

Toward understanding and forecasting the research trends in process mining: An analysis using text mining and machine learning

Gyunam Park^a, Jiyeon Lee^b, Minsu Cho^{b,*}

^aProcess and Data Science Group (PADS), RWTH Aachen University, Ahornstraße 55, Aachen, 52074, North Rhine-Westphalia, Germany

^bSchool of Information Convergence, Kwangwoon University, Kwangwoon-ro 20, Seoul, 01897, South Korea

Abstract

Process mining is a relatively new discipline focusing on gaining process-centric knowledge from event logs collected by enterprise systems. Various techniques have been developed to automatically discover process models, analyze the compliance of real-life processes to the process models, predict operational frictions, recommend possible actions to mitigate emerging risks, etc. In recent years, the techniques have been adopted in various industries, such as healthcare, manufacturing, logistics, and finance. Now there are over 40 commercial products based on process mining techniques, such as Celonis, SAP Signavio, Minit (Microsoft), MyInvenio (IBM), and open-source tools such as ProM, PM4Py, and bupaR. In this work, we analyze the process mining literature in depth using text mining and machine learning techniques. More in detail, we 1) analyze the main research topics in process mining and their trends, 2) investigate the relationship between the topics, and 3) predict the expansion of the topics in the near future. To that end, we analyzed 2,207 process mining articles from 2002 to 2021 using a range of techniques such as topic modeling, exploratory data analysis, network and community detection, and future prediction.

Keywords: process mining, text mining, network analysis, literature reviews, emerging trends, predictions

1. Introduction

The development of SMACIT (i.e., Social, Mobile, Analytics, Cloud, and Internet of Things) and other digital technologies have enabled the growth of new organizational values with the aim of digital transformation [1, 2, 3]. In particular, there have been many efforts to identify and improve operational processes based on IT-enabled technologies, and the increase in the availability of data has expedited it [4, 5]. Process mining has emerged in such an environment and played a crucial role in implementing digital transformation [5, 6].

Process mining is a relatively young discipline focused on deriving process-oriented knowledge from event logs collected by enterprise systems [7]. The discipline leverages process science, a family of methods for improving operational processes, including business process management and simulation, and data science deriving value from data, such as machine learning and artificial intelligence [7].

*Corresponding author

Email addresses: gnpark@pads.rwth-aachen.de (Gyunam Park), dbad10923@naver.com (Jiyeon Lee), mcho@kw.ac.kr (Minsu Cho)

Under the circumstances, there have been many approaches to three research branches: (i) technical, (ii) application, and (iii) supportive aspects. Regarding the technical viewpoint, many efforts have been devoted to developing process discovery techniques [8] that automatically produce process models from event logs without prior knowledge and presenting innovative works on conformance checking [9], process monitoring [10], and predictions [11]. From the view of the application, i.e., the business perspective, this discipline has provided availability to be applied in multiple domains, e.g., healthcare, education, services, and manufacturing, and the ability to create a new implication [12, 13, 6]. Also, from the supportive aspect, process mining has been utilized to enhance efficiency and improve the accuracy of other disciplines, such as simulation, business process management, and machine learning [14, 7, 15].

Such interests in process mining and researchers' efforts have guided the advancement and expansion of this research field. A process mining research community, i.e., the IEEE task force on process mining, has emerged, and much research has been published in the independent international conference on process mining series. This well-marked progress requires an in-depth understanding of the research scopes, specific research topics, fashionable techniques, relationships of research topics, and future trends of process mining. Therefore, it requires a comprehensive review to understand the discipline in essence.

There have been several comprehensive reviews on process mining [8, 16, 17, 18, 19, 20, 21, 22]. Corallo et al. [21] employ a systematic mapping method to describe research matters, techniques, and applications in process mining. Zerbino et al. [22] reconcile process mining research based on the multiple viewpoints of contributions, algorithms, perspectives, sectors, business, and management. In addition, there is research work to detail a specific topic, including process discovery [8], conformance checking [16], event abstraction [17], and concept drift [18], while others have handled the applications for a specific context, such as healthcare [19] and education [20].

However, the existing studies employ a qualitative review by a group of experts centered on the authors. Although knowledge-based reviews by experts have produced mature and worthwhile results, they have some disadvantages in that it is based on subjective judgments and discussions. As a result, bias or preconceived notions may arise in recognizing research topics and identifying meaningful trends, and there may be limitations in predicting the potential for future technology development. As a means of overcoming these limitations, this study attempts to analyze the process mining literature in depth using text mining and machine learning techniques. That is, we propose a quantitative literature review method. To the best of our knowledge, this study is the first case of discovering new values and knowledge by directly utilizing process mining-related literature as data.

In this paper, we prepared the following research questions to clarify the purpose and direction of data analysis.

- RQ1. What are the main research topics of process mining, and how has the trend changed over time?
- RQ2. What process mining techniques are mutually cooperative, and how does each research topic relate to the other?
- RQ3. How much will process mining evolve and expand in the near future, and what research topics will be the focus?

To resolve these research questions, we collected 2,207 process mining articles and conference proceedings from 2002 to 2021. Such documents have been explored with the pre-trained model-based topic modeling, exploratory data analysis with key features, network and community detection, and future prediction for research trends with a deep learning model. This research has the following contributions:

- More straightforward results, e.g., research topics and trends, are presented through quantitative analysis, such as text mining and machine learning.
- Rather than simply enumerating the process mining review results, we describe in-depth analysis outcomes, such as discovering relationships between research topics and identifying research themes.
- Finally, we suggest a method for identifying the potential of process mining and each specific research topic through a time-series forecasting model based on historical data.

This paper is organized as follows. A series of methods for deriving process mining research topics and meaningful bibliometric results are provided in Section 2. Section 3 presents topic modeling results and describes the details of process mining research fields. Then, Section 4 investigates the discovered research fields with the common aspects of publications. Section 5 explains the interrelationships between research topics, and Section 6 identifies the future of process mining. Section 7 discusses the results, and Section 8 concludes this paper by preparing future works.

2. Method

This section proposes a quantitative method for understanding the process mining discipline by analyzing research fields and trends based on text mining and machine learning. As shown in Figure 1, the proposed framework consists of five steps: (1) data preparation, (2) research field (RF) identification, (3) RF investigation, (4) RF relation identification, and (5) future prediction for RF. In step 1, we prepare a collection of process mining research documents using a bibliographic search. Then, the feasible data is constructed based on a series of data preprocessing techniques. Step 2 aims to identify research fields by employing the latest topic modeling approach. Afterward, step 3 investigates the derived research fields on yearly trends, publication types, and nationality. Step 4 identifies the relations between research fields based on network analysis. Finally, we predict the future of each research field by employing a machine learning technique in step 5.

2.1. Data Preparation

The data preparation aims to build feasible data to understand process mining. To this end, we first exploit a bibliographic search from the *Web of Science* database produced by *Clarivate*. The bibliographic search signifies identifying a suitable collection of research papers by combining publications' diverse fields, such as authors, keywords, and abstracts, with search words relating to a specific research discipline. In this research, we retrieve research documents containing *process mining* and *workflow mining* in the title, keyword, and abstract. This is because process mining is

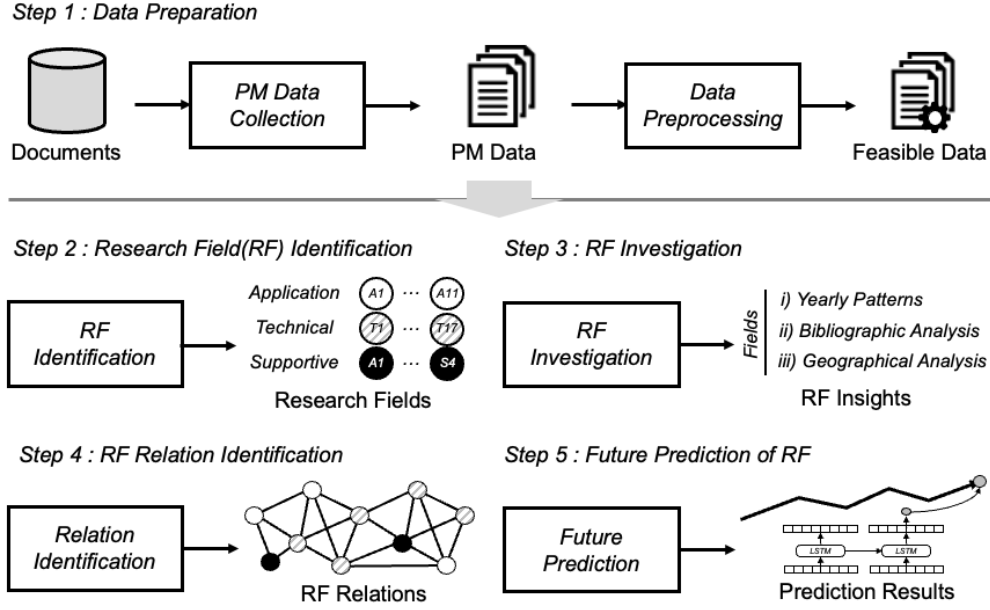


Figure 1: The five steps of the proposed method

the discipline directly connected to this study, while workflow mining is a similar area of research conducted before process mining was first introduced. Also, only publications written in English in the form of articles and conference proceedings are extracted. As a result of this, we collect 2,271 research papers.

In the data preprocessing phase, we apply a series of methods to the collected data. First, duplicated records with the same title and authors are removed from the collected data. These include cases of deriving a partial development from the previously conducted research or publishing the same article on multiple sites. As such, we delete these 10 redundant records. Second, we remove records that do not contain essential fields of publications, i.e., title, source, publication type, authorities, abstract, addresses, and year. As such, 55 documents that do not contain any of the fields are removed from the data, leading to a total of 2,207 publications as the feasible dataset. Also, we eliminate numbers, symbols, and stopwords in the abstract of the papers since these are the factors hindering effective topic extractions when a topic modeling approach is applied. In addition to this, we apply case conversion and lemmatization, i.e., the most fundamental preprocessing for text data, to reduce unnecessary noise and increase the accuracy of the results.

2.2. Research Field Identification

Based on the feasible data related to the process mining discipline derived from the previous step, this step aims to derive specific research fields based on topic modeling techniques. Topic modeling is a text mining technique that identifies the topics in a collection of words or documents by discovering the hidden semantic structure of documents [23]. There are several popular techniques, such as *Latent Semantic Analysis* [24], *Latent Dirichlet Allocation* [25], and *BERTopic* [26]. In this work, we employ *BERTopic* to utilize a pre-trained language model, e.g.,

BERT(Bidirectional Encoder Representations from Transformers) [27], to compensate for the short history of the process mining field.

BERTopic consists of three steps: document embedding, clustering, and topic representation. First, the document embedding produces real-valued vectors for given words and sentences using a BERT-based pre-trained language model. A BERT framework is an approach that only employs the self-attention of the encoder in the transformer structure, which is one of the representative generation-based self-supervised learning methods that derive semantic features of language using masked sentences as input. Based on this framework, numerous pre-trained language models have been constructed, and we utilize *sentence-BERT(SBERT) pre-trained model* to embed our data into vector spaces. Second, the embedded data are clustered to group documents with similar topics into the same clusters. To that end, we apply dimension reduction techniques [28], e.g., *PCA*, *t-SNE*, and *UMAP*, and clustering techniques [29], including *K-means* and *HDBSCAN*. Finally, the topic representation step produces a collection of important keywords for each cluster to understand the topics. A class-based TF-IDF is used for this purpose and formalized as follows:

$$W_{x,c} = tf_{x,c} \cdot \log \left(1 + \frac{A}{tf_x} \right) \quad (1)$$

, where c is a cluster derived as a result of clustering and x a term included in cluster c . Here, $tf_{x,c}$ indicates the Term Frequency (TF) of x in cluster c . tf_x and A represent the TF of x in the whole document before clustering and the average number of words per class, respectively. Then, the Inverse of Document Frequency (IDF) becomes the logarithm of A divided by tf_x . Note that 1 is added to hinder the output of the negative value. By combining these two values, i.e., TF and IDF, we measure the importance of the term x in cluster c .

Based on a set of clusters and the critical keywords for each cluster, we transform the discovered clusters into meaningful research fields. As such, a qualitative method based on expert reviews is used, and as a first step, candidates of research fields are derived by analyzing keywords included in each cluster. After that, we reconstruct and group the research fields in consideration of the goals and surroundings and finalize a set of the research fields of a specific discipline. The results will be explained in Section 3.

2.3. Research Field Investigation

This step investigates the derived research fields to obtain meaningful insights by identifying common aspects of them. In this regard, this study focuses on the three viewpoints among the multitude: yearly pattern analysis, bibliographic analysis, and geographical analysis.

First, the yearly patterns attempt to find out the trends for process mining and the specific research fields over time. To this end, it is possible to determine the exemplars of the advancement of research topics, such as dominant groups that have been in the spotlight from the past to the present, skyrocketing groups that have recently gained attention, and deep-cut groups that do not currently receive interests unlike in the past.

Second, the bibliographic analysis investigates the publication types of the process mining discipline. In detail, we can figure out what type of publications the researchers prefer, i.e., journals or conference proceedings, for the whole discipline and each category. Also, it delivers the most popular sources of research papers in the process mining field, which helps to understand what sources people consider to be a target for publication on a particular topic.

The geographical analysis aims to identify which geographical areas are the most advanced to study in process mining. To this end, we only consider the nationality of the corresponding authors of articles to make it more clear. The results for the yearly pattern analysis, bibliographic analysis, and geographical analysis will be presented in Section 4.

2.4. Relation Identification

This section presents a network analysis method to obtain insights into the relationship between the research fields. The network analysis method used in this study is similar to a typical social network. In other words, the network is composed of research topics as nodes and the similarity between topics as edges [30]. To accomplish this, we perform a series of steps: keyword extraction for each research topic, similarity measurement between research topics based on the keywords, network derivation, and community detection.

The keyword extraction aims to derive key terms of research topics that are used as the inputs of the similarity measurement. To this end, we do not consider the results from topic modeling because the goals for performing these tasks are different. In the research field identification step, the method, i.e., class-based TF-IDF, aims at eliciting keywords representing only a specific topic but not included in other groups. However, it is necessary to investigate keywords that can describe a particular group rather than comparing different groups for this similarity analysis. For this reason, we employ an alternative method to resolve these issues. It is possible to use all kinds of effective methods, including *YAKE!* [31], *KP-Miner* [32], *RakUn* [33], *LexRank* [34], and *SIFRank+* [35]. Among them, we utilize *YAKE!*, i.e., an unsupervised lightweight statistical method that derives the most relevant keywords based on the statistical text features due to the advantage of few restrictions on document size and a high-quality of extracted keywords.

Next, the similarity measurement aims to investigate the noticeable relations between research fields to build a network. To that end, we adopt the Jaccard similarity commonly employed in the natural language processing area [36]. The Jaccard similarity is defined as follows:

$$J(K_A, K_B) = \frac{|K_A \cap K_B|}{|K_A \cup K_B|} = \frac{|K_A \cap K_B|}{|K_A| + |K_B| - |K_A \cap K_B|} \quad (2)$$

, where K_A is the set of keywords for research topic A and K_B is the set of keywords for research topic B extracted by *YAKE!*. K_A and K_B are directly compared to assess how much they overlap quantitatively. Note that the similarity is calculated as 1 when all the items of two fields are overlapped and 0 in the opposite case.

Next, the network derivation aims to derive graph $G = (V, E, \#, w)$ based on the similarity between research topics. The set of nodes $V = RF \times T$ is a set of pairs of a research field and a time step, e.g., $(rf, t) \in V$ is the node for

the research field $rf \in RF$ at time $t \in T$. The set of edges $E \subseteq V \times V$ is a set of node pairs such that, for any $e = ((rf_1, t_1), (rf_2, t_2)) \in E$, $rf_1 = rf_2$. A count function $\# \in V \rightarrow \mathbb{N}$ maps nodes to publication counts. A weight function $w \in E \rightarrow \mathbb{R}$ maps edges to Jaccard similarities, i.e., for any $e = (v_1, v_2) \in E$, $w(e) = J(kw(v_1), kw(v_2))$, where kw is a function mapping a node to a set of keywords.

The derived network basically has the advantage of grasping the relationship between research topics at a glance. Nevertheless, our study intends to in-depth explore the relationship among the research fields based on a community detection method. Community detection refers to assembling nodes with high connection density into a single group [37]. Through this, it is possible to analyze which communities exist on the entire network and what characteristics each community has. In this study, we employ *Louvain algorithm* that can derive a community in a quick and effective manner [38]. The Louvain algorithm utilizes modularity to compare the density of links of inner- and inter-community. More in specific, this algorithm uses the modularity delta concept defined in Equation 3, which aims to determine whether shifting a specific node v into a community C is feasible.

$$\Delta Q(v \rightarrow C) = \left[\frac{\sum_{in} + k_{v,in}}{2m} - \left(\frac{\sum_{tot} + k_v}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_v}{2m} \right)^2 \right] \quad (3)$$

In the formula, m , $\sum_{in}$, and $\sum_{tot}$ indicate the sum of all link weights in the network, the sum of link weights between nodes in community C , and the sum of all link weights of nodes in C , respectively. Also, $k_{v,in}$ is the sum of link weights between nodes in C and node v , whereas k_v is the sum of all link weights of node v . The modularity delta, ΔQ , represents the difference in modularity contribution before and after merging node v in the community C , i.e., optimization. This phase is repeated until the modularity is optimized, and the finally obtained communities are compressed into super-nodes, i.e., aggregation. After that, the optimization of the modularity delta at the super-nodes level continues. We obtain a set of communities by persisting these two optimization and aggregation steps.

The results for the similarity measurement, network detection, and community detection will be presented in Section 5.

2.5. Future Prediction

This section aims to predict future research trends for the identified research fields. To this end, this work employs the results of previous steps in the method part. More in detail, we build a prediction model to predict the number of papers for each research field in the future based on the given historical number of papers for fields and their similarities. Also, we employ Long-Short Term Memory neural networks (LSTM) since it is robust for learning temporal dependency [39].

In order to predict the number of publications in a specific research field in the future, we use a three-dimensional publication presentation matrix $R \in \mathbb{R}^{|RF| \times |RF| \times |T|}$ where RF and T are the set of research fields and time steps, respectively. As shown in Figure 2(a), an element of the matrix, $r_{i,j,t} \in R$, refers to the contribution value of research field rf_i at time step $t - 1$ to research field rf_j at time step t . Given network $G = (V, E, \#, w)$ used in Subsection 2.4, $r_{i,j,t} \in R$ is formalized as follows:

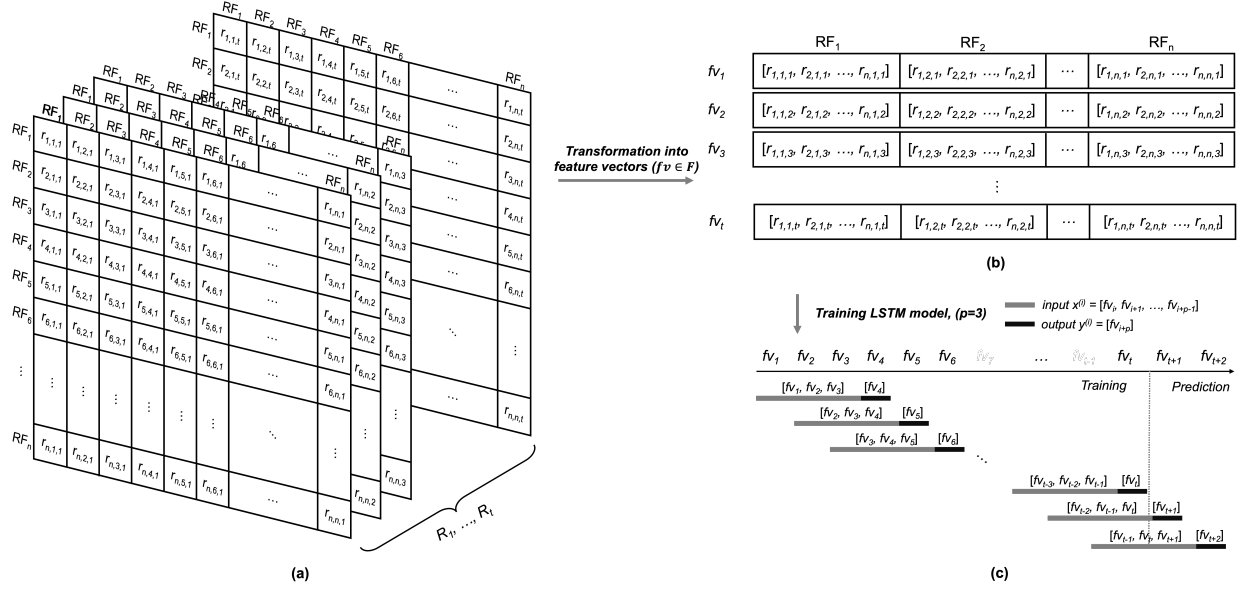


Figure 2: (a) Three-dimensional publication representation matrix, (b) Feature vector, (c) Training and prediction procedures

$$r_{i,j,t} = \frac{w((rf_i, t-1), (rf_j, t))}{\sum_{1 \leq k \leq |RF|} w((rf_k, t-1), (rf_j, t))} \times \#((rf_j, t)) \quad (4)$$

, where $(rf_i, t-1) \in V$ indicates research field rf_i at time step $t-1$ and $(rf_j, t) \in V$ indicates research field rf_j at time step t .

Next, we transform the three-dimensional publication presentation matrix $R \in \mathbb{R}^{|RF| \times |RF| \times |T|}$ into a feature vector $F \subseteq \mathbb{R}^{|RF|}$ such that, for any $f_{vt} \in F$, $f_{vt} = [f_{v1,t}, f_{v2,t}, \dots, f_{v|RF|,t}]$ where $f_{v_{j,t}} = [r_{i,j,t} | 1 \leq i \leq |RF|]$ for $1 \leq j \leq |RF|$, as shown in Figure 2(b). For instance, $f_{v1} = [f_{v1,1}, f_{v2,1}, \dots, f_{v|RF|,1}]$ is a feature vector at time $t = 1$ where $f_{v1,1} = [r_{1,1,1}, r_{2,1,1}, \dots]$ indicates the contribution values of all research fields to research field rf_1 at $t = 1$, $f_{v2,1} = [r_{1,2,1}, r_{2,2,1}, \dots]$ indicates the contribution values of all research fields to research field rf_2 at $t = 1$, etc.

As shown in Figure 2(c), the LSTM model is trained to predict the feature vector at $t+1$, given the last $1 \leq p \leq t$ feature vectors. Therefore, the i -th training sample consists of $x^{(i)} = [f_{v_i}, f_{v_{i+1}}, \dots, f_{v_{i+p-1}}]$ and $y^{(i)} = f_{v_{i+p}}$. For instance, given $p = 3$, the first training sample consists of $x^{(1)} = [f_{v_1}, f_{v_2}, f_{v_3}]$ and $y^{(1)} = f_{v_4}$, while the second training sample consists of $x^{(2)} = [f_{v_2}, f_{v_3}, f_{v_4}]$ and $y^{(2)} = f_{v_5}$. Based on the predicted feature vector at $t+1$, i.e., $f_{v_{t+1}}$, we compute the predicted number of papers, for research field rf_j at $t+1$, i.e., $R_{j,t+1}$, as $\sum_{1 \leq i \leq |RF|} [r_{i,j,t} \in f_{v_{j,t+1}}]$.

3. Identification of research fields in process mining

In this section, we explain the topics discovered from the topic modeling presented in Subsection 2.2. First, we map each discovered topic into the existing field of process mining by analyzing the keywords and group the topics

into three categories by qualitatively analyzing the similarity between topics. Next, we explain the individual topics in detail according to the categories.

3.1. Topic Mapping & Grouping

First, we map the discovered topics from text mining to the existing fields in process mining by analyzing the keywords representing each topic. Figure 3(a) shows the discovered topics mapped to the existing fields in process mining. For instance, *topic 6* is mapped to *general process discovery* since it contains keywords such as *discovery*, *discovered*, *discovering*, etc. *Topic 7* is mapped to *conformance checking* since its keywords include *conformance*, *compliance*, *checking*, etc. Moreover, *Topic 29* is mapped to *healthcare* since it includes keywords such as *patient*, *clinic*, *hospital*, etc.

Next, we categorize the topics 1) by merging the topics that share the same goals and 2) by classifying them into three categories. We merge the topics based on the similarities in the associated fields, as shown in Figure 3(b), resulting in 32 merged topics. For instance, *workflow mining*, *genetic algorithm-based mining*, *region-based mining*, *pattern mining*, *Petri-nets*, and *general process discovery* are merged into *process discovery* since they all aim to discover process models from event logs (but in different formalism with different approaches). *Conformance checking* and *alignments* are merged into *conformance checking* since *alignments* is an approach to *conformance checking*. Moreover, *smart artifacts* and *Internet of Things (IoT) security* are merged into *IoT*.

