2017). *New Big Data Trend Tracks 'Digital Footprints'* [Online]. Available: https://www.ft.com/content/402553f4-c4a4-11e7-b30e-a7c1c7c13aab

А.Є. Батюк, В.В. Войтишин

ПРОЦЕС-МАЙНИНГ: ПРИКЛАДНА ДИСЦИПЛІНА ТА ПРОГРАМНІ РЕАЛІЗАЦІЇ

Проблематика. Розуміння того, як бізнес-процеси (в сенсі тлумачення, даного Andrea Burattin) виконуються на практиці, життєво необхідне для сучасної організації, оскільки це показує реальну ситуацію, виявляючи недоліки та "вузькі" місця. Процес-майнинг (англ. process mining), або інтелектуальний аналіз процесів, – це дисципліна, яка займається дослідженням бізнес-процесів, використовуючи як вхідні дані записи із журналів подій (англ. event logs або event data), які за своєю суттю є цифровим відбитком (англ. digital footprint), залишеним в IT-системах як результат виконання бізнес-процесів.

Мета дослідження. Метою дослідження є огляд поточного стану процес-майнингу та визначення актуальних наукових і практичних задач у цій галузі, а також обґрунтування і формалізація вимог до інформаційних технологій, що могли б реалізувати окреслене коло прикладних задач процес-майнингу.

Методика реалізації. Методика, застосована авторами для підготовки цього огляду, складалася з таких кроків: (а) аналіз інформаційних джерел; (б) визначення та формалізація актуальних наукових і практичних задач; (в) опис вимог до інформаційних технологій, метою яких є реалізація окреслених актуальних задач.

Результати дослідження. Встановлено, що процес-майнинг як прикладна дисципліна активно розвивається впродовж останніх 20 років; значний вклад у розробку наукової бази процес-майнингу було зроблено у Технічному університеті Ейндховена (англ. Eindhoven Technical University) під керівництвом професора Wil M.P. van der Aalst. Визначено основні актуальні наукові та практичні задачі процес-майнингу: підготовка даних (англ. event data preparation); опрацювання зміни моделі бізнес-процесу (англ. concept drift); надання операційної підтримки для бізнес-процесів (англ. operational support); обробка потоків даних (англ. event data streams); опрацювання великих обсягів даних (англ. big event data); представлення моделей бізнес-процесів із різних точок зору (англ. mining from different perspectives); покращення зручності (англ. usability) програмних засобів процес-майнингу для кінцевих користувачів. Формалізовано вимоги та описано атрибути якості інформаційних технологій, які розв'язують ці актуальні задачі, та запропоновано їх архітектуру.

Висновки. На сьогодні теоретичне ядро процес-майнингу в основному вже сформоване і достатньо структуроване. Однак незважаючи на те, що математичне та програмне забезпечення процес-майнингу успішно використовується на практиці упродовж останніх кількох років, потреба в інтелектуальному аналізі бізнес-процесів реалізована ще далеко не повністю. Авторами встановлено, що актуальною є розробка інформаційних технологій, основними функціями, яких є обробка потоків даних і забезпечення операційної підтримки бізнес-процесів, що перебувають на стадії виконання.

Ключові слова: процес-майнинг; інформаційна технологія; інтелектуальний аналіз процесів; бізнес-процес менеджмент; XES; ProM; Disco; Celonis.

А.Е. Батюк, В.В. Войтишин

ПРОЦЕСС-МАЙНИНГ: ПРИКЛАДНАЯ ДИСЦИПЛИНА И ПРОГРАММНЫЕ РЕАЛИЗАЦИИ

Проблематика. Понимание того, как бизнес-процессы (в смысле толкования, данного Andrea Burattin) выполняются на практике, жизненно необходимо для современной организации, так как это отображает реальную ситуацию, демонстрируя недостатки и “слабые” места. Процесс-майнинг (англ. process mining), или интеллектуальный анализ процессов, – это дисциплина, которая занимается исследованием бизнес-процессов, используя как входные данные записи из журналов событий (англ. event logs или event data), которые по своей сути представляют цифровой отпечаток (англ. digital footprint), оставленный в ИТ-системах как результат выполнения бизнес-процессов.

Цель исследования. Целью исследования является обзор текущего состояния процесс-майнинга, определение актуальных научных и практических задач в этой отрасли, а также обоснование и формализация требований к информационным технологиям, которые могли бы реализовать очерченный круг прикладных задач процесс-майнинга.

Методика реализации. Методика, примененная авторами для подготовки этого обзора, состояла из таких шагов: (а) анализ информационных источников; (б) определение и формализация актуальных научных и практических задач; (в) описание требований к информационным технологиям, целью которых является реализация очерченных актуальных задач.

Результаты исследования. Установлено, что процесс-майнинг как прикладная дисциплина активно развивается в течении последних 20 лет; значительный вклад в разработку научной базы процесс-майнинга был сделан в Техническом университете Эйндховена (англ. Eindhoven Technical University) под руководством профессора Wil M.P. van der Aalst. Определены основные актуальные задачи процесс-майнинга: подготовка данных (англ. event data preparation); обработка изменения модели бизнес-процесса (англ. concept drift); предоставление операционной поддержки для бизнес-процессов (англ. operational support); анализ потоков данных (англ. event data streams); обработка больших объемов данных (англ. big event data); представление моделей бизнес-процессов с разных точек зрения (англ. mining from different perspectives); улучшение удобства (англ. usability) программных инструментов процесс-майнинга для конечных пользователей. Формализованы требования и описаны атрибуты качества информационных технологий, которые решают эти актуальные задачи, а также предложена их архитектура.

Выводы. На сегодня теоретическое ядро процесс-майнинга в основном уже сформировано и достаточно структурировано. Тем не менее невзирая на то, что математическое и программное обеспечение процесс-майнинга успешно используется на практике в течении последних нескольких лет, потребность в интеллектуальном анализе бизнес-процессов реализована еще далеко не полностью. Авторами установлено, что актуальной является разработка информационных технологий, основными функциями которых являются обработка потоков данных и предоставление операционной поддержки для бизнес-процессов, пребывающих на стадии выполнения.

Ключевые слова: процесс-майнинг; информационная технология; интеллектуальный анализ процессов; бизнес-процесс менеджмент; XES; ProM; Disco; Celonis.

Рекомендована Радою
факультету прикладної математики
КПІ ім. Ігоря Сікорського

Надійшла до редакції
28 серпня 2018 року

Прийнята до публікації
6 вересня 2018 року