We group the merged fields into three categories: *Technique*, *Application*, and *Support*. The technique category contains process mining techniques, whereas the application category contains topics that focus on applying such process mining techniques in different domains such as healthcare, software, education, manufacturing, etc. In addition, the support category contains topics that provide the fundamentals for the development of process mining techniques. Figure 3(c) shows the grouping of different fields. For instance, *process discovery*, *conformance checking*, and *performance-driven process mining* are classified into the technique category, whereas *healthcare*, *software*, and *education* are grouped into the application category. Moreover, *machine learning*, *business intelligence*, and *business process management* are categorized into the support category.

3.2. Topic Description

In this subsection, we explain the 32 merged topics from the previous subsection based on the three categories.

3.2.1. Technique Category

Figure 4 shows a word cloud representing keywords that explain topics that belong to this category. The keywords include *workflow*, *discovery*, *algorithm*, *behavior*, *time*, and *trace*.

Below are the topics in the technique category:

- **T1: Process Discovery (451 articles)** aims to discover process models from event logs. Different articles propose different process discovery techniques by varying the modeling formalism of the discovered process

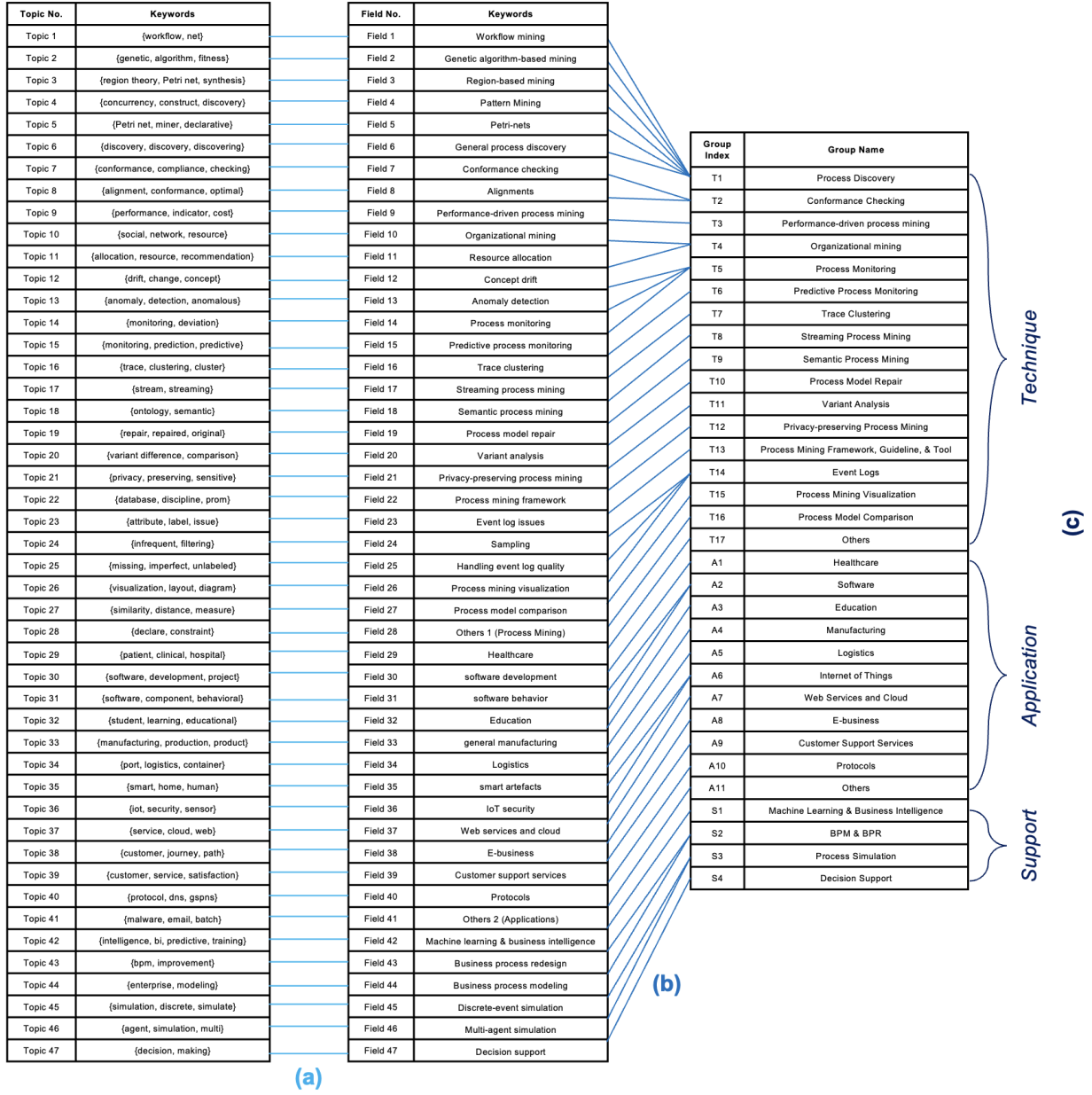


Figure 3: (a) Mapping 47 discovered topics to 47 existing fields of process mining based on the associated keywords, (b) Merging 47 research fields into 32 fields, (c) Classifying 32 research fields into three categories

model, such as Petri nets, BPMNs, UML diagrams, etc. Moreover, for the same modeling formalism (e.g., Petri nets), different articles propose different approaches (e.g., region-based approaches, genetic algorithm-based approaches, etc.).

- **T2: Conformance Checking (136 articles)** aims to compute the conformance of event logs to process models using various conformance measures. Different articles propose different techniques to tackle different

logs. Different techniques focus on 1) deriving process features that represent process instances, 2) computing coherent clusters using different clustering algorithms, 3) and explaining the discovered clusters.

- **T8: Streaming Process Mining (17 articles)** aims to analyze streaming event logs, unlike other process mining techniques that analyze historical event logs. Different techniques focus on 1) storing streaming event logs, 2) discovering process models from streaming event logs, and 3) checking the conformance of running process instances.
- **T9: Semantic Process Mining (25 articles)** aims to leverage semantic information in the form of knowledge graphs and process ontology to enhance the existing process mining techniques. Different techniques focus on enhancing different process mining techniques using different forms of semantic information.
- **T10: Process Model Repair (18 articles)** aims to support repairing process models to reflect reality. Existing techniques first uncover differences between observed behavior (i.e., event logs) and modeled behavior (i.e., process model) and then modify process models while ensuring the soundness of the model.
- **T11: Variant Analysis (17 articles)** aims to analyze different forms of variants existing in business processes. First, process variant analysis aims to improve the original reference model such that it fits variant process models. Second, trace variant analysis aims to analyze unique sequences of activities by different process instances based on event logs.
- **T12: Privacy-preserving Process Mining (15 articles)** aims to protect event data privacy such that the data conform to privacy regulations on the technical design and the organizational application of process mining techniques. Existing work provides privacy-preserving data publishing techniques to anonymize event data while maintaining the utility of the data.
- **T13: Process Mining Framework, Guideline, & Tool (72 articles)** aims to provide frameworks to develop process mining techniques, guidelines to apply process mining techniques in various domains, and tools to support the implementation of process mining techniques. The ProM framework provides a software framework to support the process mining tooling, and the process mining manifesto describes opportunities and challenges in process mining.
- **T14: Event Logs (101 articles)** provides various techniques to support the management of event logs. First, extraction techniques support the extraction of event logs from source systems (e.g., SAP ERP systems). Second, sampling techniques support the sampling of event logs to enhance the performance of various process mining techniques such as conformance checking. Finally, log quality techniques support the assessment of log qualities and their enhancement, e.g., by handling missing information.

- **T15: Process Mining Visualization (18 articles)** aims to provide various visualizations of process mining insights. Existing techniques focus on improving the visualization of process models by applying graph layout algorithms and supporting the visualization of social networks using various diagrams, e.g., Chord diagram.
- **T16: Process Model Comparison (13 articles)** aims to analyze the similarity of process models. Some of the techniques analyze the structure of process models, whereas others analyze the semantics of process models.
- **T17: Others (56 articles)** contains articles that provide the design/implementation of process mining techniques but do not belong to a specific existing field. Such articles deal with *Big Data*, agile development, natural language processing, etc.

Figure 5 shows a word cloud representing keywords that explain topics that belong to the application category. The keywords include *case study*, *healthcare*, *knowledge*, and *application*.

Figure 5: Word cloud representing keywords associated with the technique category

- **A1. Healthcare (267 articles)** aims to analyze healthcare-related business processes using process mining techniques. Existing articles present clinical pathway patterns from medical behaviors and analyses of the length of hospital stay based on process discovery and performance-driven process mining techniques.
- **A2. Software (103 articles)** aims to analyze software development processes and software behavior using process mining techniques. Whereas the former focuses on analyzing business processes related to software developments, the latter focuses on the system/user behavior of the developed software.

- **A3: Education (78 articles)** aims to analyze the learning processes of students in different formats of education, e.g., online learning, self-regulated learning, etc. Process discovery techniques are often deployed to provide an overview of students' learning behavior.
- **A4: Manufacturing (51 articles)** aims to analyze end-to-end manufacturing processes, including planning (i.e., balancing demand and supply), sourcing (i.e., procuring materials), making (i.e., transforming materials to finished products), delivering (i.e., providing finished products to retailers), and returning (i.e., managing used products) processes. Various manufacturing domains have been studied, such as semiconductor manufacturing, make-to-order manufacturing, the automotive industry, etc.
- **A5: Logistics (12 articles)** aims to analyze business processes related to logistics and storage management. The application focuses on predicting the Estimated Time of Arrival (ETA), scheduling transport, improving warehouse layout, etc. The proposed applications rely on tagging systems such as the Radio Frequency Identification (RFID) system to identify the movement of goods.
- **A6: Internet of Things (50 articles)** aims to analyze the data generated or captured by the Internet of Things (IoT) using process mining techniques. On the one hand, the existing application focuses on analyzing human behavior patterns, e.g., the behavior pattern of residents in a nursing home based on a wireless tracking system. On the other hand, the application concerns the detection of misbehavior in IoT-based systems to enhance the security of the system.
- **A7: Web and Cloud Services (32 articles)** aims to analyze processes in Web and Cloud services, e.g., Google Cluster, to optimize resources. To that end, process mining techniques are used to discover user behavior by analyzing execution logs and monitoring the violation of business constraints. Furthermore, the insights from process mining techniques are used to optimize the scheduling of resources.
- **A8: E-business (18 articles)** aims to analyze customer journeys in online business processes on the web. The existing application uses the click stream data of users to automatically discover customer journey maps. Comparative process mining techniques are applied to compare the behavior of different types of customers, e.g., the behavior of bargain shoppers and surgical shoppers.
- **A9: Customer Support Services (15 articles)** aims to analyze business processes of help desks and customer services that resolve reported issues. As well as the insights from process mining are used to improve the process, they are used to develop new customer support services, e.g., adaptive automated teller machines based on the insights from the usage analysis of automated teller machines.
- **A10: Protocols (13 articles)** aims to analyze protocols of software systems. On the one hand, the existing application aims to discover protocols of proprietary legacy software systems, e.g., TCP/IP-Protocol. On the

other hand, the application concerns the detection of anomalies in protocol traces, e.g., a mail bonnet attack in DNS protocol traces.

- **A11: Others (121 articles)** contains articles that employ process mining but do not belong to a specific domain. Such articles are engaged in the following domains: e-mail, *smart manufacturing*, *industry 4.0*, *robotic process automation*), *blockchain*, etc.

3.2.3. Support Category

Figure 6 shows a word cloud that represents keywords explaining topics in this category. The keywords include *user*, *resource*, *knowledge*, and *modeling*.



Figure 6: Word cloud representing keywords associated with the technique category

Below are the topics in the support category:

- **S1: Machine Learning and Business Intelligence (73 articles)** relates techniques for machine learning and business intelligence to enable process mining techniques such as process discovery, conformance checking, performance analysis, etc. For instance, statistical relational learning supports the discovery of process models, and frequent itemset mining supports organizational mining.
- **S2: Business Process Management and Business Process Redesign (77 articles)** relates techniques in business process management and business process redesign. The formal provides the fundamentals of process models that are used as modeling formalism employed in process discovery techniques, whereas the latter introduces redesign patterns in business processes that support the process improvement based on the insights from process mining techniques.
- **S3: Process Simulation (40 articles)** refers to the study related to discrete event simulation of business processes. Based on the existing simulation techniques, process mining provides techniques for discovering such

simulation models to automate the design and implementation of the simulation models. Some techniques automatically discover simulation models with more advanced constructs, such as the presence of multitasking and availability constraints.

- **S4: Decision Support (16 articles)** refers to the design and implementation of decision support systems. The suggested decision support systems support the adoption of process mining techniques to realize informed decisions based on data-driven insights and automatically generate recommendations (i.e., decisions).

4. Investigation into research fields in process mining

This section investigates the research fields derived in Section 3 based on the relevant common aspects: yearly patterns, bibliographic analysis, and geographical analysis.

4.1. Yearly Patterns

Figure 7 presents the number of publications in each research field by year. As depicted in the figure, the process mining discipline was introduced in the early 2000s and has been gradually developed due to many efforts. In particular, the number of publications has been over 100 since 2013, and almost half of the papers were published in the last five years. In other words, this field has recently received more attention. Also, compared to the support category, both technique and application domains have been in the same flow by year. That is, a bunch of applications specialized in various industries has appeared with the development of numerous process mining techniques.

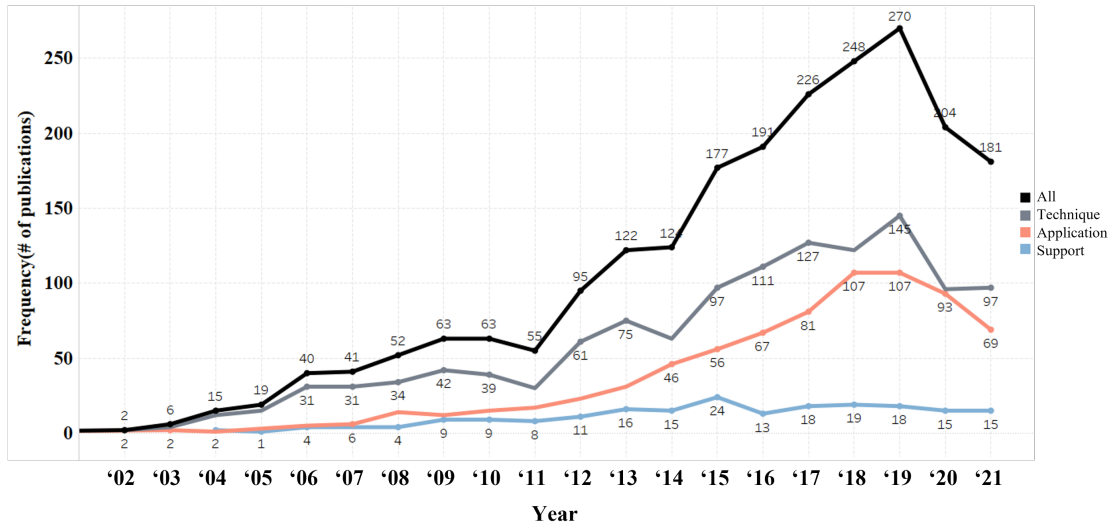
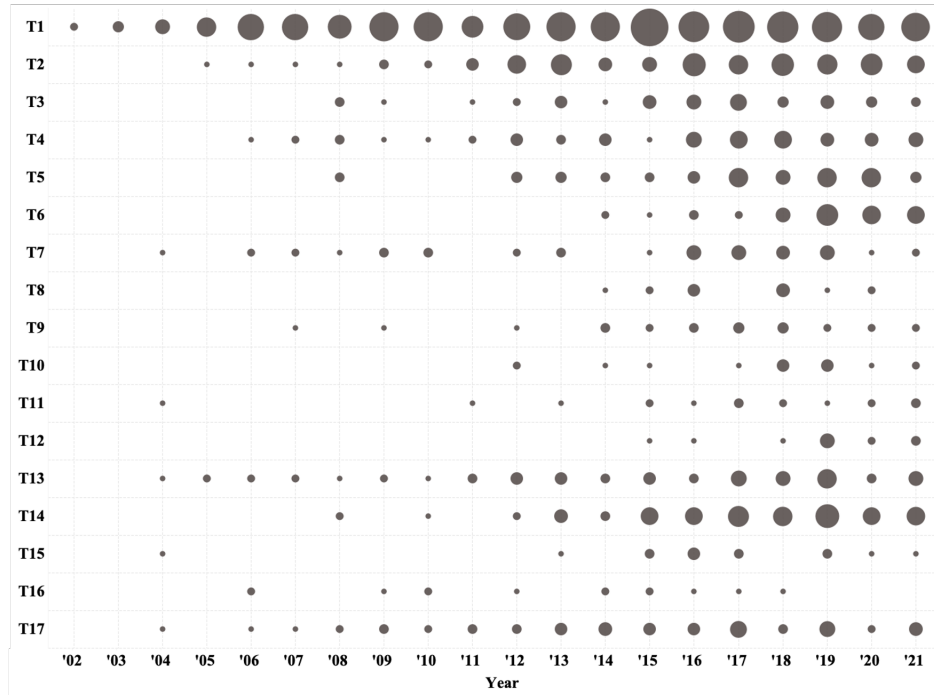
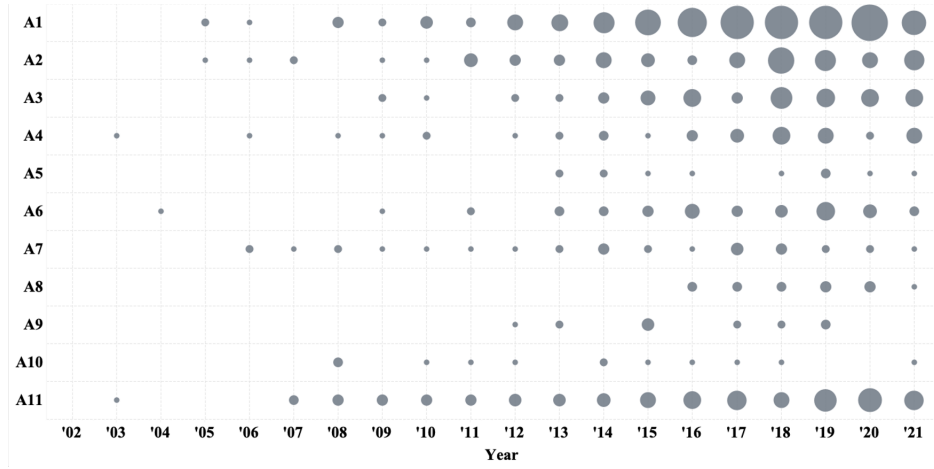


Figure 7: The number of publications in each category by year

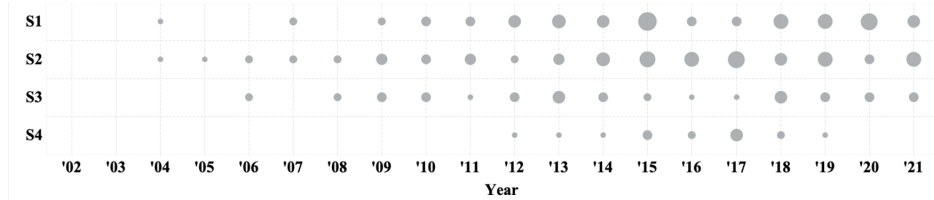
Next, we explore the yearly patterns for each research field in the fine-grained viewpoints, which aims at classifying them into the following archetypes: dominant, skyrocketing, and deep-cut patterns. Figure 8(a) presents the number of publications by year for each research field in the technique category.



(a) Technique category



(b) Application category



(c) Support category

Figure 8: The number of publications in each research field by year

In the technique category, process discovery (T1), conformance checking (T2), and process mining framework, guideline, & tool (T13) belong to the dominant groups. For example, T1 is the most fundamental field in the process mining discipline. In other words, it has been constantly studied every year, and new algorithms have been developed even recently. More in detail, this field first appeared as workflow mining techniques, such as *alpha-algorithm*. Also, it covered both initial process discovery algorithms, e.g., *heuristic mining* and *fuzzy mining*, and modern techniques, e.g., *inductive mining*.

Moreover, researchers in the process mining community have taken a keen interest in conformance checking, i.e., T2. It first emerged with initial research on measuring conformance based on the token replay approach in 2008. Since then, it has been steadily researched and stands as an essential research topic in process mining with modern techniques, such as multi-perspective, stochastic, and online conformance checking. Regarding T13, we can easily understand that they belong to the dominant group since *process mining framework, guideline, and tools* are the essential components of process mining.

As far as the skyrocketing patterns are concerned, we can consider process monitoring (T5), predictive process monitoring (T6), privacy-preserving process mining (T12), and event logs (T14) in the technique category. Process monitoring and predictive process monitoring, i.e., research areas that have expanded from the naive performance analysis, have achieved much growth with the development of machine learning-related technologies in recent years. Research on T5 has been more actively published since 2017 with the aim of catching sudden changes in performance measurements in the process environment with machine learning approaches, including anomaly detection and concept drift. On the other hand, predictive process monitoring, i.e., T6, has received much attention since 2019 based on multiple deep learning algorithms to predict performance at the process instance level. Regarding T12, there have been some efforts to preserve the privacy of event logs in recent years, with a trend that data security has become essential in the data science field. Also, T14 has been steadily progressing in research in various directions, including event abstraction, complex event processing, data cleansing, and quality. Furthermore, it has recently received strong attention due to the emergence of a new type of event logs, e.g., object-centric event logs.

Lastly, we classify the process model comparison topic (T16) into the deep-cut group. This topic aims to compare the two given process models structurally, and recent studies have been reduced due to limitations in their utilization and, thus, low demands for new technological developments.

The yearly pattern results for the application and support categories are depicted in Figure 8(b) and Figure 8(c), respectively. Among the research fields, the followings are included in the dominant group: healthcare (A1), software (A2), education (A3), manufacturing (A4), machine learning and business intelligence (S1), and business process management and business process redesign (S2). The healthcare domain is one of the representative fields where process mining is actively and steadily employed. The articles in this field have covered various healthcare processes, including outpatient, inpatient, and emergency departments, and they triggered the publication of review papers on process mining in healthcare. In addition, process mining has been consistently applied to the software, education, and manufacturing areas from the past to the present. In the support category, the number of publications in S1 and

S2, i.e., machine learning and business process management, is insufficient compared to other classes, but these also have been popular subjects for researchers. Also, we identify only two research fields in the application category, i.e., internet of things (A6) and e-business (A8) are considered as the skyrocketing groups, whereas customer support services (A9), protocols (A10), and decision support (S4) are the topics deep-cut.

4.2. Bibliographic analysis

This section aims to obtain new insights about bibliographic features by research fields and the whole process mining discipline. In more detail, we investigate the proportions of publication types and specific sources mainly published in process mining.

First, Figure 9 shows the result of comparing the proportions of articles according to the publication type for each category and the whole. The process mining discipline has published 32% of journal articles and 68% of conference proceedings, which is similar to the distribution of general computer science disciplines. In addition, by exploring the results at the category level, we can identify that application-oriented articles have been published more in journals than those in the technique and support categories.

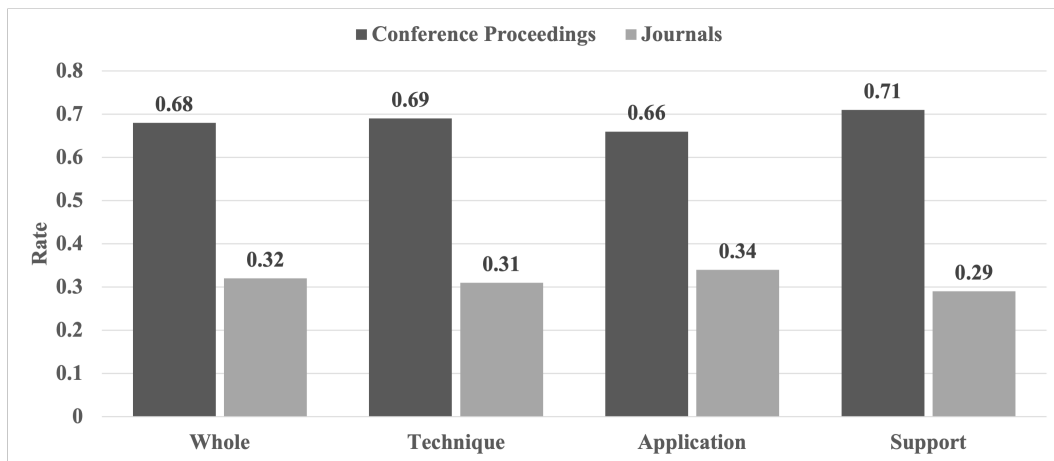


Figure 9: The proportions of publication types in process mining

Next, we perform the same analysis for each research field, and Figure 10 presents its results. It is recognizable that the number of conference proceedings exceeds that of journal articles in most cases. Specifically, semantic process mining (T9), process mining framework, guideline, & tool (T13), customer support services (A9), and decision support (S4) have a significantly higher proportion of conference proceedings. On the other hand, the number of journal articles is more than the number of conference proceedings in logistics (A5) and process model repair (T10).

Next, we identify the most frequently published publication sources in journals and conference proceedings for each category and the whole discipline. Table 1 describes the results to identify the top seven sources of publications. Many journal articles have been published in the disciplines of computer science, information systems, software

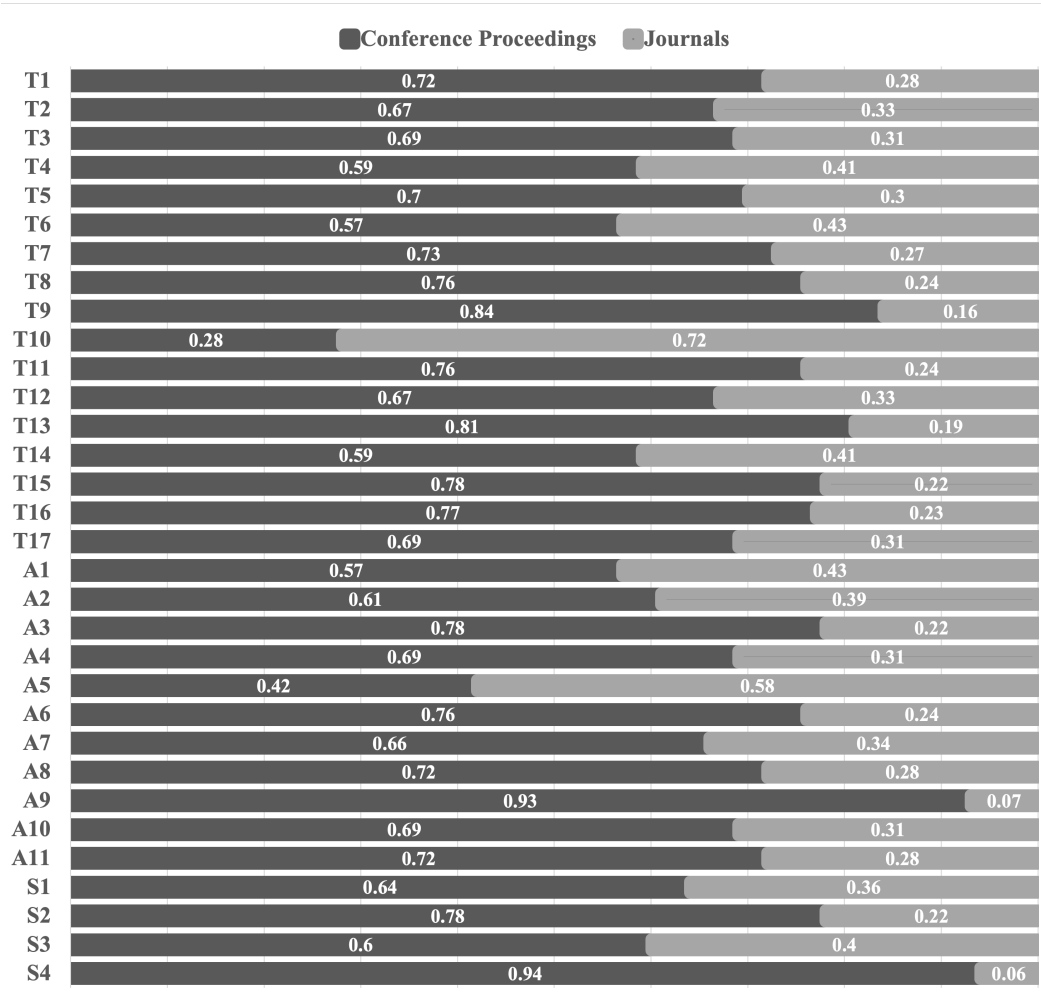


Figure 10: The proportions of publication types for each research field

engineering, and management science, such as *Information Systems*, *Expert Systems with Applications*, *Decision Support Systems*, and *Software and Systems Modeling*. This tendency is similar in the technique and support categories. However, the application-oriented articles show disparate results, i.e., they have been published in multi-disciplinary journals focusing on producing case studies of process mining, such as *IEEE Access*, *Applied Sciences*, and *Sensors*. Furthermore, domain-specific journals, e.g., the healthcare-specialized, are listed in Table 1 as well, including *International Journal of Environmental Research and Public Health* and *Journal of Biomedical Informatics*.

As far as the conference proceedings are concerned, there is no significant discrepancy in results by the whole and each category. Most of the research papers have been published in conferences on *Business Process Management (BPM)*, *Advanced Information Systems Engineering (CAiSE)*, *Enterprise Information Systems (ICEIS)*, and *On The Move To Meaningful Internet Systems (OTM)*. After the initialization of the *Process Mining Conference (ICPM)* in 2019, a large number of up-to-date research articles have been published through this.

Table 1: The top 7 journals and conference proceedings by the whole and categories

	Journal	Conference Proceedings
Whole	Information Systems IEEE Access Expert Systems with Applications Decision Support Systems Software and Systems Modeling IEEE Transactions on Services Computing Applied Sciences	Business Process Management (BPM) Advanced Information Systems Engineering (CAiSE) Enterprise Information Systems (ICEIS) ICT and Knowledge Engineering (ICT-KE) On The Move To Meaningful Internet Systems (OTM) Process Mining (ICPM) Enterprise Distributed Object Computing (EDOC)
Technique	Information Systems IEEE Access Expert Systems with Applications Software and Systems Modeling Decision Support Systems IEEE Transactions on Services Computing Data and Knowledge Engineering	Business Process Management (BPM) On The Move To Meaningful Internet Systems (OTM) Advanced Information Systems Engineering (CAiSE) Enterprise Information Systems (ICEIS) Enterprise Distributed Object Computing (EDOC) Process Mining (ICPM) Conceptual Modeling (ER)
Application	International Journal of Environmental Research and Public Health IEEE Access Expert Systems with Applications Journal of Biomedical Informatics Applied Sciences Sensors Information Systems	Business Process Management (BPM) ICT and Knowledge Engineering (ICT-KE) Enterprise Information Systems (ICEIS) Biomedical Engineering Systems and Technologies (BIOSTEC) Learning Analytics & Knowledge (LAK) Advanced Information Systems Engineering (CAiSE) Process Mining (ICPM)
Support	Information Systems Software and Systems Modeling IEEE Transactions on Services Computing Expert Systems with Applications Decision Support Systems Data and Knowledge Engineering Computers in Industry	Business Process Management (BPM) Advanced Information Systems Engineering (CAiSE) Enterprise Information Systems (ICEIS) Simulation Conference (WSC) Process Mining (ICPM) Enterprise Distributed Object Computing (EDOC) Business Information Systems (BIS)

4.3. Geographic Analysis

The final goal of investigating research articles in process mining is to identify the discrepancy of geographical aspects by research fields. To this end, we uncover the leading countries that frequently publish process mining articles. To obtain more explicit insights, we analyze the country where the affiliation of the corresponding author of each paper is located. Table 2 describes the results of this.

Table 2: The top 7 countries by the whole and categories

	Whole	Technique	Application	Support
Country	Netherlands	Netherlands	Italy	Germany
	China	China	China	Netherlands
	Italy	Germany	Netherlands	China
	Germany	Italy	Spain	Brazil
	Spain	Australia	France	France
	South Korea	South Korea	Germany	Spain
	France	U.S.	Thailand	Belgium

Based on the results, we can identify that the Netherlands has occupied the most dominant position in the process mining discipline. In addition, China, Italy, and Germany have remarkably contributed to the development of the process mining field as well. These four countries have published almost 40% of the whole research articles, especially the Netherlands and China occupied 12.9% and 10.7% of the total, respectively.

The technique category shows a similar pattern to the whole process mining discipline. The Netherlands, China, Germany, and Italy have also contributed significantly to technique-related publications. However, unlike the overall distribution, many techniques have been developed in Australia and the United States as well. The application-oriented articles have been published by authors in Italy, China, and the Netherlands. Besides, Spain and France also have contributed to the emergence of the application of process mining to various domains. In the support category, Germany and the Netherlands have occupied a large part of this. In addition, Brazil and Belgium are also included in the list as the key regions for the support field.

5. Relation identification between research fields in process mining

This section aims at discovering noticeable relations between research fields in process mining. To achieve this objective, we employed a series of methods, e.g., similarity analysis, network construction, and community detection, as introduced in Subsection 2.4.

5.1. Similarity analysis

Before measuring the similarity between research fields, we apply YAKE! presented in Subsection 2.4 to extract principal terms to represent them. Then, we measured the similarity between research fields based on the Jaccard similarity method. Figure 11 describes the result of the Jaccard similarity measurements between research topics.

		Technique																	Application											Support			
		T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	S1	S2	S3	S4
Technique	T1	1.000	0.232	0.254	0.227	0.208	0.193	0.222	0.219	0.204	0.206	0.217	0.197	0.247	0.235	0.215	0.200	0.258	0.193	0.194	0.185	0.195	0.169	0.191	0.197	0.179	0.177	0.194	0.175	0.227	0.226	0.227	0.207
	T2	0.232	1.000	0.200	0.190	0.218	0.202	0.212	0.189	0.213	0.207	0.225	0.183	0.231	0.214	0.218	0.189	0.224	0.186	0.191	0.178	0.169	0.177	0.169	0.211	0.170	0.173	0.193	0.175	0.202	0.229	0.199	0.178
	T3	0.254	0.200	1.000	0.216	0.229	0.224	0.225	0.213	0.222	0.209	0.216	0.190	0.229	0.218	0.218	0.190	0.229	0.206	0.216	0.182	0.215	0.189	0.190	0.196	0.177	0.186	0.188	0.182	0.229	0.236	0.251	0.216
	T4	0.227	0.190	0.216	1.000	0.207	0.203	0.214	0.210	0.217	0.181	0.195	0.195	0.224	0.223	0.209	0.187	0.212	0.208	0.196	0.202	0.196	0.177	0.196	0.200	0.193	0.194	0.164	0.194	0.224	0.221	0.213	0.216
	T5	0.208	0.218	0.229	0.207	1.000	0.192	0.229	0.210	0.221	0.191	0.207	0.179	0.214	0.202	0.203	0.175	0.199	0.185	0.193	0.191	0.194	0.168	0.213	0.204	0.188	0.164	0.176	0.179	0.209	0.205	0.208	0.185
	T6	0.193	0.202	0.224	0.203	0.192	1.000	0.230	0.207	0.210	0.174	0.211	0.178	0.201	0.193	0.211	0.181	0.210	0.183	0.173	0.183	0.199	0.167	0.186	0.170	0.185	0.172	0.174	0.183	0.211	0.214	0.211	0.184
	T7	0.222	0.212	0.225	0.214	0.229	0.230	1.000	0.200	0.206	0.194	0.239	0.196	0.220	0.225	0.217	0.201	0.211	0.193	0.200	0.206	0.190	0.185	0.201	0.178	0.198	0.173	0.182	0.187	0.214	0.229	0.209	0.195
	T8	0.219	0.189	0.213	0.210	0.210	0.207	0.200	1.000	0.213	0.190	0.205	0.208	0.235	0.209	0.195	0.177	0.213	0.191	0.190	0.172	0.194	0.174	0.197	0.196	0.178	0.174	0.175	0.183	0.198	0.202	0.191	0.188
	T9	0.204	0.213	0.222	0.217	0.221	0.210	0.206	0.213	1.000	0.184	0.224	0.188	0.235	0.198	0.208	0.204	0.229	0.188	0.187	0.201	0.198	0.174	0.220	0.201	0.190	0.179	0.186	0.185	0.216	0.225	0.203	0.189
	T10	0.206	0.207	0.209	0.211	0.191	0.174	0.194	0.190	0.184	1.000	0.195	0.148	0.180	0.200	0.204	0.176	0.192	0.168	0.170	0.162	0.158	0.158	0.155	0.158	0.152	0.161	0.183	0.153	0.179	0.179	0.174	0.158
	T11	0.217	0.225	0.216	0.195	0.207	0.211	0.239	0.205	0.224	0.195	1.000	0.197	0.206	0.212	0.213	0.214	0.231	0.202	0.205	0.190	0.188	0.193	0.190	0.188	0.200	0.191	0.188	0.183	0.208	0.200	0.201	0.193
	T12	0.197	0.183	0.190	0.195	0.179	0.178	0.196	0.208	0.188	0.148	0.197	1.000	0.204	0.182	0.214	0.181	0.189	0.193	0.177	0.166	0.179	0.182	0.205	0.205	0.193	0.179	0.182	0.187	0.189	0.188	0.175	0.197
	T13	0.247	0.231	0.229	0.224	0.214	0.201	0.220	0.235	0.235	0.180	0.206	0.204	1.000	0.219	0.233	0.197	0.253	0.200	0.205	0.204	0.216	0.199	0.202	0.196	0.192	0.191	0.180	0.200	0.226	0.243	0.227	0.186
	T14	0.235	0.214	0.218	0.223	0.202	0.193	0.225	0.209	0.198	0.200	0.212	0.182	0.219	1.000	0.195	0.188	0.221	0.183	0.190	0.180	0.200	0.172	0.203	0.171	0.175	0.179	0.166	0.189	0.209	0.212	0.195	0.178
	T15	0.215	0.218	0.218	0.209	0.203	0.211	0.217	0.195	0.208	0.204	0.213	0.214	0.233	0.195	1.000	0.196	0.220	0.207	0.204	0.204	0.188	0.205	0.199	0.191	0.231	0.197	0.185	0.201	0.225	0.234	0.219	0.224
	T16	0.200	0.189	0.190	0.187	0.175	0.181	0.201	0.177	0.204	0.176	0.214	0.181	0.197	0.188	0.196	1.000	0.198	0.164	0.173	0.179	0.174	0.167	0.181	0.176	0.177	0.154	0.173	0.175	0.205	0.196	0.176	0.196
	T17	0.258	0.224	0.229	0.212	0.199	0.210	0.211	0.213	0.229	0.192	0.231	0.189	0.253	0.221	0.220	0.198	1.000	0.214	0.221	0.193	0.190	0.189	0.201	0.188	0.181	0.182	0.192	0.196	0.220	0.245	0.211	0.194
Application	A1	0.193	0.186	0.206	0.208	0.185	0.183	0.193	0.191	0.188	0.168	0.202	0.193	0.200	0.183	0.207	0.164	0.214	1.000	0.194	0.198	0.189	0.200	0.186	0.191	0.183	0.176	0.168	0.193	0.202	0.208	0.189	0.190
	A2	0.194	0.191	0.216	0.196	0.193	0.173	0.200	0.190	0.187	0.170	0.205	0.177	0.205	0.190	0.204	0.173	0.221	0.194	1.000	0.199	0.202	0.175	0.185	0.182	0.169	0.176	0.168	0.183	0.199	0.213	0.201	0.184
	A3	0.185	0.178	0.182	0.202	0.191	0.183	0.206	0.172	0.201	0.162	0.190	0.166	0.204	0.180	0.204	0.179	0.193	0.198	0.199	1.000	0.176	0.176	0.182	0.166	0.171	0.200	0.148	0.189	0.197	0.182	0.196	0.190
	A4	0.195	0.169	0.215	0.196	0.194	0.199	0.190	0.194	0.198	0.158	0.188	0.179	0.216	0.200	0.188	0.174	0.190	0.189	0.202	0.176	1.000	0.205	0.218	0.191	0.176	0.195	0.167	0.198	0.206	0.203	0.195	0.199
	A5	0.169	0.177	0.189	0.177	0.168	0.167	0.185	0.174	0.174	0.158	0.193	0.182	0.199	0.172	0.205	0.167	0.189	0.200	0.175	0.176	0.205	1.000	0.168	0.153	0.171	0.175	0.142	0.181	0.184	0.185	0.160	0.187
	A6	0.191	0.169	0.190	0.196	0.213	0.186	0.201	0.197	0.220	0.155	0.190	0.205	0.202	0.203	0.199	0.181	0.201	0.186	0.185	0.182	0.218	0.168	1.000	0.209	0.186	0.185	0.169	0.182	0.197	0.200	0.200	0.181
	A7	0.197	0.211	0.196	0.200	0.204	0.170	0.178	0.196	0.201	0.158	0.188	0.205	0.196	0.171	0.191	0.176	0.188	0.191	0.182	0.166	0.191	0.153	0.209	1.000	0.169	0.163	0.170	0.174	0.182	0.222	0.202	0.189
	A8	0.179	0.170	0.177	0.193	0.188	0.185	0.198	0.178	0.190	0.152	0.200	0.193	0.192	0.175	0.231	0.177	0.181	0.183	0.169	0.171	0.176	0.171	0.186	0.169	1.000	0.179	0.150	0.166	0.180	0.192	0.175	0.181
	A9	0.177	0.173	0.186	0.194	0.164	0.172	0.173	0.174	0.179	0.161	0.191	0.179	0.191	0.179	0.197	0.154	0.182	0.176	0.176	0.200	0.195	0.175	0.185	0.163	0.179	1.000	0.158	0.178	0.187	0.171	0.182	0.177
	A10	0.194	0.193	0.188	0.164	0.176	0.174	0.182	0.175	0.186	0.183	0.188	0.182	0.180	0.166	0.185	0.173	0.192	0.168	0.168	0.148	0.167	0.142	0.169	0.170	0.150	0.158	1.000	0.165	0.175	0.187	0.197	0.176
	A11	0.175	0.175	0.182	0.194	0.179	0.183	0.187	0.183	0.185	0.153	0.183	0.187	0.200	0.189	0.201	0.175	0.196	0.193	0.183	0.189	0.198	0.181	0.182	0.174	0.166	0.178	0.165	1.000	0.193	0.197	0.191	0.177
Support	S1	0.227	0.202	0.229	0.224	0.209	0.211	0.214	0.198	0.216	0.179	0.208	0.189	0.226	0.209	0.225	0.205	0.220	0.202	0.199	0.197	0.206	0.184	0.197	0.182	0.180	0.187	0.175	0.193	1.000	0.231	0.211	0.212
	S2	0.226	0.229	0.236	0.221	0.205	0.214	0.229	0.202	0.225	0.179	0.200	0.188	0.248	0.212	0.234	0.196	0.245	0.208	0.213	0.182	0.203	0.185	0.200	0.222	0.192	0.171	0.187	0.197	0.231	1.000	0.224	0.208
	S3	0.227	0.199	0.251	0.213	0.208	0.211	0.209	0.191	0.203	0.174	0.201	0.175	0.227	0.195	0.219	0.176	0.211	0.189	0.201	0.196	0.195	0.160	0.200	0.202	0.175	0.182	0.197	0.191	0.211	0.224	1.000	0.196
	S4	0.207	0.178	0.216	0.216	0.185	0.184	0.195	0.188	0.189	0.158	0.193	0.197	0.186	0.178	0.224	0.196	0.194	0.190	0.184	0.190	0.199	0.187	0.181	0.189	0.181	0.177	0.176	0.177	0.212	0.208	0.196	1.000

Figure 11: The Jaccard similarity results between research fields

In Figure 11, grey-filled items represent relations that have relatively high similarity values compared to others. Regarding the intra-relations of research topics belonging to the same category, we can identify high similarities between topics within the technique category and the support category. However, compared to these, the research topics in the application category have weak connectivity among each other. That is, applications in different domains apply different goals, techniques, and tools, leading to less intersection with each other.

Regarding the inter-relations of research fields between the different categories, we can identify that the topics in the technique and support categories have high similarities. On the contrary, it is hard to find meaningful relations between the application and other categories. This is because the extracted keywords for the application category are mainly relevant to a specific domain.

Next, we focus on the similarities between the individual research fields in the technique category. First, there is a strong relationship between process discovery (T1) and other fields, such as conformance checking (T2), performance-driven process mining (T3), process mining framework, guideline, & tool (T13), event logs (T14), and others (T17). This indicates that process discovery has played a pivotal role in process mining, linked to other research topics such as conformance checking, event logs, and performance-driven techniques. Also, process mining framework, guideline, and tool (T13) has high similarity values with not only process discovery (T1) but also conformance checking (T2), semantic process mining (T9), and process mining visualization (T15). Thus, various technical algorithms have been

developed to utilize a process mining framework or tool as a standard. In addition, it is noteworthy that there is a high similarity between trace clustering (T7) and variant analysis (T11) since both topics pursue the similar goal of analyzing processes at a fine-grained level.

In the application category, there is no notable identification of the relationships, whereas the research fields in the support show high similarities with multiple topics in the technique category. More in specific, a couple of research topics, e.g., machine learning & business intelligence (S1) and BPM & BPR (S2), show a high similarity with the technical algorithms, e.g., process discovery, organization mining, and process mining visualization. Also, BPM & BPR (S2) and process simulation (S3) have been studied with performance-driven process mining, indicating that it plays an essential role in implementing business processes.

5.2. Network derivation and community detection

Based on the results obtained in previous sections, we first discover a network representing relations of research fields. Figure 12(a) presents the derived network.

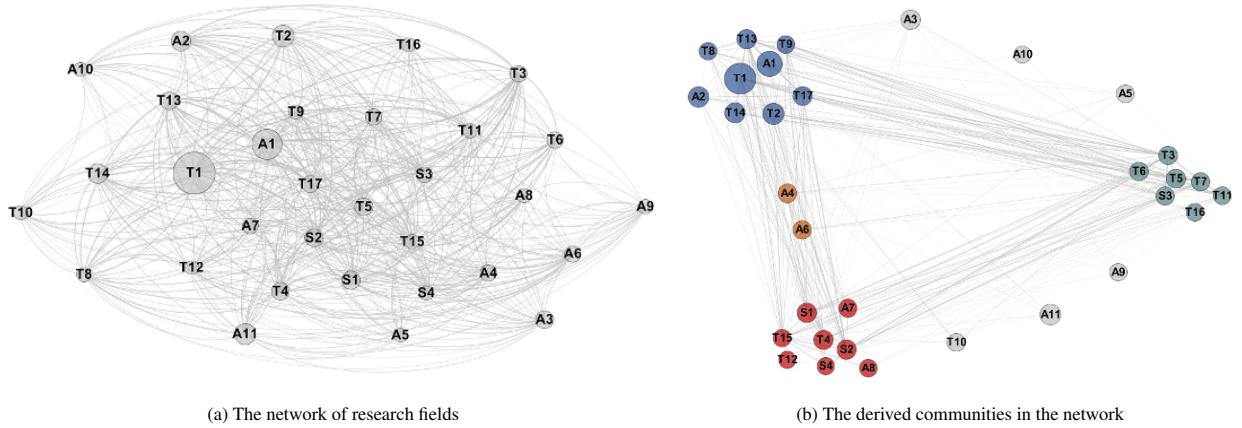


Figure 12: The results for network derivation and community detection

In the network, nodes indicate the research fields identified in Section 3, whereas edges indicate the similarity between the research fields measured in Subsection 5.1. Also, the node size indicates the number of publications belonging to each research field, while the thickness of the edge represents the degree of similarity. Based on the derived network, we can recognize that many studies have focused on process discovery (T1), conformance checking (T2), healthcare (A1), and software (A2) topics. Also, process discovery (T1) and process mining framework, guideline, and tool (T13) have been the central points to connect other research fields, as discussed in Subsection 5.1.

Next, we derive research communities by applying the Louvain algorithm on the basis of the network. Figure 12(b) depicts the derived communities. Applying the modularity delta threshold of 2.21 in discovering communities, we found four groups, e.g., representative, organizational, performance-centric, and manufacturing & IoT communities, and six individual topics.

- **Representative community** is composed of process discovery (T1), conformance checking (T2), streaming process mining (T8), semantic process mining (T9), process mining framework, guideline, & tool (T13), event logs (T14), others (T17), healthcare (A1), and software (A2). This community includes process discovery, conformance checking, healthcare, and software research fields where active research has been steadily conducted since the past. In addition, research on event logs and process mining frameworks, guidelines & tools are included in this community. Also, we can identify a couple of cutting-edge process mining techniques, including semantic process mining and streaming process mining. These are emerging research fields derived from process discovery and event log topics. Based on these results, we understand that this community represents process mining and consists of the most fundamental components of this discipline.
- **Resource-oriented community** is composed of organizational mining (T4), privacy-preserving process mining (T12), process mining visualization (T15), web and cloud services (A7), e-business (A8), machine learning and business intelligence (S1), Business process management and business process redesign (S2), and decision support (S4). This community represents the research topics from the business process management perspective, centered on resources and organizations. Specifically, it includes organizational mining and process mining visualization, e.g., social networks, in the technique category. In addition, a couple of support research fields, e.g., business intelligence and BPM, belong to this community, where resource-oriented data are actively utilized. Also, this community holds the web service and E-business fields in the application area.
- **Performance-oriented community** consists of performance-driven process mining (T3), process monitoring (T5), predictive process monitoring (T6), trace clustering (T7), variant analysis (T11), process model comparison (T16), and process simulation (S3). Unlike the prior two communities, the technical research topics comprise most of this community. In particular, performance-driven process mining plays a key role, and other relevant research topics are connected to this. For example, a series of connections of similar technologies, such as performance-driven process mining, process monitoring, and predictive process monitoring, show changes in technological trends and advancements. In addition, trace clustering, variant analysis, and process model comparison technologies have mutual similarities. In more detail, sub-event logs derived through trace clustering have been utilized as input for the variant analysis and process model comparison. Finally, process simulation aligns with process mining-based performance measurement and scenario analysis.
- **Manufacturing & IoT community** involves manufacturing (A4) and internet of things (A6). This community implies that sensor-based event logs in various domains, especially manufacturing environments, have been collected, and numerous case studies based on these have been conducted to analyze the end-to-end manufacturing processes.

Also, six individuals are not included in the communities: process model repair (T10), education (A3), manufacturing (A5), customer support services (A9), protocols (A10), and others (A11). Most of these research fields belong

to the application category, and they selectively utilize different algorithms to resolve problems for a specific domain. Also, the process model repair is independent of the communities because it has only been conducted recently and is slightly linked to many other topics.

6. Forecasting future trends of research fields in process mining

This section aims to analyze future trends for the research topics. To that end, we predict the number of publications in a research field for the upcoming years (e.g., the number of publications in T1, i.e., process discovery, for 2023) by leveraging 1) the historical publication records of the research field (e.g., the number of T1’s publications for the last three years) and 2) the historical records of other research fields that are highly correlated to the research field (e.g., the number of publications for conformance checking and performance-driven process mining).

6.1. Prediction Model

As presented in Subsection 2.5, we use LSTM models due to their robustness to learning temporal dependencies [39]. In the following, we describe the training procedure and validations of the model.

The LSTM model consists of 2 layers where the number of hidden states in the first layer is 2048 and the number of hidden states in the second layer is 4096. We train the model using the training data over 14 years with Adam optimizer algorithm [40] such that RMSE(Root Mean Squared Error) and MAE(Mean Absolute Error) between the predicted value $\hat{y}$ and the actual value y are minimized [41]. RMSE represents the root mean of squared errors (i.e., $(\hat{y} - y)^2$) as below:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (5)$$

MAE represents the mean of absolute errors (i.e., $|\hat{y} - y|$) as below:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (6)$$

To improve model generalizability, we use Dropout [42] and Batch Normalization [43].

To evaluate the performance of the trained model, we adopt time-series cross validation [44] since k-fold validation has limitations in validating time series data (e.g., due to its random selection of independent data points). We use the test data over four years. Table 3 shows the RMSE and MAE resulting from the time-series cross-validation.

Table 3: RMSE and MAE of the trained model that results from the time-series cross-validation

RMSE	MAE
5.63 ± 1.02	3.98 ± 0.55

6.2. Prediction Results

Based on the prediction model, we predict the number of publications in all the identified research fields for the upcoming two years. To that end, we first predict the number of publications in 2022 using the historical records and then predict the number of publications in 2023 using the predicted record in 2022 as well as the historical records.

Figure 13 shows the prediction results for the research fields in the technique category. For instance, the left-top chart depicts the predicted number of publications for T1 (process discovery) in 2022 and 2023, as well as the historical records from 2019 to 2020. The number of publications in 2019 is 29, decreasing to 22 in 2020 and increasing to 26 in 2021. The predicted number of publications in 2022 is around 29, and the one in 2023 is around 31, showing an increasing trend in the research field.

Most of the research fields in the technique category are forecasted to be strengthened in 2022 compared to 2021. In particular, T2 (conformance checking), T3 (performance-driven process mining), T5 (process monitoring), and T8 (streaming process mining) are predicted to show tremendous increases in the number of publications in 2022. However, a few research fields are anticipated to show decreasing advances with fewer publications, e.g., T11 (variant analysis) and T13 (process mining framework, guideline, & tool).

In 2023, most of the research fields are predicted to have fewer publications than in 2022. For instance, T2 (conformance checking), T3 (performance-driven process mining), T5 (process monitoring), T7 (trace clustering), and T8 (streaming process mining) are anticipated to have fewer publications in 2023 compared to 2022. However, such research fields still show the predicted number of publications higher than the average number of publications in the last three years, demonstrating the ongoing expansion of the research fields. For instance, the average number of publications of T2 from 2020 to 2022 is around 13, whereas the predicted number of publications in 2023 is above 14.

Figure 14 shows the prediction results for the research fields that belong to the application category. For instance, the chart on the left-top corner describes the predicted number of publications for A1 (healthcare) in 2022 and 2023, as well as the historical records of the last three years. The number of publications in 2019, 2020, and 2021 is 35, 52, and 19, respectively. The predicted number of publications in 2022 and 2023 is around 35 and 34, respectively.

Similar to the technique category, the research fields in the application category are anticipated to expand in terms of the number of publications in 2021 compared to the previous year. For instance, the number of publications in A1 (healthcare) is forecasted to increase from 19 in 2021 to around 35 in 2022, and the number of publications in A3 (education) is anticipated to increase from 10 to 12. The predicted number of publications in 2023 remains similar to 2022, only showing slight increases and decreases depending on research fields. The number of publications for A1 is predicted to decrease marginally in 2023 compared to 2022, whereas the predicted number of publications for A2 is slightly higher than the one in 2022.

Figure 15 describes the predicted number of publications for the research fields in the support category. For instance, the left-top chart shows the number of publications for S1 (machine learning and business intelligence) in

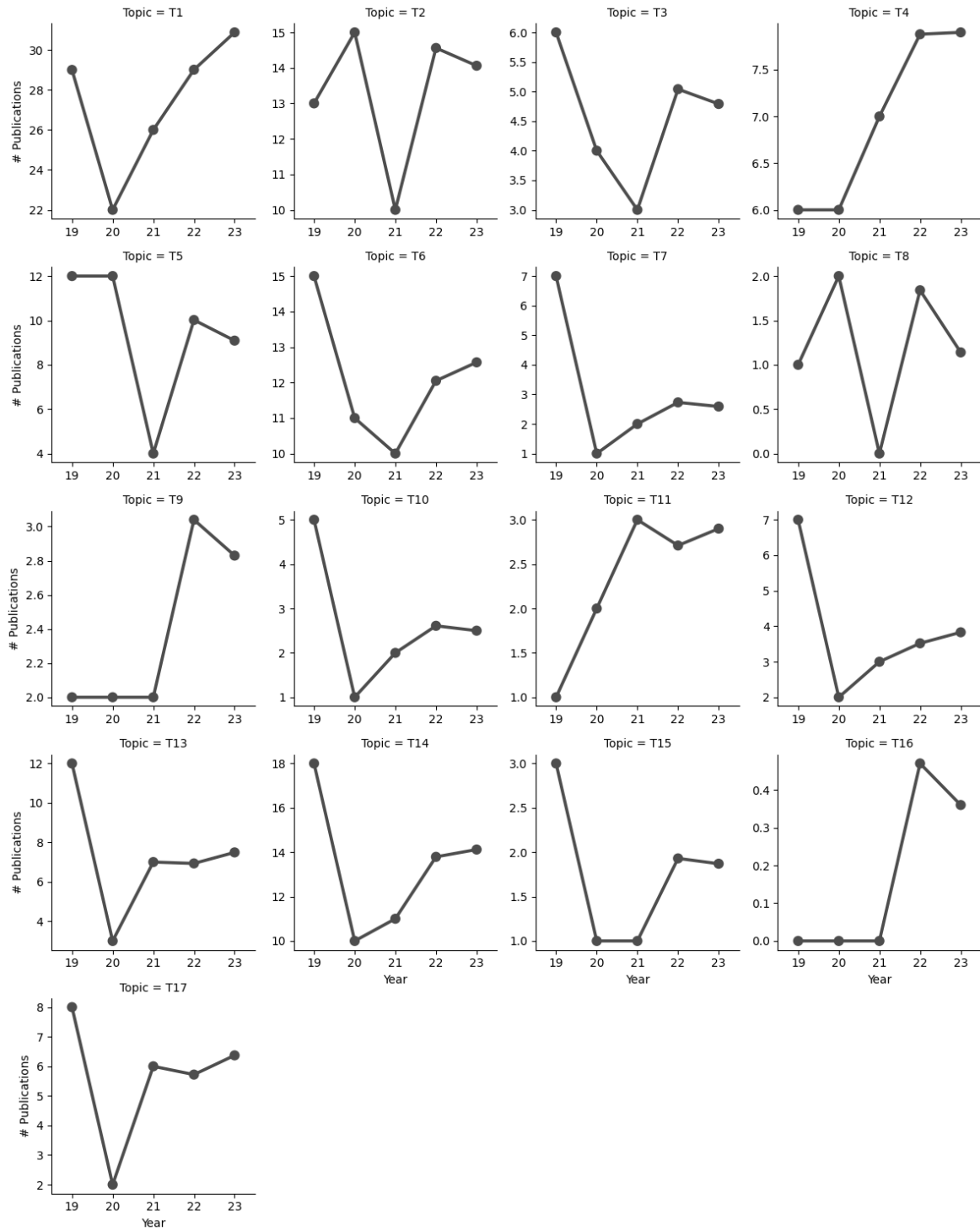


Figure 13: The predicted number of publications in 2022 and 2023 for the research fields of the technique category, along with the historical number of publications from 2019 to 2021

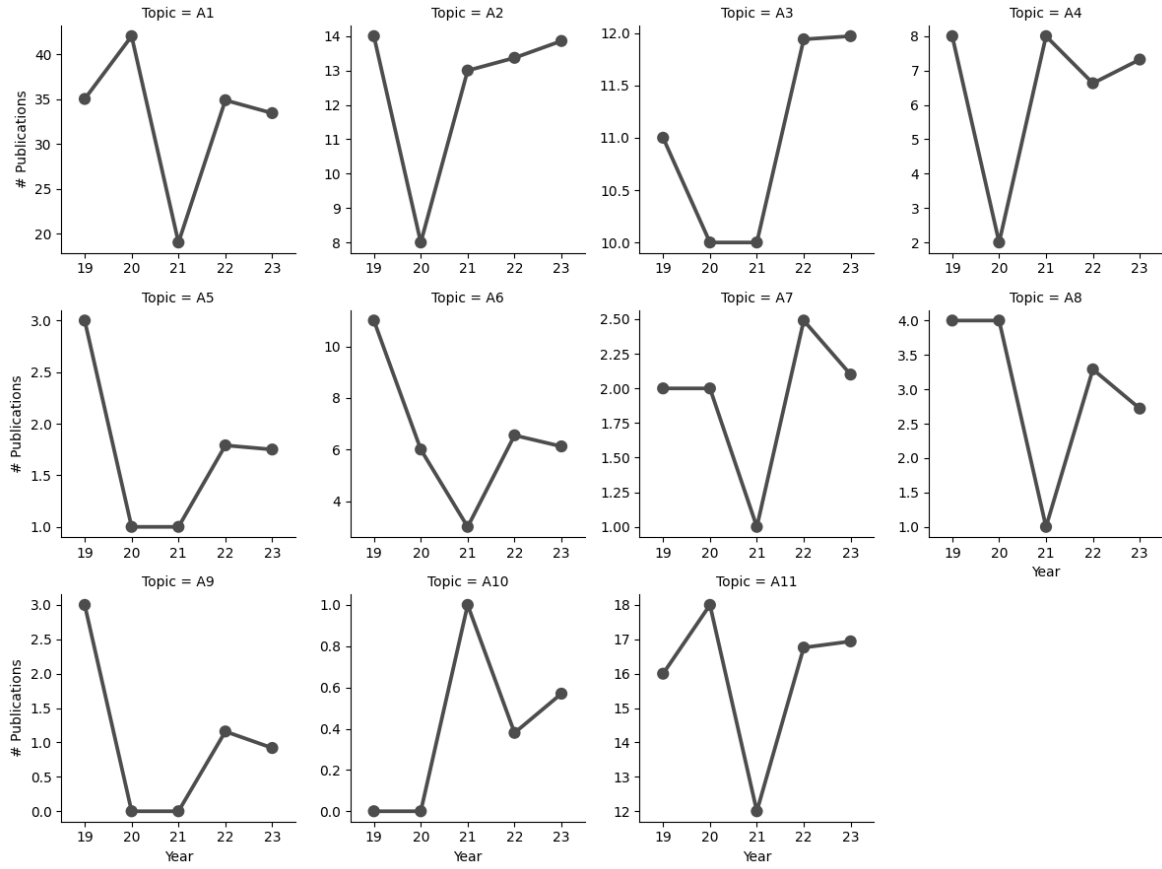


Figure 14: The predicted number of publications in 2022 and 2023 for the research fields of the application category, along with the historical number of publications from 2019 to 2021

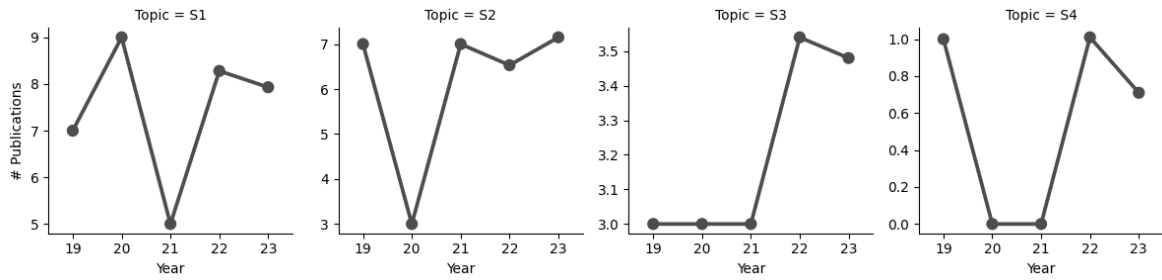


Figure 15: The predicted number of publications in 2022 and 2023 for the research fields of the support category, along with the historical number of publications from 2019 to 2021

2022 and 2023, along with the historical records of the last three years. The number of publications in 2019, 2020, and 2021 is 7, 9, and 5, respectively. The predicted number of publications in 2022 and 2023 is around 8 in both years.

The number of publications in S1 (machine learning and business intelligence) and S2 (business process management and business process reengineering) is predicted to increase in 2022 and remain similar in 2023. Although

the predicted number of publications in S3 (process simulation) and S4 (decision support) shows similar trends, the absolute number of publications remains nearly unchanged; the numbers of publications in S3 and S4 are predicted to be around 3 and 1, respectively.

7. Discussion

This section aims to answer the research questions raised in Section 1 by discussing the results presented in the earlier sections. Moreover, we compare the obtained knowledge with the existing trends in process mining both from academic and practical standpoints.

7.1. Answers to the research questions

Regarding the first research question, we identified and investigated research topics in process mining with a series of approaches, as presented in Section 3 and Section 4. The research topics in the technique, application, and support categories have been classified into the dominant, skyrocketing, and deep-cut based on history. Process discovery and conformance checking steadily and dominantly have occupied a large part in the process mining discipline. Also, popular domains in applying process mining have been healthcare, software, education, and manufacturing from the past to the present. It is noteworthy that some research fields, e.g., predictive process monitoring, concept drift, and privacy-preserving process mining, have emerged rapidly in recent years and are highly related to machine learning and big data. On the contrary, process model comparison techniques and protocol applications have not gained much attention in the process mining community. In addition, we should pay attention to the research fields on other techniques (T17) and other applications (A11), which include innovative research papers with novel ideas on emergent topics in process mining. For example, there have been articles in T17 about process mining connecting to key technologies, such as natural language processing, big data, and blockchains. Until now, only a couple of papers have been published for each topic, but they may constitute a new process mining research field in the next few years.

To resolve the second research question, we discovered the topic network and detected the communities of research fields in process mining, as presented in Section 5. As a result, process discovery has been placed at the central point in the network, which indicates that it has strong relationships with other topics. Besides, the research fields in the application category have much less resemblance with other topics compared to the fields in the technique and support. With the derived communities, we uncovered how the research fields have interacted with each other and how they have developed. The representative community demonstrated which process mining techniques have mainly integrated for process mining projects in most domains. In addition, regarding the performance-related community, we observed that performance-driven process mining had become the cornerstone for the emergence of the latest monitoring techniques, such as process monitoring and predictive process monitoring. The connection network between research topics covered in this study and the community derived through it has explained the emergence of new research fields from the existing fields. In other words, it is necessary to explore the possibility of deriving a new research field within process mining by utilizing a predictive network model such as a graph neural network.

Regarding the third question, the prediction model suggested that the topics in most of the research fields will be expanded in 2022 and 2023 with respect to the number of their publications. Especially most of the research fields will show quantum leaps in the number of publications in 2022 compared to 2021 when the number of publications was significantly lower than in the previous years. In 2023, the number of publications will remain mostly the same as in 2022, with different topics showing non-significant increases or decreases. However, a few research fields will show fewer publications in 2022 and 2023, such as T11 (variant analysis) and T13 (process mining framework, guideline & tool). Due to the changes in trends from the emergence of new techniques and the connectivity with other disciplines, it is hard to get accurate values for the next few years. However, our work contributes to understanding future technology trends and development potential.

7.2. Further discussions

In this subsection, we further discuss the obtained research fields and their future trends in relation to the existing trends in process mining, both from academic and practical standpoints. First, to relate the existing academic trends, we analyze the workshop topics in the renowned conferences presented in Section 4 such as ICPM, BPM, and CAiSE. As of 2022, the workshops held with ICPM, BPM, and CAiSE conferences are as follows.

- Business Process Intelligence (BPI)
- DEClarative DECision and Hybrid approaches to processes (DEC2H)
- Streaming Analytics for Process Mining (SA4PM)
- Quantitative Aspects of Process Mining (QAPM)
- Social and Human Aspects of Business Process Management (BPMS2)
- Leveraging Machine Learning in Process Mining (ML4PM)
- Artificial Intelligence for Business Process Management (AI4BPM)
- Responsible Process Mining (RPM)
- Event Data and Behavioral Analytics (EdbA)
- Process Querying, Manipulation, and Intelligence (PQMI)
- Data Quality and Transformation in Process Mining (DQT-PM)
- Process-Oriented Data Science for Healthcare (PODS4H)
- Education meets Process Mining (EduPM)
- Process Management in Digital Production (PMDiPro)

- Business Processes Meet Internet-of-Things (BP-meet-IoT)

BPI has covered the entire research area of process mining, including process discovery, conformance checking, etc. The other workshops are mapped to specific research fields that are derived from this study. DEC2H, which covers declarative process mining, belongs to a subtopic of process discovery (T1). Articles on organizational mining (T4), streaming process mining (T8), and privacy-preserving process mining (T12) has been published through BPMS2, SA4PM, and RPM, respectively. In addition, the performance-oriented community, consisting of performance-related topics such as performance-driven process mining (T3), process monitoring (T5), and predictive process monitoring (T6), has published articles in QAPM. As performance-driven prediction has become essential, there has been in-depth attention to workshops covering machine learning and artificial intelligence, such as ML4PM and AI4BPM. Moreover, research on process model repair (T10) and event logs (T14) have been discussed in PQMI, EdbA, and DQT-PM, respectively. There have also been some workshops focused on specific applications domains, including PODS4H (healthcare (A1)), Edu4PM (education (A3)), PMDiPro (manufacturing (A4)), and BP-meet-IoT (IoT (A6)).

Second, to relate the existing practical trends in process mining, we analyze the functionalities supported by the renowned process mining tools that are evaluated as leaders and star performers in the process mining market by the *Everest group report* [45]: *Celonis*, *Software AG*, *Minit*, and *UiPath*. Moreover, we analyze the functionalities provided by the most representative open-source tool in process mining, *ProM*; The tool is contributed by process mining researchers, and most of the techniques developed so far have been implemented in the tool. Furthermore, we include *Fluxicon Disco*, which continuously received much attention for its compact and practical functionalities.

Table 4 presents whether or not the tools provide the functionalities corresponding to the research fields presented in this paper. It is noticeable that all tools cover process discovery (T1) and performance-driven process mining (T13). In other words, process discovery is the most fundamental topic and technique in process mining, and it is worthwhile to present performance measurements on the discovered process model. In addition, conformance checking (T2) and process model comparison (T16) are the functions all tools provide except for Disco. Although process model comparison is classified as deep-cut, it is a practically relevant technique. Moreover, organizational mining (T4) is only covered by Celonis and Minit, which indicates that the control-flow perspective is more influential than the resource perspective. Also, most tools provide process monitoring (T5), e.g., anomaly detection with statistical analysis, whereas they have not yet implemented predictive process monitoring (T6) that requires machine learning and deep learning. Also, relatively new algorithms, such as streaming process mining (T8), semantic process mining (T9), and process model repair (T10), are yet to be implemented in most tools except Celonis. Likewise, process mining commercial tools still do not cover privacy-preserving process mining (T12).

8. Conclusion

In this paper, we analyzed the literature on process mining using text mining and machine learning techniques. First, we collected 2,207 process mining articles and conference proceedings from 2002 to 2021. Next, we analyzed

Table 4: Key functionalities of process mining tools

	ProM	Celonis	Software AG	MINIT	UIPath	DISCO
T1	✓	✓	✓	✓	✓	✓
T2	✓	✓	✓	✓	✓	
T3	✓	✓	✓	✓	✓	✓
T4	✓	✓		✓		
T5	✓	✓	✓	✓		
T6	✓	✓				
T7	✓					
T8	✓	✓				
T9	✓	✓				
T10	✓	✓				
T11	✓	✓	✓	✓	✓	✓
T12	✓					
T13	✓	✓	✓	✓	✓	✓
T14	✓	✓	✓	✓	✓	✓
T15	✓	✓	✓	✓	✓	✓
T16	✓	✓	✓	✓	✓	

the with a pre-trained model-based topic modeling, exploratory data analysis with key features, network and community detection, and future prediction with a deep learning model. As a result, we provided knowledge on the main research topics in process mining and their temporal trends, the relationships between the research topics, and the predictions on the expansion of the topics. Moreover, we compared the provided knowledge with the trends in both academics and practices.

Our work has several limitations. First, the presented quantitative analysis based on text mining and machine learning techniques still require some manual steps. For instance, the pre-processing of collected documents is manually conducted, while the identification of research fields requires qualitative evaluations based on expert knowledge. Second, the presented research topics are by no means exhaustive. We only focus on the topics identified by text mining algorithms, and thus, some of the essential topics may not be a part of the further analysis, e.g., network detection, community detection, yearly patterns, etc. For instance, object-centric process mining receives much attention from the community as an essential research topic, but it is not part of our analysis since it is not identified as a research topic in the topic modeling algorithm. Finally, since our analysis relies on historical data, we cannot present topics that do not exist in history. The discovery and analysis of newly-emerging topics remain as the judgments of readers.

Author contribution statement

Gyunam Park: Conceptualization, Methodology, Software, Writing - original draft. **Jiyeon Lee:** Data curation, Software, Investigation. **Minsu Cho:** Conceptualization, Methodology, Supervision, Writing – original draft, review & editing.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by Ministry of Science and ICT (No. NRF-2021R1G1A1094019) and the Research Grant of Kwangwoon University in 2021.

References

- [1] I. M. Sebastian, J. W. Ross, C. Beath, M. Mocker, K. G. Moloney, N. O. Fonstad, How big old companies navigate digital transformation, *MIS Quarterly Executive* 16 (3) (2017) 6.
- [2] C. Matt, T. Hess, A. Benlian, Digital transformation strategies, *Business & information systems engineering* 57 (5) (2015) 339–343.
- [3] P. C. Verhoef, T. Broekhuizen, Y. Bart, A. Bhattacharya, J. Q. Dong, N. Fabian, M. Haenlein, Digital transformation: A multidisciplinary reflection and research agenda, *Journal of Business Research* 122 (2021) 889–901.
- [4] J. vom Brocke, M. Jans, J. Mendling, H. A. Reijers, A five-level framework for research on process mining, *Business & Information Systems Engineering* 63 (5) (2021) 483–490.
- [5] T. Grisold, J. Mendling, M. Otto, J. vom Brocke, Adoption, use and management of process mining in practice, *Business Process Management Journal* (2020).
- [6] L. Reinkemeyer, *Process mining in action, Process Mining in Action Principles, Use Cases and Outlook* (2020).
- [7] W. M. P. van der Aalst, *Process Mining - Data Science in Action, Second Edition*, Springer, 2016.
- [8] A. Augusto, R. Conforti, M. Dumas, M. La Rosa, F. M. Maggi, A. Marrella, M. Mecella, A. Soo, Automated discovery of process models from event logs: review and benchmark, *IEEE transactions on knowledge and data engineering* 31 (4) (2018) 686–705.
- [9] A. Rozinat, W. M. P. van der Aalst, Conformance checking of processes based on monitoring real behavior, *Information Systems* 33 (1) (2008) 64–95.
- [10] W. M. P. van der Aalst, M. H. Schonenberg, M. Song, Time prediction based on process mining, *Information systems* 36 (2) (2011) 450–475.
- [11] I. Teinmaa, M. Dumas, M. L. Rosa, F. M. Maggi, Outcome-oriented predictive process monitoring: review and benchmark, *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13 (2) (2019) 1–57.
- [12] W. M. P. van der Aalst, H. A. Reijers, A. J. Weijters, B. F. van Dongen, A. A. De Medeiros, M. Song, H. Verbeek, Business process mining: An industrial application, *Information Systems* 32 (5) (2007) 713–732.
- [13] W. M. P. van der Aalst, J. Carmona, *Process mining handbook* (2022).
- [14] A. Rozinat, R. S. Mans, M. Song, W. M. P. van der Aalst, Discovering simulation models, *Information systems* 34 (3) (2009) 305–327.
- [15] M. Cho, M. Song, M. Comuzzi, S. Yoo, Evaluating the effect of best practices for business process redesign: An evidence-based approach based on process mining techniques, *Decision Support Systems* 104 (2017) 92–103.
- [16] V. Naderifar, S. Sahran, Z. Shukur, A review on conformance checking technique for the evaluation of process mining algorithms, *TEM Journal* 8 (4) (2019) 1232.
- [17] S. J. van Zelst, F. Mannhardt, M. de Leoni, A. Koschmider, Event abstraction in process mining: literature review and taxonomy, *Granular Computing* 6 (3) (2021) 719–736.
- [18] D. M. V. Sato, S. C. De Freitas, J. P. Barddal, E. E. Scalabrin, A survey on concept drift in process mining, *ACM Computing Surveys (CSUR)* 54 (9) (2021) 1–38.

- [19] E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, Process mining in healthcare: A literature review, *Journal of biomedical informatics* 61 (2016) 224–236.
- [20] A. Bogarín, R. Cerezo, C. Romero, A survey on educational process mining, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8 (1) (2018) e1230.
- [21] A. Corallo, M. Lazoi, F. Striani, Process mining and industrial applications: A systematic literature review, *Knowledge and Process Management* 27 (3) (2020) 225–233.
- [22] P. Zerbino, A. Stefanini, D. Aloini, Process science in action: A literature review on process mining in business management, *Technological Forecasting and Social Change* 172 (2021) 121021.
- [23] S. Kim, H. Park, J. Lee, Word2vec-based latent semantic analysis (w2v-lsa) for topic modeling: A study on blockchain technology trend analysis, *Expert Systems with Applications* 152 (2020) 113401.
- [24] T. K. Landauer, P. W. Foltz, D. Laham, An introduction to latent semantic analysis, *Discourse processes* 25 (2-3) (1998) 259–284.
- [25] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *Journal of machine Learning research* 3 (Jan) (2003) 993–1022.
- [26] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, *arXiv preprint arXiv:2203.05794* (2022).
- [27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [28] G. T. Reddy, M. P. K. Reddy, K. Lakshmana, R. Kaluri, D. S. Rajput, G. Srivastava, T. Baker, Analysis of dimensionality reduction techniques on big data, *IEEE Access* 8 (2020) 54776–54788.
- [29] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, C.-T. Lin, A review of clustering techniques and developments, *Neurocomputing* 267 (2017) 664–681.
- [30] J. Scott, *What is social network analysis?*, Bloomsbury Academic, 2012.
- [31] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! keyword extraction from single documents using multiple local features, *Information Sciences* 509 (2020) 257–289.
- [32] S. R. El-Beltagy, A. Rafea, Kp-miner: A keyphrase extraction system for english and arabic documents, *Information systems* 34 (1) (2009) 132–144.
- [33] B. Škrlj, A. Repar, S. Pollak, Rakun: Rank-based keyword extraction via unsupervised learning and meta vertex aggregation, in: *International Conference on Statistical Language and Speech Processing*, Springer, 2019, pp. 311–323.
- [34] G. Erkan, D. R. Radev, Lexrank: Graph-based lexical centrality as salience in text summarization, *Journal of artificial intelligence research* 22 (2004) 457–479.
- [35] Y. Sun, H. Qiu, Y. Zheng, Z. Wang, C. Zhang, Sifrank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model, *IEEE Access* 8 (2020) 10896–10906.
- [36] S. Bag, S. K. Kumar, M. K. Tiwari, An efficient recommendation generation using relevant jaccard similarity, *Information Sciences* 483 (2019) 53–64.
- [37] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, *Journal of statistical mechanics: theory and experiment* 2008 (10) (2008) P10008.
- [38] P. De Meo, E. Ferrara, G. Fiumara, A. Provetti, Generalized louvain method for community detection in large networks, in: *2011 11th international conference on intelligent systems design and applications*, IEEE, 2011, pp. 88–93.
- [39] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: Lstm cells and network architectures, *Neural computation* 31 (7) (2019) 1235–1270.
- [40] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
- [41] C. J. Willmott, K. Matsuura, Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, *Climate research* 30 (1) (2005) 79–82.
- [42] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* 15 (1) (2014) 1929–1958.

- [43] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F. R. Bach, D. M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, Vol. 37 of JMLR Workshop and Conference Proceedings, JMLR.org, 2015, pp. 448–456.
- [44] C. Bergmeir, J. M. Benítez, On the use of cross-validation for time series predictor evaluation, *Inf. Sci.* 191 (2012) 192–213.
- [45] A. Modi, H. K. Makan, S. Kumar, A. Nath, Process mining – technology provider landscape with products peak matrix assessment 2022, Tech. rep., Everest Group (2022).

Gyunam Park



Gyunam Park received the M.S. degree in the Department of Industrial & Management Engineering at POSTECH (Pohang University of Science and Technology), Pohang, South Korea, in 2019. He is currently a Ph.D. candidate in the Department of Computer Science, RWTH Aachen University, Aachen, Germany. His research interest includes process mining, data science, online operational support, machine learning, and deep learning. He has published scientific papers in journals such as Decision Support Systems, IEEE Transactions on Semiconductor Manufacturing, and Progress In Artificial Intelligence and conferences including International Conference on Business Process Management, International Conference on Conceptual Modeling, International Conference on Service-Oriented Computing, and International Conference on Process Mining.

Jiyeon Lee



Jiyeon Lee is currently a B.S. candidate in the School of Information Convergence at Kwangwoon University, Seoul, South Korea. Her research interest includes process mining, text mining, data science, deep learning, and recommendation systems.

Minsu Cho



Minsu Cho received his Ph.D. degree in the Department of Management Engineering at UNIST (Ulsan National Institute of Science and Technology), Ulsan, South Korea, in 2018. He is currently an assistant professor at the School of Information Convergence at Kwangwoon University. Prior to this, he was a post-doctoral researcher at POSTECH (Pohang University of Science and Technology) and a senior researcher at KITECH (Korea Institute of Science and Technology), respectively. His

research interest includes process mining, business analytics, text mining, data science, machine learning, deep learning, and technological forecasting. He has published scientific papers in journals such as Decision Support Systems, International Journal of Medical Informatics, IEEE Transactions on Semiconductor Manufacturing, International Journal of Environmental Research and Public Health, IEEE Access, and others.

Highlights

Toward understanding and forecasting the research trends in process mining: An analysis using text mining and machine learning

- Process mining research has been conducted into three research branches: technique, application, and support aspects.
- Process discovery and conformance checking dominantly have occupied a large part, while predictive process monitoring, concept drift, and privacy-preserving process mining belong to skyrocketing groups.
- Most of the research fields will have a trend of quantum leaps in the number of publications in a few years.

Toward understanding and forecasting the research trends in process mining: An analysis using text mining and machine learning

Abstract

Process mining is a relatively new discipline focusing on gaining process-centric knowledge from event logs collected by enterprise systems. Various techniques have been developed to automatically discover process models, analyze the compliance of real-life processes to the process models, predict operational frictions, recommend possible actions to mitigate emerging risks, etc. In recent years, the techniques have been adopted in various industries, such as healthcare, manufacturing, logistics, and finance. Now there are over 40 commercial products based on process mining techniques, such as Celonis, SAP Signavio, Minit (Microsoft), MyInvenio (IBM), and open-source tools such as ProM, PM4Py, and bupaR. In this work, we analyze the process mining literature in depth using text mining and machine learning techniques. More in detail, we 1) analyze the main research topics in process mining and their trends, 2) investigate the relationship between the topics, and 3) predict the expansion of the topics in the near future. To that end, we analyzed 2,207 process mining articles from 2002 to 2021 using a range of techniques such as topic modeling, exploratory data analysis, network and community detection, and future prediction.

Keywords: process mining, text mining, network analysis, literature reviews, emerging trends, predictions

1. Introduction

The development of SMACIT (i.e., Social, Mobile, Analytics, Cloud, and Internet of Things) and other digital technologies have enabled the growth of new organizational values with the aim of digital transformation [1, 2, 3]. In particular, there have been many efforts to identify and improve operational processes based on IT-enabled technologies, and the increase in the availability of data has expedited it [4, 5]. Process mining has emerged in such an environment and played a crucial role in implementing digital transformation [5, 6].

Process mining is a relatively young discipline focused on deriving process-oriented knowledge from event logs collected by enterprise systems [7]. The discipline leverages process science, a family of methods for improving operational processes, including business process management and simulation, and data science deriving value from data, such as machine learning and artificial intelligence [7].

Under the circumstances, there have been many approaches to three research branches: (i) technical, (ii) application, and (iii) supportive aspects. Regarding the technical viewpoint, many efforts have been devoted to developing process discovery techniques [8] that automatically produce process models from event logs without prior knowledge and presenting innovative works on conformance checking [9], process monitoring [10], and predictions [11]. From

the view of the application, i.e., the business perspective, this discipline has provided availability to be applied in multiple domains, e.g., healthcare, education, services, and manufacturing, and the ability to create a new implication [12, 13, 6]. Also, from the supportive aspect, process mining has been utilized to enhance efficiency and improve the accuracy of other disciplines, such as simulation, business process management, and machine learning [14, 7, 15].

Such interests in process mining and researchers' efforts have guided the advancement and expansion of this research field. A process mining research community, i.e., the IEEE task force on process mining, has emerged, and much research has been published in the independent international conference on process mining series. This well-marked progress requires an in-depth understanding of the research scopes, specific research topics, fashionable techniques, relationships of research topics, and future trends of process mining. Therefore, it requires a comprehensive review to understand the discipline in essence.

There have been several comprehensive reviews on process mining [8, 16, 17, 18, 19, 20, 21, 22]. Corallo et al. [21] employ a systematic mapping method to describe research matters, techniques, and applications in process mining. Zerbino et al. [22] reconcile process mining research based on the multiple viewpoints of contributions, algorithms, perspectives, sectors, business, and management. In addition, there is research work to detail a specific topic, including process discovery [8], conformance checking [16], event abstraction [17], and concept drift [18], while others have handled the applications for a specific context, such as healthcare [19] and education [20].

However, the existing studies employ a qualitative review by a group of experts centered on the authors. Although knowledge-based reviews by experts have produced mature and worthwhile results, they have some disadvantages in that it is based on subjective judgments and discussions. As a result, bias or preconceived notions may arise in recognizing research topics and identifying meaningful trends, and there may be limitations in predicting the potential for future technology development. As a means of overcoming these limitations, this study attempts to analyze the process mining literature in depth using text mining and machine learning techniques. That is, we propose a quantitative literature review method. To the best of our knowledge, this study is the first case of discovering new values and knowledge by directly utilizing process mining-related literature as data.

In this paper, we prepared the following research questions to clarify the purpose and direction of data analysis.

- RQ1. What are the main research topics of process mining, and how has the trend changed over time?
- RQ2. What process mining techniques are mutually cooperative, and how does each research topic relate to the other?
- RQ3. How much will process mining evolve and expand in the near future, and what research topics will be the focus?

To resolve these research questions, we collected 2,207 process mining articles and conference proceedings from 2002 to 2021. Such documents have been explored with the pre-trained model-based topic modeling, exploratory data

analysis with key features, network and community detection, and future prediction for research trends with a deep learning model. This research has the following contributions:

- More straightforward results, e.g., research topics and trends, are presented through quantitative analysis, such as text mining and machine learning.
- Rather than simply enumerating the process mining review results, we describe in-depth analysis outcomes, such as discovering relationships between research topics and identifying research themes.
- Finally, we suggest a method for identifying the potential of process mining and each specific research topic through a time-series forecasting model based on historical data.

This paper is organized as follows. A series of methods for deriving process mining research topics and meaningful bibliometric results are provided in Section 2. Section 3 presents topic modeling results and describes the details of process mining research fields. Then, Section 4 investigates the discovered research fields with the common aspects of publications. Section 5 explains the interrelationships between research topics, and Section 6 identifies the future of process mining. Section 7 discusses the results, and Section 8 concludes this paper by preparing future works.

2. Method

This section proposes a quantitative method for understanding the process mining discipline by analyzing research fields and trends based on text mining and machine learning. As shown in Figure 1, the proposed framework consists of five steps: (1) data preparation, (2) research field (RF) identification, (3) RF investigation, (4) RF relation identification, and (5) future prediction for RF. In step 1, we prepare a collection of process mining research documents using a bibliographic search. Then, the feasible data is constructed based on a series of data preprocessing techniques. Step 2 aims to identify research fields by employing the latest topic modeling approach. Afterward, step 3 investigates the derived research fields on yearly trends, publication types, and nationality. Step 4 identifies the relations between research fields based on network analysis. Finally, we predict the future of each research field by employing a machine learning technique in step 5.

2.1. Data Preparation

The data preparation aims to build feasible data to understand process mining. To this end, we first exploit a bibliographic search from the *Web of Science* database produced by *Clarivate*. The bibliographic search signifies identifying a suitable collection of research papers by combining publications' diverse fields, such as authors, keywords, and abstracts, with search words relating to a specific research discipline. In this research, we retrieve research documents containing *process mining* and *workflow mining* in the title, keyword, and abstract. This is because process mining is the discipline directly connected to this study, while workflow mining is a similar area of research conducted before

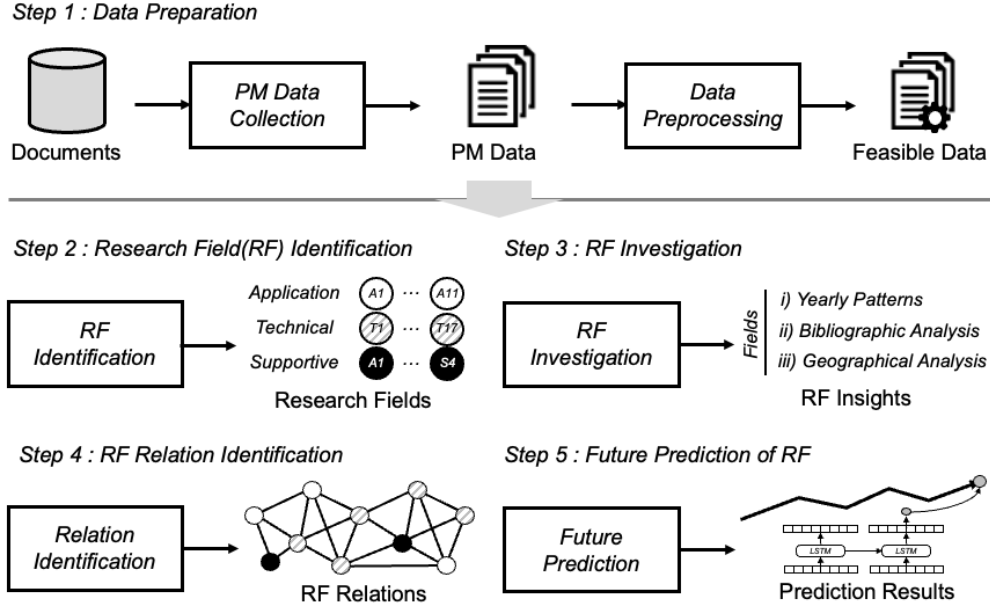


Figure 1: The five steps of the proposed method

process mining was first introduced. Also, only publications written in English in the form of articles and conference proceedings are extracted. As a result of this, we collect 2,271 research papers.

In the data preprocessing phase, we apply a series of methods to the collected data. First, duplicated records with the same title and authors are removed from the collected data. These include cases of deriving a partial development from the previously conducted research or publishing the same article on multiple sites. As such, we delete these 10 redundant records. Second, we remove records that do not contain essential fields of publications, i.e., title, source, publication type, authorities, abstract, addresses, and year. As such, 55 documents that do not contain any of the fields are removed from the data, leading to a total of 2,207 publications as the feasible dataset. Also, we eliminate numbers, symbols, and stopwords in the abstract of the papers since these are the factors hindering effective topic extractions when a topic modeling approach is applied. In addition to this, we apply case conversion and lemmatization, i.e., the most fundamental preprocessing for text data, to reduce unnecessary noise and increase the accuracy of the results.

2.2. Research Field Identification

Based on the feasible data related to the process mining discipline derived from the previous step, this step aims to derive specific research fields based on topic modeling techniques. Topic modeling is a text mining technique that identifies the topics in a collection of words or documents by discovering the hidden semantic structure of documents [23]. There are several popular techniques, such as *Latent Semantic Analysis* [24], *Latent Dirichlet Allocation* [25], and *BERTopic* [26]. In this work, we employ *BERTopic* to utilize a pre-trained language model, e.g., BERT(Bidirectional Encoder Representations from Transformers) [27], to compensate for the short history of the process mining field.

BERTopic consists of three steps: document embedding, clustering, and topic representation. First, the document embedding produces real-valued vectors for given words and sentences using a BERT-based pre-trained language model. A BERT framework is an approach that only employs the self-attention of the encoder in the transformer structure, which is one of the representative generation-based self-supervised learning methods that derive semantic features of language using masked sentences as input. Based on this framework, numerous pre-trained language models have been constructed, and we utilize *sentence-BERT(SBERT) pre-trained model* to embed our data into vector spaces. Second, the embedded data are clustered to group documents with similar topics into the same clusters. To that end, we apply dimension reduction techniques [28], e.g., *PCA*, *t-SNE*, and *UMAP*, and clustering techniques [29], including *K-means* and *HDBSCAN*. Finally, the topic representation step produces a collection of important keywords for each cluster to understand the topics. A class-based TF-IDF is used for this purpose and formalized as follows:

$$W_{x,c} = tf_{x,c} \cdot \log \left(1 + \frac{A}{tf_x} \right) \quad (1)$$

, where c is a cluster derived as a result of clustering and x a term included in cluster c . Here, $tf_{x,c}$ indicates the Term Frequency (TF) of x in cluster c . tf_x and A represent the TF of x in the whole document before clustering and the average number of words per class, respectively. Then, the Inverse of Document Frequency (IDF) becomes the logarithm of A divided by tf_x . Note that 1 is added to hinder the output of the negative value. By combining these two values, i.e., TF and IDF, we measure the importance of the term x in cluster c .

Based on a set of clusters and the critical keywords for each cluster, we transform the discovered clusters into meaningful research fields. As such, a qualitative method based on expert reviews is used, and as a first step, candidates of research fields are derived by analyzing keywords included in each cluster. After that, we reconstruct and group the research fields in consideration of the goals and surroundings and finalize a set of the research fields of a specific discipline. The results will be explained in Section 3.

2.3. Research Field Investigation

This step investigates the derived research fields to obtain meaningful insights by identifying common aspects of them. In this regard, this study focuses on the three viewpoints among the multitude: yearly pattern analysis, bibliographic analysis, and geographical analysis.

First, the yearly patterns attempt to find out the trends for process mining and the specific research fields over time. To this end, it is possible to determine the exemplars of the advancement of research topics, such as dominant groups that have been in the spotlight from the past to the present, skyrocketing groups that have recently gained attention, and deep-cut groups that do not currently receive interests unlike in the past.

Second, the bibliographic analysis investigates the publication types of the process mining discipline. In detail, we can figure out what type of publications the researchers prefer, i.e., journals or conference proceedings, for the whole

discipline and each category. Also, it delivers the most popular sources of research papers in the process mining field, which helps to understand what sources people consider to be a target for publication on a particular topic.

The geographical analysis aims to identify which geographical areas are the most advanced to study in process mining. To this end, we only consider the nationality of the corresponding authors of articles to make it more clear. The results for the yearly pattern analysis, bibliographic analysis, and geographical analysis will be presented in Section 4.

2.4. Relation Identification

This section presents a network analysis method to obtain insights into the relationship between the research fields. The network analysis method used in this study is similar to a typical social network. In other words, the network is composed of research topics as nodes and the similarity between topics as edges [30]. To accomplish this, we perform a series of steps: keyword extraction for each research topic, similarity measurement between research topics based on the keywords, network derivation, and community detection.

The keyword extraction aims to derive key terms of research topics that are used as the inputs of the similarity measurement. To this end, we do not consider the results from topic modeling because the goals for performing these tasks are different. In the research field identification step, the method, i.e., class-based TF-IDF, aims at eliciting keywords representing only a specific topic but not included in other groups. However, it is necessary to investigate keywords that can describe a particular group rather than comparing different groups for this similarity analysis. For this reason, we employ an alternative method to resolve these issues. It is possible to use all kinds of effective methods, including *YAKE!* [31], *KP-Miner* [32], *RakUn* [33], *LexRank* [34], and *SIFRank+* [35]. Among them, we utilize *YAKE!*, i.e., an unsupervised lightweight statistical method that derives the most relevant keywords based on the statistical text features due to the advantage of few restrictions on document size and a high-quality of extracted keywords.

Next, the similarity measurement aims to investigate the noticeable relations between research fields to build a network. To that end, we adopt the Jaccard similarity commonly employed in the natural language processing area [36]. The Jaccard similarity is defined as follows:

$$J(K_A, K_B) = \frac{|K_A \cap K_B|}{|K_A \cup K_B|} = \frac{|K_A \cap K_B|}{|K_A| + |K_B| - |K_A \cap K_B|} \quad (2)$$

, where K_A is the set of keywords for research topic A and K_B is the set of keywords for research topic B extracted by *YAKE!*. K_A and K_B are directly compared to assess how much they overlap quantitatively. Note that the similarity is calculated as 1 when all the items of two fields are overlapped and 0 in the opposite case.

Next, the network derivation aims to derive graph $G = (V, E, \#, w)$ based on the similarity between research topics. The set of nodes $V = RF \times T$ is a set of pairs of a research field and a time step, e.g., $(rf, t) \in V$ is the node for the research field $rf \in RF$ at time $t \in T$. The set of edges $E \subseteq V \times V$ is a set of node pairs such that, for any $e = ((rf_1, t_1), (rf_2, t_2)) \in E$, $rf_1 = rf_2$. A count function $\# \in V \rightarrow \mathbb{N}$ maps nodes to publication counts. A weight

function $w \in E \rightarrow \mathbb{R}$ maps edges to Jaccard similarities, i.e., for any $e = (v_1, v_2) \in E$, $w(e) = J(kw(v_1), kw(v_2))$, where kw is a function mapping a node to a set of keywords.

The derived network basically has the advantage of grasping the relationship between research topics at a glance. Nevertheless, our study intends to in-depth explore the relationship among the research fields based on a community detection method. Community detection refers to assembling nodes with high connection density into a single group [37]. Through this, it is possible to analyze which communities exist on the entire network and what characteristics each community has. In this study, we employ *Louvain algorithm* that can derive a community in a quick and effective manner [38]. The Louvain algorithm utilizes modularity to compare the density of links of inner- and inter-community. More in specific, this algorithm uses the modularity delta concept defined in Equation 3, which aims to determine whether shifting a specific node v into a community C is feasible.

$$\Delta Q(v \rightarrow C) = \left[\frac{\sum_{in} + k_{v,in}}{2m} - \left(\frac{\sum_{tot} + k_v}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_v}{2m} \right)^2 \right] \quad (3)$$

In the formula, m , $\sum_{in}$, and $\sum_{tot}$ indicate the sum of all link weights in the network, the sum of link weights between nodes in community C , and the sum of all link weights of nodes in C , respectively. Also, $k_{v,in}$ is the sum of link weights between nodes in C and node v , whereas k_v is the sum of all link weights of node v . The modularity delta, ΔQ , represents the difference in modularity contribution before and after merging node v in the community C , i.e., optimization. This phase is repeated until the modularity is optimized, and the finally obtained communities are compressed into super-nodes, i.e., aggregation. After that, the optimization of the modularity delta at the super-nodes level continues. We obtain a set of communities by persisting these two optimization and aggregation steps.

The results for the similarity measurement, network detection, and community detection will be presented in Section 5.

2.5. Future Prediction

This section aims to predict future research trends for the identified research fields. To this end, this work employs the results of previous steps in the method part. More in detail, we build a prediction model to predict the number of papers for each research field in the future based on the given historical number of papers for fields and their similarities. Also, we employ Long-Short Term Memory neural networks (LSTM) since it is robust for learning temporal dependency [39].

In order to predict the number of publications in a specific research field in the future, we use a three-dimensional publication presentation matrix $R \in \mathbb{R}^{|RF| \times |RF| \times |T|}$ where RF and T are the set of research fields and time steps, respectively. As shown in Figure 2(a), an element of the matrix, $r_{i,j,t} \in R$, refers to the contribution value of research field r_i at time step $t - 1$ to research field r_j at time step t . Given network $G = (V, E, \#, w)$ used in Subsection 2.4, $r_{i,j,t} \in R$ is formalized as follows:

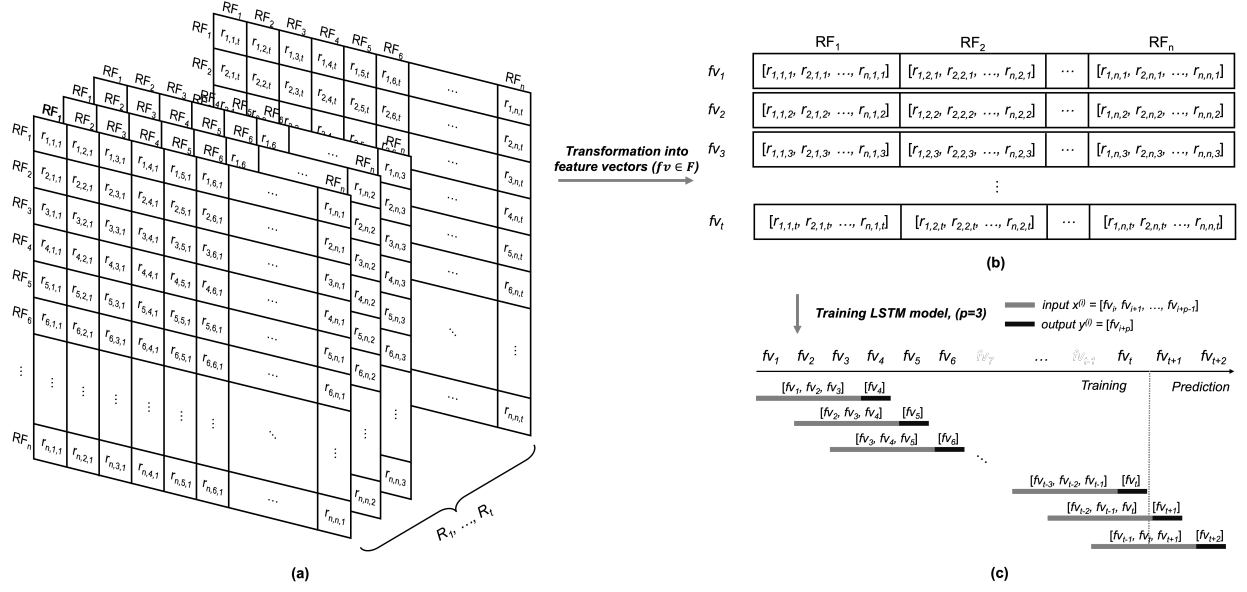


Figure 2: (a) Three-dimensional publication representation matrix, (b) Feature vector, (c) Training and prediction procedures

$$r_{i,j,t} = \frac{w((rf_i, t-1), (rf_j, t))}{\sum_{1 \leq k \leq |RF|} w((rf_k, t-1), (rf_j, t))} \times \#((rf_j, t)) \quad (4)$$

, where $(rf_i, t-1) \in V$ indicates research field rf_i at time step $t-1$ and $(rf_j, t) \in V$ indicates research field rf_j at time step t .

Next, we transform the three-dimensional publication presentation matrix $R \in \mathbb{R}^{|RF| \times |RF| \times |T|}$ into a feature vector $F \subseteq \mathbb{R}^{|RF|}$ such that, for any $f_{v_t} \in F$, $f_{v_t} = [f_{v_{1,t}}, f_{v_{2,t}}, \dots, f_{v_{|RF|,t}}]$ where $f_{v_{j,t}} = [r_{i,j,t} | 1 \leq i \leq |RF|]$ for $1 \leq j \leq |RF|$, as shown in Figure 2(b). For instance, $f_{v_1} = [f_{v_{1,1}}, f_{v_{2,1}}, \dots, f_{v_{|RF|,1}}]$ is a feature vector at time $t = 1$ where $f_{v_{1,1}} = [r_{1,1,1}, r_{2,1,1}, \dots]$ indicates the contribution values of all research fields to research field rf_1 at $t = 1$, $f_{v_{2,1}} = [r_{1,2,1}, r_{2,2,1}, \dots]$ indicates the contribution values of all research fields to research field rf_2 at $t = 1$, etc.

As shown in Figure 2(c), the LSTM model is trained to predict the feature vector at $t+1$, given the last $1 \leq p \leq t$ feature vectors. Therefore, the i -th training sample consists of $x^{(i)} = [f_{v_i}, f_{v_{i+1}}, \dots, f_{v_{i+p-1}}]$ and $y^{(i)} = f_{v_{i+p}}$. For instance, given $p = 3$, the first training sample consists of $x^{(1)} = [f_{v_1}, f_{v_2}, f_{v_3}]$ and $y^{(1)} = f_{v_4}$, while the second training sample consists of $x^{(2)} = [f_{v_2}, f_{v_3}, f_{v_4}]$ and $y^{(2)} = f_{v_5}$. Based on the predicted feature vector at $t+1$, i.e., $f_{v_{t+1}}$, we compute the predicted number of papers, for research field rf_j at $t+1$, i.e., $R_{j,t+1}$, as $\sum_{1 \leq i \leq |RF|} [r_{i,j,t} \in f_{v_{j,t+1}}]$.

3. Identification of research fields in process mining

In this section, we explain the topics discovered from the topic modeling presented in Subsection 2.2. First, we map each discovered topic into the existing field of process mining by analyzing the keywords and group the topics

into three categories by qualitatively analyzing the similarity between topics. Next, we explain the individual topics in detail according to the categories.

3.1. Topic Mapping & Grouping

First, we map the discovered topics from text mining to the existing fields in process mining by analyzing the keywords representing each topic. Figure 3(a) shows the discovered topics mapped to the existing fields in process mining. For instance, *topic 6* is mapped to *general process discovery* since it contains keywords such as *discovery*, *discovered*, *discovering*, etc. *Topic 7* is mapped to *conformance checking* since its keywords include *conformance*, *compliance*, *checking*, etc. Moreover, *Topic 29* is mapped to *healthcare* since it includes keywords such as *patient*, *clinic*, *hospital*, etc.

Next, we categorize the topics 1) by merging the topics that share the same goals and 2) by classifying them into three categories. We merge the topics based on the similarities in the associated fields, as shown in Figure 3(b), resulting in 32 merged topics. For instance, *workflow mining*, *genetic algorithm-based mining*, *region-based mining*, *pattern mining*, *Petri-nets*, and *general process discovery* are merged into *process discovery* since they all aim to discover process models from event logs (but in different formalism with different approaches). *Conformance checking* and *alignments* are merged into *conformance checking* since *alignments* is an approach to *conformance checking*. Moreover, *smart artifacts* and *Internet of Things (IoT) security* are merged into *IoT*.

We group the merged fields into three categories: *Technique*, *Application*, and *Support*. The technique category contains process mining techniques, whereas the application category contains topics that focus on applying such process mining techniques in different domains such as healthcare, software, education, manufacturing, etc. In addition, the support category contains topics that provide the fundamentals for the development of process mining techniques. Figure 3(c) shows the grouping of different fields. For instance, *process discovery*, *conformance checking*, and *performance-driven process mining* are classified into the technique category, whereas *healthcare*, *software*, and *education* are grouped into the application category. Moreover, *machine learning*, *business intelligence*, and *business process management* are categorized into the support category.

3.2. Topic Description

In this subsection, we explain the 32 merged topics from the previous subsection based on the three categories.

3.2.1. Technique Category

Figure 4 shows a word cloud representing keywords that explain topics that belong to this category. The keywords include *workflow*, *discovery*, *algorithm*, *behavior*, *time*, and *trace*.

Below are the topics in the technique category:

- **T1: Process Discovery (451 articles)** aims to discover process models from event logs. Different articles propose different process discovery techniques by varying the modeling formalism of the discovered process

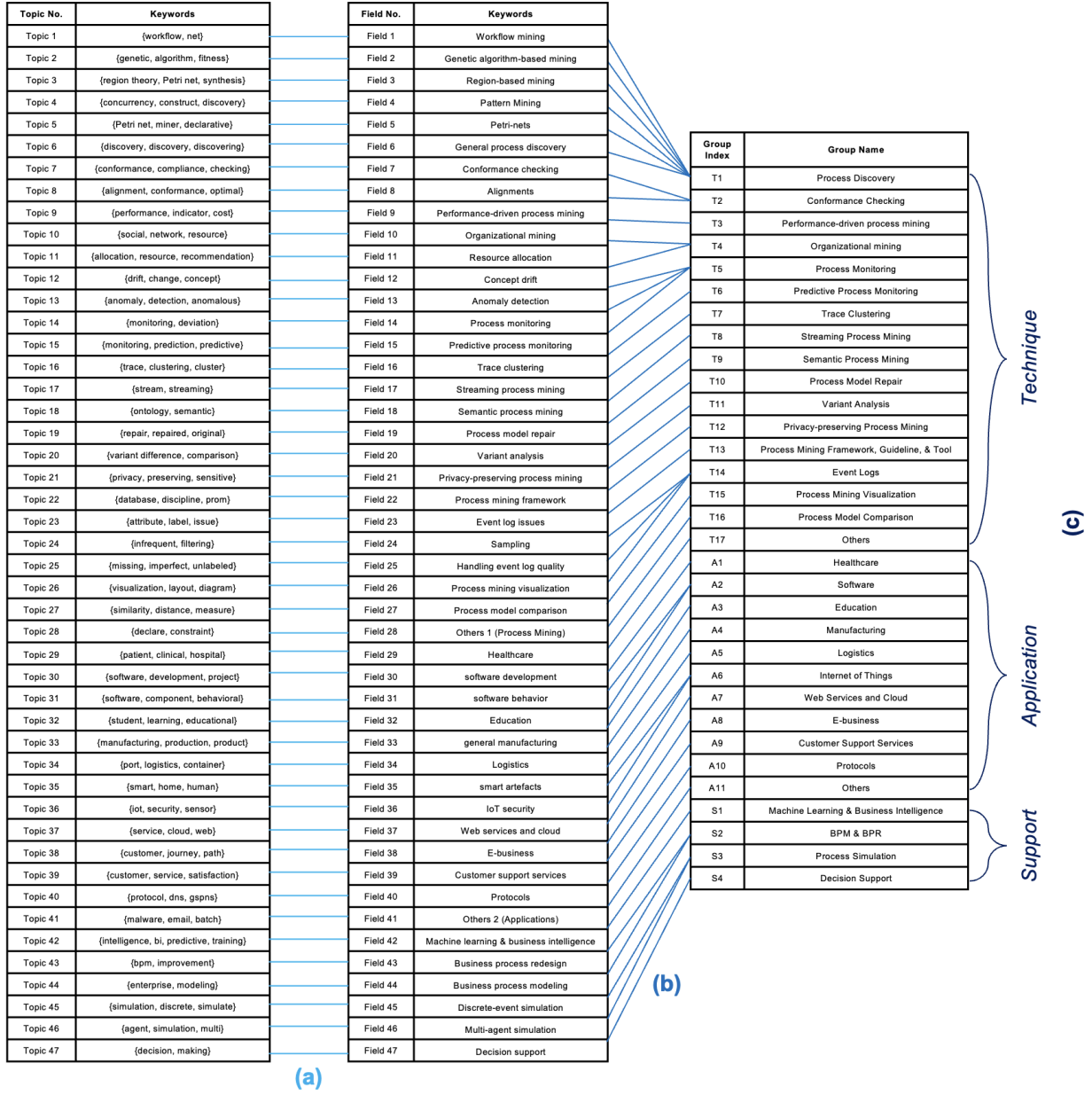


Figure 3: (a) Mapping 47 discovered topics to 47 existing fields of process mining based on the associated keywords, (b) Merging 47 research fields into 32 fields, (c) Classifying 32 research fields into three categories

model, such as Petri nets, BPMNs, UML diagrams, etc. Moreover, for the same modeling formalism (e.g., Petri nets), different articles propose different approaches (e.g., region-based approaches, genetic algorithm-based approaches, etc.).

- **T2: Conformance Checking (136 articles)** aims to compute the conformance of event logs to process models using various conformance measures. Different articles propose different techniques to tackle different

logs. Different techniques focus on 1) deriving process features that represent process instances, 2) computing coherent clusters using different clustering algorithms, 3) and explaining the discovered clusters.

- **T8: Streaming Process Mining (17 articles)** aims to analyze streaming event logs, unlike other process mining techniques that analyze historical event logs. Different techniques focus on 1) storing streaming event logs, 2) discovering process models from streaming event logs, and 3) checking the conformance of running process instances.
- **T9: Semantic Process Mining (25 articles)** aims to leverage semantic information in the form of knowledge graphs and process ontology to enhance the existing process mining techniques. Different techniques focus on enhancing different process mining techniques using different forms of semantic information.
- **T10: Process Model Repair (18 articles)** aims to support repairing process models to reflect reality. Existing techniques first uncover differences between observed behavior (i.e., event logs) and modeled behavior (i.e., process model) and then modify process models while ensuring the soundness of the model.
- **T11: Variant Analysis (17 articles)** aims to analyze different forms of variants existing in business processes. First, process variant analysis aims to improve the original reference model such that it fits variant process models. Second, trace variant analysis aims to analyze unique sequences of activities by different process instances based on event logs.
- **T12: Privacy-preserving Process Mining (15 articles)** aims to protect event data privacy such that the data conform to privacy regulations on the technical design and the organizational application of process mining techniques. Existing work provides privacy-preserving data publishing techniques to anonymize event data while maintaining the utility of the data.
- **T13: Process Mining Framework, Guideline, & Tool (72 articles)** aims to provide frameworks to develop process mining techniques, guidelines to apply process mining techniques in various domains, and tools to support the implementation of process mining techniques. The ProM framework provides a software framework to support the process mining tooling, and the process mining manifesto describes opportunities and challenges in process mining.
- **T14: Event Logs (101 articles)** provides various techniques to support the management of event logs. First, extraction techniques support the extraction of event logs from source systems (e.g., SAP ERP systems). Second, sampling techniques support the sampling of event logs to enhance the performance of various process mining techniques such as conformance checking. Finally, log quality techniques support the assessment of log qualities and their enhancement, e.g., by handling missing information.

- **T15: Process Mining Visualization (18 articles)** aims to provide various visualizations of process mining insights. Existing techniques focus on improving the visualization of process models by applying graph layout algorithms and supporting the visualization of social networks using various diagrams, e.g., Chord diagram.
- **T16: Process Model Comparison (13 articles)** aims to analyze the similarity of process models. Some of the techniques analyze the structure of process models, whereas others analyze the semantics of process models.
- **T17: Others (56 articles)** contains articles that provide the design/implementation of process mining techniques but do not belong to a specific existing field. Such articles deal with *Big Data*, agile development, natural language processing, etc.

Figure 5 shows a word cloud representing keywords that explain topics that belong to the application category. The keywords include *case study*, *healthcare*, *knowledge*, and *application*.

Figure 5: Word cloud representing keywords associated with the technique category

- **A1. Healthcare (267 articles)** aims to analyze healthcare-related business processes using process mining techniques. Existing articles present clinical pathway patterns from medical behaviors and analyses of the length of hospital stay based on process discovery and performance-driven process mining techniques.
- **A2. Software (103 articles)** aims to analyze software development processes and software behavior using process mining techniques. Whereas the former focuses on analyzing business processes related to software developments, the latter focuses on the system/user behavior of the developed software.

- **A3: Education (78 articles)** aims to analyze the learning processes of students in different formats of education, e.g., online learning, self-regulated learning, etc. Process discovery techniques are often deployed to provide an overview of students' learning behavior.
- **A4: Manufacturing (51 articles)** aims to analyze end-to-end manufacturing processes, including planning (i.e., balancing demand and supply), sourcing (i.e., procuring materials), making (i.e., transforming materials to finished products), delivering (i.e., providing finished products to retailers), and returning (i.e., managing used products) processes. Various manufacturing domains have been studied, such as semiconductor manufacturing, make-to-order manufacturing, the automotive industry, etc.
- **A5: Logistics (12 articles)** aims to analyze business processes related to logistics and storage management. The application focuses on predicting the Estimated Time of Arrival (ETA), scheduling transport, improving warehouse layout, etc. The proposed applications rely on tagging systems such as the Radio Frequency Identification (RFID) system to identify the movement of goods.
- **A6: Internet of Things (50 articles)** aims to analyze the data generated or captured by the Internet of Things (IoT) using process mining techniques. On the one hand, the existing application focuses on analyzing human behavior patterns, e.g., the behavior pattern of residents in a nursing home based on a wireless tracking system. On the other hand, the application concerns the detection of misbehavior in IoT-based systems to enhance the security of the system.
- **A7: Web and Cloud Services (32 articles)** aims to analyze processes in Web and Cloud services, e.g., Google Cluster, to optimize resources. To that end, process mining techniques are used to discover user behavior by analyzing execution logs and monitoring the violation of business constraints. Furthermore, the insights from process mining techniques are used to optimize the scheduling of resources.
- **A8: E-business (18 articles)** aims to analyze customer journeys in online business processes on the web. The existing application uses the click stream data of users to automatically discover customer journey maps. Comparative process mining techniques are applied to compare the behavior of different types of customers, e.g., the behavior of bargain shoppers and surgical shoppers.
- **A9: Customer Support Services (15 articles)** aims to analyze business processes of help desks and customer services that resolve reported issues. As well as the insights from process mining are used to improve the process, they are used to develop new customer support services, e.g., adaptive automated teller machines based on the insights from the usage analysis of automated teller machines.
- **A10: Protocols (13 articles)** aims to analyze protocols of software systems. On the one hand, the existing application aims to discover protocols of proprietary legacy software systems, e.g., TCP/IP-Protocol. On the

other hand, the application concerns the detection of anomalies in protocol traces, e.g., a mail bonnet attack in DNS protocol traces.

- **A11: Others (121 articles)** contains articles that employ process mining but do not belong to a specific domain. Such articles are engaged in the following domains: e-mail, *smart manufacturing*, *industry 4.0*, *robotic process automation*), *blockchain*, etc.

3.2.3. Support Category

Figure 6 shows a word cloud that represents keywords explaining topics in this category. The keywords include *user*, *resource*, *knowledge*, and *modeling*.



Figure 6: Word cloud representing keywords associated with the technique category

Below are the topics in the support category:

- **S1: Machine Learning and Business Intelligence (73 articles)** relates techniques for machine learning and business intelligence to enable process mining techniques such as process discovery, conformance checking, performance analysis, etc. For instance, statistical relational learning supports the discovery of process models, and frequent itemset mining supports organizational mining.
- **S2: Business Process Management and Business Process Redesign (77 articles)** relates techniques in business process management and business process redesign. The formal provides the fundamentals of process models that are used as modeling formalism employed in process discovery techniques, whereas the latter introduces redesign patterns in business processes that support the process improvement based on the insights from process mining techniques.
- **S3: Process Simulation (40 articles)** refers to the study related to discrete event simulation of business processes. Based on the existing simulation techniques, process mining provides techniques for discovering such

simulation models to automate the design and implementation of the simulation models. Some techniques automatically discover simulation models with more advanced constructs, such as the presence of multitasking and availability constraints.

- **S4: Decision Support (16 articles)** refers to the design and implementation of decision support systems. The suggested decision support systems support the adoption of process mining techniques to realize informed decisions based on data-driven insights and automatically generate recommendations (i.e., decisions).

4. Investigation into research fields in process mining

This section investigates the research fields derived in Section 3 based on the relevant common aspects: yearly patterns, bibliographic analysis, and geographical analysis.

4.1. Yearly Patterns

Figure 7 presents the number of publications in each research field by year. As depicted in the figure, the process mining discipline was introduced in the early 2000s and has been gradually developed due to many efforts. In particular, the number of publications has been over 100 since 2013, and almost half of the papers were published in the last five years. In other words, this field has recently received more attention. Also, compared to the support category, both technique and application domains have been in the same flow by year. That is, a bunch of applications specialized in various industries has appeared with the development of numerous process mining techniques.

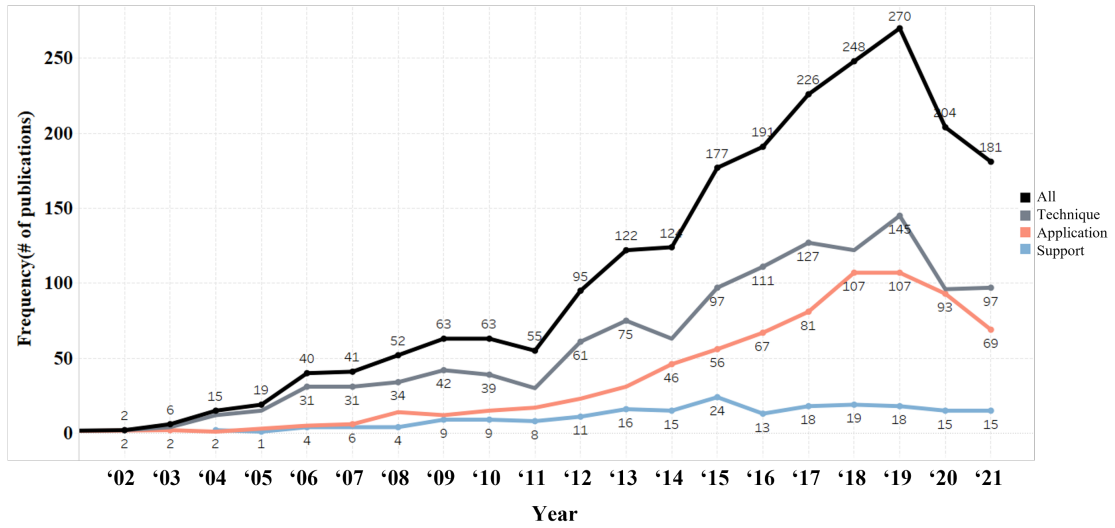
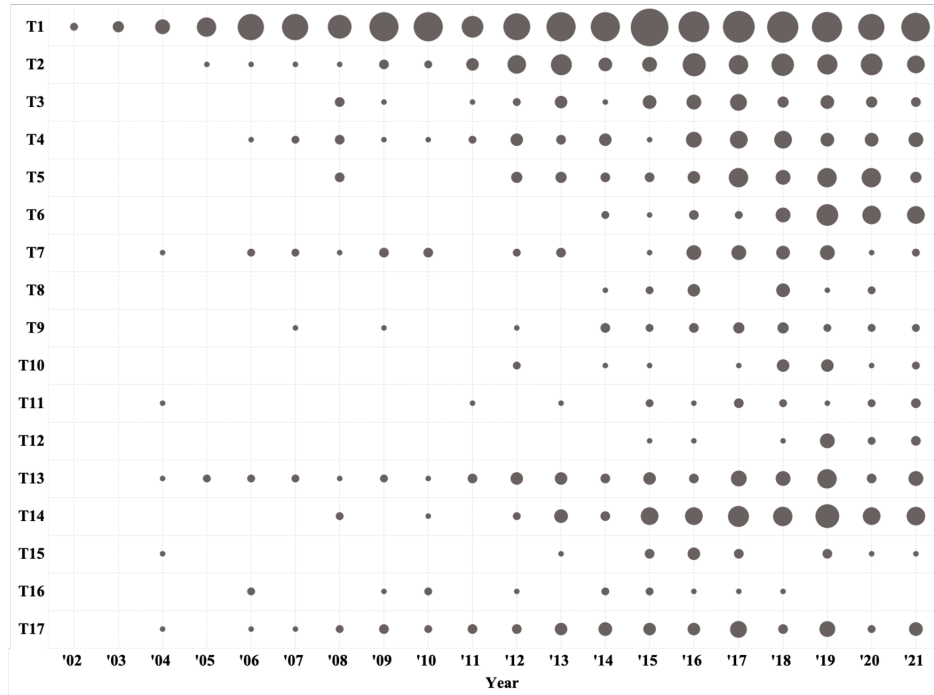
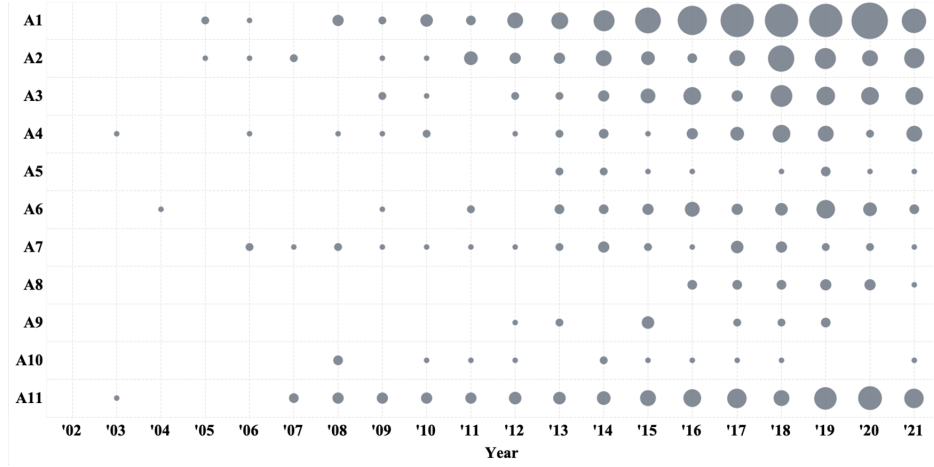


Figure 7: The number of publications in each category by year

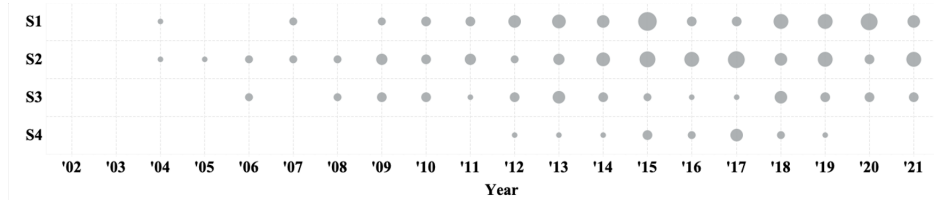
Next, we explore the yearly patterns for each research field in the fine-grained viewpoints, which aims at classifying them into the following archetypes: dominant, skyrocketing, and deep-cut patterns. Figure 8(a) presents the number of publications by year for each research field in the technique category.



(a) Technique category



(b) Application category



(c) Support category

Figure 8: The number of publications in each research field by year

In the technique category, process discovery (T1), conformance checking (T2), and process mining framework, guideline, & tool (T13) belong to the dominant groups. For example, T1 is the most fundamental field in the process mining discipline. In other words, it has been constantly studied every year, and new algorithms have been developed even recently. More in detail, this field first appeared as workflow mining techniques, such as *alpha-algorithm*. Also, it covered both initial process discovery algorithms, e.g., *heuristic mining* and *fuzzy mining*, and modern techniques, e.g., *inductive mining*.

Moreover, researchers in the process mining community have taken a keen interest in conformance checking, i.e., T2. It first emerged with initial research on measuring conformance based on the token replay approach in 2008. Since then, it has been steadily researched and stands as an essential research topic in process mining with modern techniques, such as multi-perspective, stochastic, and online conformance checking. Regarding T13, we can easily understand that they belong to the dominant group since *process mining framework, guideline, and tools* are the essential components of process mining.

As far as the skyrocketing patterns are concerned, we can consider process monitoring (T5), predictive process monitoring (T6), privacy-preserving process mining (T12), and event logs (T14) in the technique category. Process monitoring and predictive process monitoring, i.e., research areas that have expanded from the naive performance analysis, have achieved much growth with the development of machine learning-related technologies in recent years. Research on T5 has been more actively published since 2017 with the aim of catching sudden changes in performance measurements in the process environment with machine learning approaches, including anomaly detection and concept drift. On the other hand, predictive process monitoring, i.e., T6, has received much attention since 2019 based on multiple deep learning algorithms to predict performance at the process instance level. Regarding T12, there have been some efforts to preserve the privacy of event logs in recent years, with a trend that data security has become essential in the data science field. Also, T14 has been steadily progressing in research in various directions, including event abstraction, complex event processing, data cleansing, and quality. Furthermore, it has recently received strong attention due to the emergence of a new type of event logs, e.g., object-centric event logs.

Lastly, we classify the process model comparison topic (T16) into the deep-cut group. This topic aims to compare the two given process models structurally, and recent studies have been reduced due to limitations in their utilization and, thus, low demands for new technological developments.

The yearly pattern results for the application and support categories are depicted in Figure 8(b) and Figure 8(c), respectively. Among the research fields, the followings are included in the dominant group: healthcare (A1), software (A2), education (A3), manufacturing (A4), machine learning and business intelligence (S1), and business process management and business process redesign (S2). The healthcare domain is one of the representative fields where process mining is actively and steadily employed. The articles in this field have covered various healthcare processes, including outpatient, inpatient, and emergency departments, and they triggered the publication of review papers on process mining in healthcare. In addition, process mining has been consistently applied to the software, education, and manufacturing areas from the past to the present. In the support category, the number of publications in S1 and

S2, i.e., machine learning and business process management, is insufficient compared to other classes, but these also have been popular subjects for researchers. Also, we identify only two research fields in the application category, i.e., internet of things (A6) and e-business (A8) are considered as the skyrocketing groups, whereas customer support services (A9), protocols (A10), and decision support (S4) are the topics deep-cut.

4.2. Bibliographic analysis

This section aims to obtain new insights about bibliographic features by research fields and the whole process mining discipline. In more detail, we investigate the proportions of publication types and specific sources mainly published in process mining.

First, Figure 9 shows the result of comparing the proportions of articles according to the publication type for each category and the whole. The process mining discipline has published 32% of journal articles and 68% of conference proceedings, which is similar to the distribution of general computer science disciplines. In addition, by exploring the results at the category level, we can identify that application-oriented articles have been published more in journals than those in the technique and support categories.

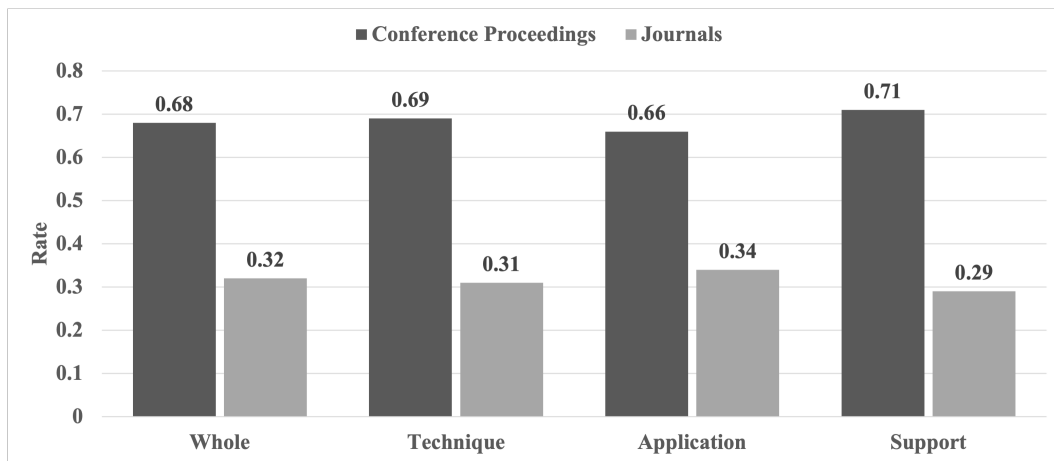


Figure 9: The proportions of publication types in process mining

Next, we perform the same analysis for each research field, and Figure 10 presents its results. It is recognizable that the number of conference proceedings exceeds that of journal articles in most cases. Specifically, semantic process mining (T9), process mining framework, guideline, & tool (T13), customer support services (A9), and decision support (S4) have a significantly higher proportion of conference proceedings. On the other hand, the number of journal articles is more than the number of conference proceedings in logistics (A5) and process model repair (T10).

Next, we identify the most frequently published publication sources in journals and conference proceedings for each category and the whole discipline. Table 1 describes the results to identify the top seven sources of publications. Many journal articles have been published in the disciplines of computer science, information systems, software

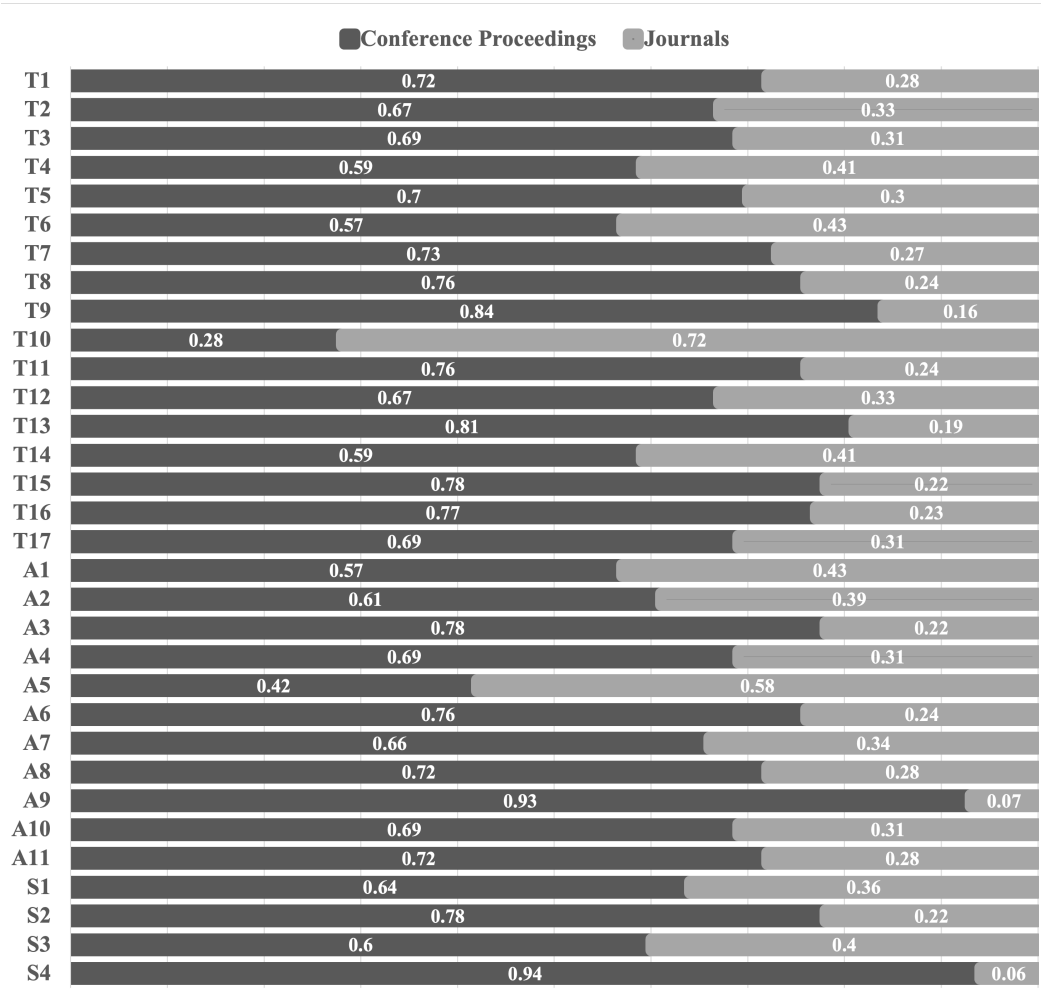


Figure 10: The proportions of publication types for each research field

engineering, and management science, such as *Information Systems*, *Expert Systems with Applications*, *Decision Support Systems*, and *Software and Systems Modeling*. This tendency is similar in the technique and support categories. However, the application-oriented articles show disparate results, i.e., they have been published in multi-disciplinary journals focusing on producing case studies of process mining, such as *IEEE Access*, *Applied Sciences*, and *Sensors*. Furthermore, domain-specific journals, e.g., the healthcare-specialized, are listed in Table 1 as well, including *International Journal of Environmental Research and Public Health* and *Journal of Biomedical Informatics*.

As far as the conference proceedings are concerned, there is no significant discrepancy in results by the whole and each category. Most of the research papers have been published in conferences on *Business Process Management (BPM)*, *Advanced Information Systems Engineering (CAiSE)*, *Enterprise Information Systems (ICEIS)*, and *On The Move To Meaningful Internet Systems (OTM)*. After the initialization of the *Process Mining Conference (ICPM)* in 2019, a large number of up-to-date research articles have been published through this.

Table 1: The top 7 journals and conference proceedings by the whole and categories

	Journal	Conference Proceedings
Whole	Information Systems IEEE Access Expert Systems with Applications Decision Support Systems Software and Systems Modeling IEEE Transactions on Services Computing Applied Sciences	Business Process Management (BPM) Advanced Information Systems Engineering (CAiSE) Enterprise Information Systems (ICEIS) ICT and Knowledge Engineering (ICT-KE) On The Move To Meaningful Internet Systems (OTM) Process Mining (ICPM) Enterprise Distributed Object Computing (EDOC)
Technique	Information Systems IEEE Access Expert Systems with Applications Software and Systems Modeling Decision Support Systems IEEE Transactions on Services Computing Data and Knowledge Engineering	Business Process Management (BPM) On The Move To Meaningful Internet Systems (OTM) Advanced Information Systems Engineering (CAiSE) Enterprise Information Systems (ICEIS) Enterprise Distributed Object Computing (EDOC) Process Mining (ICPM) Conceptual Modeling (ER)
Application	International Journal of Environmental Research and Public Health IEEE Access Expert Systems with Applications Journal of Biomedical Informatics Applied Sciences Sensors Information Systems	Business Process Management (BPM) ICT and Knowledge Engineering (ICT-KE) Enterprise Information Systems (ICEIS) Biomedical Engineering Systems and Technologies (BIOSTEC) Learning Analytics & Knowledge (LAK) Advanced Information Systems Engineering (CAiSE) Process Mining (ICPM)
Support	Information Systems Software and Systems Modeling IEEE Transactions on Services Computing Expert Systems with Applications Decision Support Systems Data and Knowledge Engineering Computers in Industry	Business Process Management (BPM) Advanced Information Systems Engineering (CAiSE) Enterprise Information Systems (ICEIS) Simulation Conference (WSC) Process Mining (ICPM) Enterprise Distributed Object Computing (EDOC) Business Information Systems (BIS)

4.3. Geographic Analysis

The final goal of investigating research articles in process mining is to identify the discrepancy of geographical aspects by research fields. To this end, we uncover the leading countries that frequently publish process mining articles. To obtain more explicit insights, we analyze the country where the affiliation of the corresponding author of each paper is located. Table 2 describes the results of this.

Table 2: The top 7 countries by the whole and categories

	Whole	Technique	Application	Support
Country	Netherlands	Netherlands	Italy	Germany
	China	China	China	Netherlands
	Italy	Germany	Netherlands	China
	Germany	Italy	Spain	Brazil
	Spain	Australia	France	France
	South Korea	South Korea	Germany	Spain
	France	U.S.	Thailand	Belgium

Based on the results, we can identify that the Netherlands has occupied the most dominant position in the process mining discipline. In addition, China, Italy, and Germany have remarkably contributed to the development of the process mining field as well. These four countries have published almost 40% of the whole research articles, especially the Netherlands and China occupied 12.9% and 10.7% of the total, respectively.

The technique category shows a similar pattern to the whole process mining discipline. The Netherlands, China, Germany, and Italy have also contributed significantly to technique-related publications. However, unlike the overall distribution, many techniques have been developed in Australia and the United States as well. The application-oriented articles have been published by authors in Italy, China, and the Netherlands. Besides, Spain and France also have contributed to the emergence of the application of process mining to various domains. In the support category, Germany and the Netherlands have occupied a large part of this. In addition, Brazil and Belgium are also included in the list as the key regions for the support field.

5. Relation identification between research fields in process mining

This section aims at discovering noticeable relations between research fields in process mining. To achieve this objective, we employed a series of methods, e.g., similarity analysis, network construction, and community detection, as introduced in Subsection 2.4.

5.1. Similarity analysis

Before measuring the similarity between research fields, we apply YAKE! presented in Subsection 2.4 to extract principal terms to represent them. Then, we measured the similarity between research fields based on the Jaccard similarity method. Figure 11 describes the result of the Jaccard similarity measurements between research topics.

		Technique																	Application											Support			
		T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	S1	S2	S3	S4
Technique	T1	1.000	0.232	0.254	0.227	0.208	0.193	0.222	0.219	0.204	0.206	0.217	0.197	0.247	0.235	0.215	0.200	0.258	0.193	0.194	0.185	0.195	0.169	0.191	0.197	0.179	0.177	0.194	0.175	0.227	0.226	0.227	0.207
	T2	0.232	1.000	0.200	0.190	0.218	0.202	0.212	0.189	0.213	0.207	0.225	0.183	0.231	0.214	0.218	0.189	0.224	0.186	0.191	0.178	0.169	0.177	0.169	0.211	0.170	0.173	0.193	0.175	0.202	0.229	0.199	0.178
	T3	0.254	0.200	1.000	0.216	0.229	0.224	0.225	0.213	0.222	0.209	0.216	0.190	0.229	0.218	0.218	0.190	0.229	0.206	0.216	0.182	0.215	0.189	0.190	0.196	0.177	0.186	0.188	0.182	0.229	0.236	0.251	0.216
	T4	0.227	0.190	0.216	1.000	0.207	0.203	0.214	0.210	0.217	0.181	0.195	0.195	0.224	0.223	0.209	0.187	0.212	0.208	0.196	0.202	0.196	0.177	0.196	0.200	0.193	0.194	0.164	0.194	0.224	0.221	0.213	0.216
	T5	0.208	0.218	0.229	0.207	1.000	0.192	0.229	0.210	0.221	0.191	0.207	0.179	0.214	0.202	0.203	0.175	0.199	0.185	0.193	0.191	0.194	0.168	0.213	0.204	0.188	0.164	0.176	0.179	0.209	0.205	0.208	0.185
	T6	0.193	0.202	0.224	0.203	0.192	1.000	0.230	0.207	0.210	0.174	0.211	0.178	0.201	0.193	0.211	0.181	0.210	0.183	0.173	0.183	0.199	0.167	0.186	0.170	0.185	0.172	0.174	0.183	0.211	0.214	0.211	0.184
	T7	0.222	0.212	0.225	0.214	0.229	0.230	1.000	0.200	0.206	0.194	0.239	0.196	0.220	0.225	0.217	0.201	0.211	0.193	0.200	0.206	0.190	0.185	0.201	0.178	0.198	0.173	0.182	0.187	0.214	0.229	0.209	0.195
	T8	0.219	0.189	0.213	0.210	0.210	0.207	0.200	1.000	0.213	0.190	0.205	0.208	0.235	0.209	0.195	0.177	0.213	0.191	0.190	0.172	0.194	0.174	0.197	0.196	0.178	0.174	0.175	0.183	0.198	0.202	0.191	0.188
	T9	0.204	0.213	0.222	0.217	0.221	0.210	0.206	0.213	1.000	0.184	0.224	0.188	0.235	0.198	0.208	0.204	0.229	0.188	0.187	0.201	0.198	0.174	0.220	0.201	0.190	0.179	0.186	0.185	0.216	0.225	0.203	0.189
	T10	0.206	0.207	0.209	0.181	0.191	0.174	0.194	0.190	0.184	1.000	0.195	0.148	0.180	0.200	0.204	0.176	0.192	0.168	0.170	0.162	0.158	0.158	0.155	0.158	0.152	0.161	0.183	0.153	0.179	0.179	0.174	0.158
	T11	0.217	0.225	0.216	0.195	0.207	0.211	0.239	0.205	0.224	0.195	1.000	0.197	0.206	0.212	0.213	0.214	0.231	0.202	0.205	0.190	0.188	0.193	0.190	0.188	0.200	0.191	0.188	0.183	0.208	0.200	0.201	0.193
	T12	0.197	0.183	0.190	0.195	0.179	0.178	0.196	0.208	0.188	0.148	0.197	1.000	0.204	0.182	0.214	0.181	0.189	0.193	0.177	0.166	0.179	0.182	0.205	0.205	0.193	0.179	0.182	0.187	0.189	0.188	0.175	0.197
	T13	0.247	0.231	0.229	0.224	0.214	0.201	0.220	0.235	0.235	0.180	0.206	0.204	1.000	0.219	0.233	0.197	0.253	0.200	0.205	0.204	0.216	0.199	0.202	0.196	0.192	0.191	0.180	0.200	0.226	0.243	0.227	0.186
	T14	0.235	0.214	0.218	0.223	0.202	0.193	0.225	0.209	0.198	0.200	0.212	0.182	0.219	1.000	0.195	0.188	0.221	0.183	0.190	0.180	0.200	0.172	0.203	0.171	0.175	0.179	0.166	0.189	0.209	0.212	0.195	0.178
	T15	0.215	0.218	0.218	0.209	0.203	0.211	0.217	0.195	0.208	0.204	0.213	0.214	0.233	0.195	1.000	0.196	0.220	0.207	0.204	0.204	0.188	0.205	0.199	0.191	0.231	0.197	0.185	0.201	0.225	0.234	0.219	0.224
	T16	0.200	0.189	0.190	0.187	0.175	0.181	0.201	0.177	0.204	0.176	0.214	0.181	0.197	0.188	0.196	1.000	0.198	0.164	0.173	0.179	0.174	0.167	0.181	0.176	0.177	0.154	0.173	0.175	0.205	0.196	0.176	0.196
	T17	0.258	0.224	0.229	0.212	0.199	0.210	0.211	0.213	0.229	0.192	0.231	0.189	0.253	0.221	0.220	0.198	1.000	0.214	0.221	0.193	0.190	0.189	0.201	0.188	0.181	0.182	0.192	0.196	0.220	0.245	0.211	0.194
Application	A1	0.193	0.186	0.206	0.208	0.185	0.183	0.193	0.191	0.188	0.168	0.202	0.193	0.200	0.183	0.207	0.164	0.214	1.000	0.194	0.198	0.189	0.200	0.186	0.191	0.183	0.176	0.168	0.193	0.202	0.208	0.189	0.190
	A2	0.194	0.191	0.216	0.196	0.193	0.173	0.200	0.190	0.187	0.170	0.205	0.177	0.205	0.190	0.204	0.173	0.221	0.194	1.000	0.199	0.202	0.175	0.185	0.182	0.169	0.176	0.168	0.183	0.199	0.213	0.201	0.184
	A3	0.185	0.178	0.182	0.202	0.191	0.183	0.206	0.172	0.201	0.162	0.190	0.166	0.204	0.180	0.204	0.179	0.193	0.198	0.199	1.000	0.176	0.176	0.182	0.166	0.171	0.200	0.148	0.189	0.197	0.182	0.196	0.190
	A4	0.195	0.169	0.215	0.196	0.194	0.199	0.190	0.194	0.198	0.158	0.188	0.179	0.216	0.200	0.188	0.174	0.190	0.189	0.202	0.176	1.000	0.205	0.218	0.191	0.176	0.195	0.167	0.198	0.206	0.203	0.195	0.199
	A5	0.169	0.177	0.189	0.177	0.168	0.167	0.185	0.174	0.174	0.158	0.193	0.182	0.199	0.172	0.205	0.167	0.189	0.200	0.175	0.176	0.205	1.000	0.168	0.153	0.171	0.175	0.142	0.181	0.184	0.185	0.160	0.187
	A6	0.191	0.169	0.190	0.196	0.213	0.186	0.201	0.197	0.220	0.155	0.190	0.205	0.202	0.203	0.199	0.181	0.201	0.186	0.185	0.182	0.218	0.168	1.000	0.209	0.186	0.185	0.169	0.182	0.197	0.200	0.200	0.181
	A7	0.197	0.211	0.196	0.200	0.204	0.170	0.178	0.196	0.201	0.158	0.188	0.205	0.196	0.171	0.191	0.176	0.188	0.191	0.182	0.166	0.191	0.153	0.209	1.000	0.169	0.163	0.170	0.174	0.182	0.222	0.202	0.189
	A8	0.179	0.170	0.177	0.193	0.188	0.185	0.198	0.178	0.190	0.152	0.200	0.193	0.192	0.175	0.231	0.177	0.181	0.183	0.169	0.171	0.176	0.171	0.186	0.169	1.000	0.179	0.150	0.166	0.180	0.192	0.175	0.181
	A9	0.177	0.173	0.186	0.194	0.164	0.172	0.173	0.174	0.179	0.161	0.191	0.179	0.191	0.179	0.197	0.154	0.182	0.176	0.176	0.200	0.195	0.175	0.185	0.163	0.179	1.000	0.158	0.178	0.187	0.171	0.182	0.177
	A10	0.194	0.193	0.188	0.164	0.176	0.174	0.182	0.175	0.186	0.183	0.188	0.182	0.180	0.166	0.185	0.173	0.192	0.168	0.168	0.148	0.167	0.142	0.169	0.170	0.150	0.158	1.000	0.165	0.175	0.187	0.197	0.176
	A11	0.175	0.175	0.182	0.194	0.179	0.183	0.187	0.183	0.185	0.153	0.183	0.187	0.200	0.189	0.201	0.175	0.196	0.193	0.183	0.189	0.198	0.181	0.182	0.174	0.166	0.178	0.165	1.000	0.193	0.197	0.191	0.177
Support	S1	0.227	0.202	0.229	0.224	0.209	0.211	0.214	0.198	0.216	0.179	0.208	0.189	0.226	0.209	0.225	0.205	0.220	0.202	0.199	0.197	0.206	0.184	0.197	0.182	0.180	0.187	0.175	0.193	1.000	0.231	0.211	0.212
	S2	0.226	0.229	0.236	0.221	0.205	0.214	0.229	0.202	0.225	0.179	0.200	0.188	0.248	0.212	0.234	0.196	0.245	0.208	0.213	0.182	0.203	0.185	0.200	0.222	0.192	0.171	0.187	0.197	0.231	1.000	0.224	0.208
	S3	0.227	0.199	0.251	0.213	0.208	0.211	0.209	0.191	0.203	0.174	0.201	0.175	0.227	0.195	0.219	0.176	0.211	0.189	0.201	0.196	0.195	0.160	0.200	0.202	0.175	0.182	0.197	0.191	0.211	0.224	1.000	0.196
	S4	0.207	0.178	0.216	0.216	0.185	0.184	0.195	0.188	0.189	0.158	0.193	0.197	0.186	0.178	0.224	0.196	0.194	0.190	0.184	0.190	0.199	0.187	0.181	0.189	0.181	0.177	0.176	0.177	0.212	0.208	0.196	1.000

Figure 11: The Jaccard similarity results between research fields

In Figure 11, grey-filled items represent relations that have relatively high similarity values compared to others. Regarding the intra-relations of research topics belonging to the same category, we can identify high similarities between topics within the technique category and the support category. However, compared to these, the research topics in the application category have weak connectivity among each other. That is, applications in different domains apply different goals, techniques, and tools, leading to less intersection with each other.

Regarding the inter-relations of research fields between the different categories, we can identify that the topics in the technique and support categories have high similarities. On the contrary, it is hard to find meaningful relations between the application and other categories. This is because the extracted keywords for the application category are mainly relevant to a specific domain.

Next, we focus on the similarities between the individual research fields in the technique category. First, there is a strong relationship between process discovery (T1) and other fields, such as conformance checking (T2), performance-driven process mining (T3), process mining framework, guideline, & tool (T13), event logs (T14), and others (T17). This indicates that process discovery has played a pivotal role in process mining, linked to other research topics such as conformance checking, event logs, and performance-driven techniques. Also, process mining framework, guideline, and tool (T13) has high similarity values with not only process discovery (T1) but also conformance checking (T2), semantic process mining (T9), and process mining visualization (T15). Thus, various technical algorithms have been

developed to utilize a process mining framework or tool as a standard. In addition, it is noteworthy that there is a high similarity between trace clustering (T7) and variant analysis (T11) since both topics pursue the similar goal of analyzing processes at a fine-grained level.

In the application category, there is no notable identification of the relationships, whereas the research fields in the support show high similarities with multiple topics in the technique category. More in specific, a couple of research topics, e.g., machine learning & business intelligence (S1) and BPM & BPR (S2), show a high similarity with the technical algorithms, e.g., process discovery, organization mining, and process mining visualization. Also, BPM & BPR (S2) and process simulation (S3) have been studied with performance-driven process mining, indicating that it plays an essential role in implementing business processes.

5.2. Network derivation and community detection

Based on the results obtained in previous sections, we first discover a network representing relations of research fields. Figure 12(a) presents the derived network.

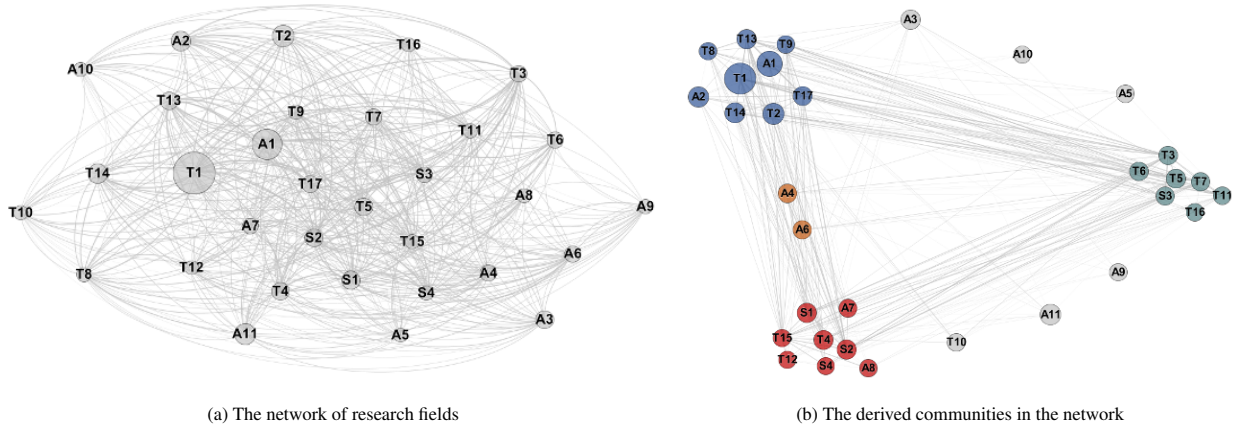


Figure 12: The results for network derivation and community detection

In the network, nodes indicate the research fields identified in Section 3, whereas edges indicate the similarity between the research fields measured in Subsection 5.1. Also, the node size indicates the number of publications belonging to each research field, while the thickness of the edge represents the degree of similarity. Based on the derived network, we can recognize that many studies have focused on process discovery (T1), conformance checking (T2), healthcare (A1), and software (A2) topics. Also, process discovery (T1) and process mining framework, guideline, and tool (T13) have been the central points to connect other research fields, as discussed in Subsection 5.1.

Next, we derive research communities by applying the Louvain algorithm on the basis of the network. Figure 12(b) depicts the derived communities. Applying the modularity delta threshold of 2.21 in discovering communities, we found four groups, e.g., representative, organizational, performance-centric, and manufacturing & IoT communities, and six individual topics.

- **Representative community** is composed of process discovery (T1), conformance checking (T2), streaming process mining (T8), semantic process mining (T9), process mining framework, guideline, & tool (T13), event logs (T14), others (T17), healthcare (A1), and software (A2). This community includes process discovery, conformance checking, healthcare, and software research fields where active research has been steadily conducted since the past. In addition, research on event logs and process mining frameworks, guidelines & tools are included in this community. Also, we can identify a couple of cutting-edge process mining techniques, including semantic process mining and streaming process mining. These are emerging research fields derived from process discovery and event log topics. Based on these results, we understand that this community represents process mining and consists of the most fundamental components of this discipline.
- **Resource-oriented community** is composed of organizational mining (T4), privacy-preserving process mining (T12), process mining visualization (T15), web and cloud services (A7), e-business (A8), machine learning and business intelligence (S1), Business process management and business process redesign (S2), and decision support (S4). This community represents the research topics from the business process management perspective, centered on resources and organizations. Specifically, it includes organizational mining and process mining visualization, e.g., social networks, in the technique category. In addition, a couple of support research fields, e.g., business intelligence and BPM, belong to this community, where resource-oriented data are actively utilized. Also, this community holds the web service and E-business fields in the application area.
- **Performance-oriented community** consists of performance-driven process mining (T3), process monitoring (T5), predictive process monitoring (T6), trace clustering (T7), variant analysis (T11), process model comparison (T16), and process simulation (S3). Unlike the prior two communities, the technical research topics comprise most of this community. In particular, performance-driven process mining plays a key role, and other relevant research topics are connected to this. For example, a series of connections of similar technologies, such as performance-driven process mining, process monitoring, and predictive process monitoring, show changes in technological trends and advancements. In addition, trace clustering, variant analysis, and process model comparison technologies have mutual similarities. In more detail, sub-event logs derived through trace clustering have been utilized as input for the variant analysis and process model comparison. Finally, process simulation aligns with process mining-based performance measurement and scenario analysis.
- **Manufacturing & IoT community** involves manufacturing (A4) and internet of things (A6). This community implies that sensor-based event logs in various domains, especially manufacturing environments, have been collected, and numerous case studies based on these have been conducted to analyze the end-to-end manufacturing processes.

Also, six individuals are not included in the communities: process model repair (T10), education (A3), manufacturing (A5), customer support services (A9), protocols (A10), and others (A11). Most of these research fields belong

to the application category, and they selectively utilize different algorithms to resolve problems for a specific domain. Also, the process model repair is independent of the communities because it has only been conducted recently and is slightly linked to many other topics.

6. Forecasting future trends of research fields in process mining

This section aims to analyze future trends for the research topics. To that end, we predict the number of publications in a research field for the upcoming years (e.g., the number of publications in T1, i.e., process discovery, for 2023) by leveraging 1) the historical publication records of the research field (e.g., the number of T1’s publications for the last three years) and 2) the historical records of other research fields that are highly correlated to the research field (e.g., the number of publications for conformance checking and performance-driven process mining).

6.1. Prediction Model

As presented in Subsection 2.5, we use LSTM models due to their robustness to learning temporal dependencies [39]. In the following, we describe the training procedure and validations of the model.

The LSTM model consists of 2 layers where the number of hidden states in the first layer is 2048 and the number of hidden states in the second layer is 4096. We train the model using the training data over 14 years with Adam optimizer algorithm [40] such that RMSE(Root Mean Squared Error) and MAE(Mean Absolute Error) between the predicted value $\hat{y}$ and the actual value y are minimized [41]. RMSE represents the root mean of squared errors (i.e., $(\hat{y} - y)^2$) as below:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (5)$$

MAE represents the mean of absolute errors (i.e., $|\hat{y} - y|$) as below:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (6)$$

To improve model generalizability, we use Dropout [42] and Batch Normalization [43].

To evaluate the performance of the trained model, we adopt time-series cross validation [44] since k-fold validation has limitations in validating time series data (e.g., due to its random selection of independent data points). We use the test data over four years. Table 3 shows the RMSE and MAE resulting from the time-series cross-validation.

Table 3: RMSE and MAE of the trained model that results from the time-series cross-validation

RMSE	MAE
5.63 ± 1.02	3.98 ± 0.55

6.2. Prediction Results

Based on the prediction model, we predict the number of publications in all the identified research fields for the upcoming two years. To that end, we first predict the number of publications in 2022 using the historical records and then predict the number of publications in 2023 using the predicted record in 2022 as well as the historical records.

Figure 13 shows the prediction results for the research fields in the technique category. For instance, the left-top chart depicts the predicted number of publications for T1 (process discovery) in 2022 and 2023, as well as the historical records from 2019 to 2020. The number of publications in 2019 is 29, decreasing to 22 in 2020 and increasing to 26 in 2021. The predicted number of publications in 2022 is around 29, and the one in 2023 is around 31, showing an increasing trend in the research field.

Most of the research fields in the technique category are forecasted to be strengthened in 2022 compared to 2021. In particular, T2 (conformance checking), T3 (performance-driven process mining), T5 (process monitoring), and T8 (streaming process mining) are predicted to show tremendous increases in the number of publications in 2022. However, a few research fields are anticipated to show decreasing advances with fewer publications, e.g., T11 (variant analysis) and T13 (process mining framework, guideline, & tool).

In 2023, most of the research fields are predicted to have fewer publications than in 2022. For instance, T2 (conformance checking), T3 (performance-driven process mining), T5 (process monitoring), T7 (trace clustering), and T8 (streaming process mining) are anticipated to have fewer publications in 2023 compared to 2022. However, such research fields still show the predicted number of publications higher than the average number of publications in the last three years, demonstrating the ongoing expansion of the research fields. For instance, the average number of publications of T2 from 2020 to 2022 is around 13, whereas the predicted number of publications in 2023 is above 14.

Figure 14 shows the prediction results for the research fields that belong to the application category. For instance, the chart on the left-top corner describes the predicted number of publications for A1 (healthcare) in 2022 and 2023, as well as the historical records of the last three years. The number of publications in 2019, 2020, and 2021 is 35, 52, and 19, respectively. The predicted number of publications in 2022 and 2023 is around 35 and 34, respectively.

Similar to the technique category, the research fields in the application category are anticipated to expand in terms of the number of publications in 2021 compared to the previous year. For instance, the number of publications in A1 (healthcare) is forecasted to increase from 19 in 2021 to around 35 in 2022, and the number of publications in A3 (education) is anticipated to increase from 10 to 12. The predicted number of publications in 2023 remains similar to 2022, only showing slight increases and decreases depending on research fields. The number of publications for A1 is predicted to decrease marginally in 2023 compared to 2022, whereas the predicted number of publications for A2 is slightly higher than the one in 2022.

Figure 15 describes the predicted number of publications for the research fields in the support category. For instance, the left-top chart shows the number of publications for S1 (machine learning and business intelligence) in

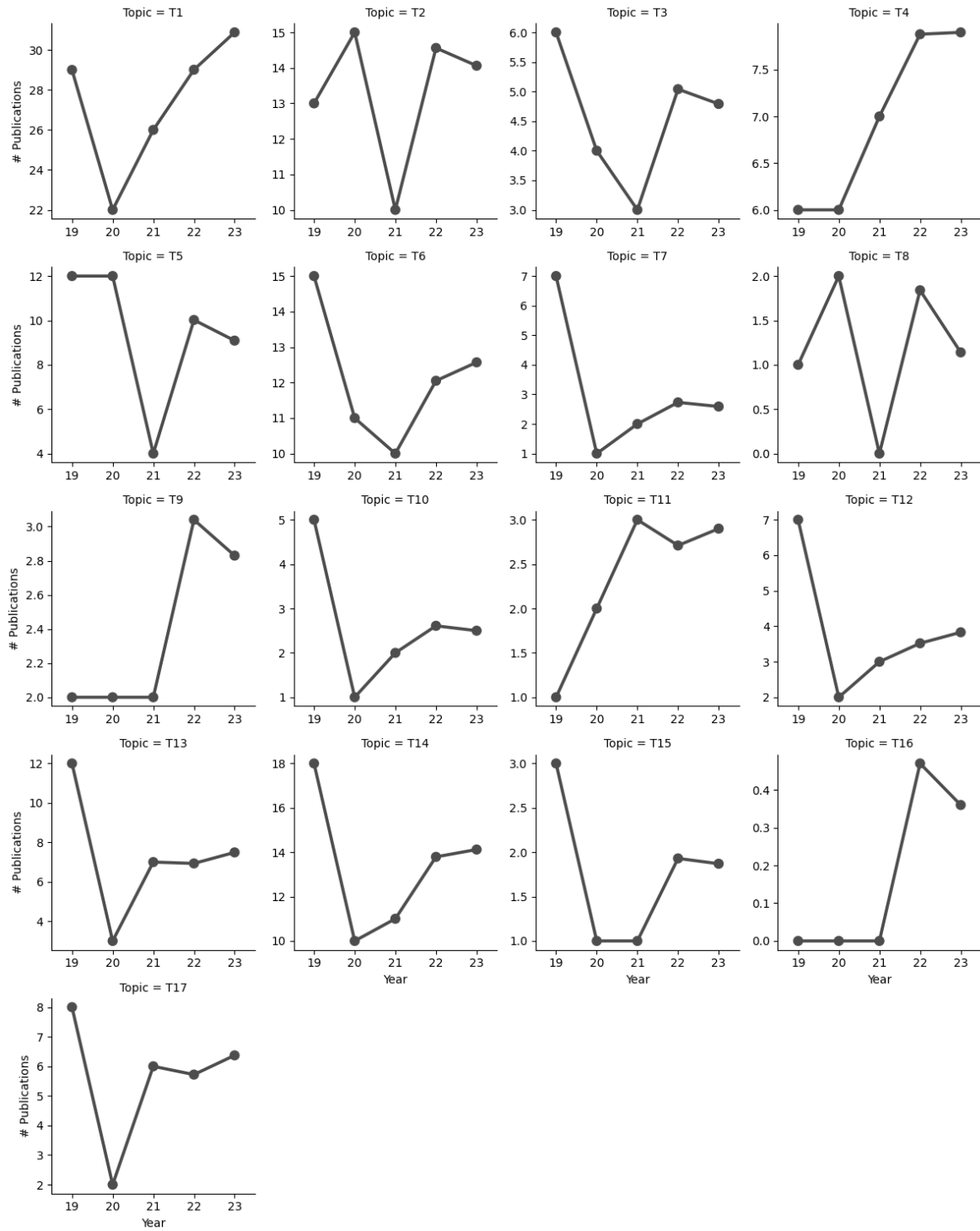


Figure 13: The predicted number of publications in 2022 and 2023 for the research fields of the technique category, along with the historical number of publications from 2019 to 2021

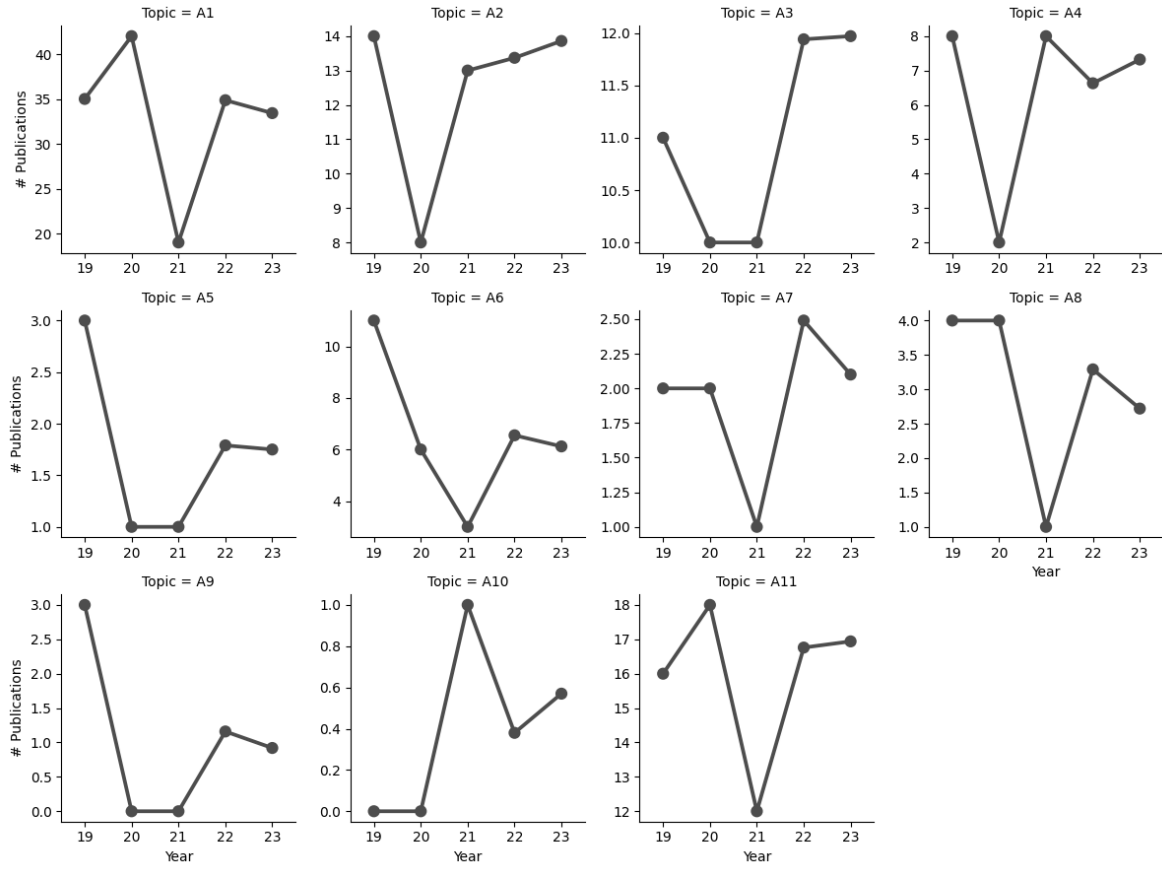


Figure 14: The predicted number of publications in 2022 and 2023 for the research fields of the application category, along with the historical number of publications from 2019 to 2021

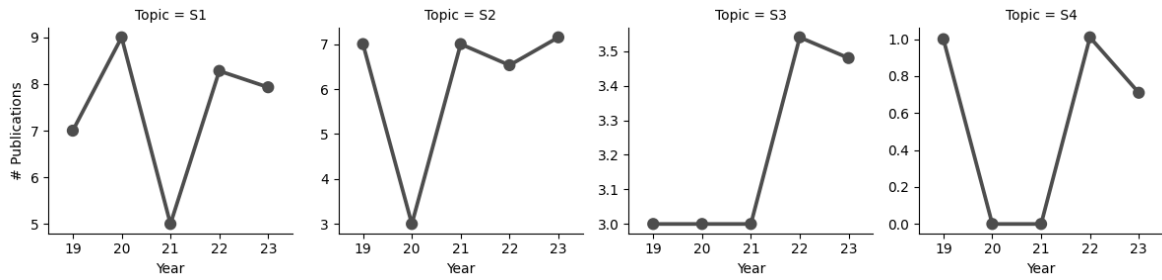


Figure 15: The predicted number of publications in 2022 and 2023 for the research fields of the support category, along with the historical number of publications from 2019 to 2021

2022 and 2023, along with the historical records of the last three years. The number of publications in 2019, 2020, and 2021 is 7, 9, and 5, respectively. The predicted number of publications in 2022 and 2023 is around 8 in both years.

The number of publications in S1 (machine learning and business intelligence) and S2 (business process management and business process reengineering) is predicted to increase in 2022 and remain similar in 2023. Although

the predicted number of publications in S3 (process simulation) and S4 (decision support) shows similar trends, the absolute number of publications remains nearly unchanged; the numbers of publications in S3 and S4 are predicted to be around 3 and 1, respectively.

7. Discussion

This section aims to answer the research questions raised in Section 1 by discussing the results presented in the earlier sections. Moreover, we compare the obtained knowledge with the existing trends in process mining both from academic and practical standpoints.

7.1. Answers to the research questions

Regarding the first research question, we identified and investigated research topics in process mining with a series of approaches, as presented in Section 3 and Section 4. The research topics in the technique, application, and support categories have been classified into the dominant, skyrocketing, and deep-cut based on history. Process discovery and conformance checking steadily and dominantly have occupied a large part in the process mining discipline. Also, popular domains in applying process mining have been healthcare, software, education, and manufacturing from the past to the present. It is noteworthy that some research fields, e.g., predictive process monitoring, concept drift, and privacy-preserving process mining, have emerged rapidly in recent years and are highly related to machine learning and big data. On the contrary, process model comparison techniques and protocol applications have not gained much attention in the process mining community. In addition, we should pay attention to the research fields on other techniques (T17) and other applications (A11), which include innovative research papers with novel ideas on emergent topics in process mining. For example, there have been articles in T17 about process mining connecting to key technologies, such as natural language processing, big data, and blockchains. Until now, only a couple of papers have been published for each topic, but they may constitute a new process mining research field in the next few years.

To resolve the second research question, we discovered the topic network and detected the communities of research fields in process mining, as presented in Section 5. As a result, process discovery has been placed at the central point in the network, which indicates that it has strong relationships with other topics. Besides, the research fields in the application category have much less resemblance with other topics compared to the fields in the technique and support. With the derived communities, we uncovered how the research fields have interacted with each other and how they have developed. The representative community demonstrated which process mining techniques have mainly integrated for process mining projects in most domains. In addition, regarding the performance-related community, we observed that performance-driven process mining had become the cornerstone for the emergence of the latest monitoring techniques, such as process monitoring and predictive process monitoring. The connection network between research topics covered in this study and the community derived through it has explained the emergence of new research fields from the existing fields. In other words, it is necessary to explore the possibility of deriving a new research field within process mining by utilizing a predictive network model such as a graph neural network.

Regarding the third question, the prediction model suggested that the topics in most of the research fields will be expanded in 2022 and 2023 with respect to the number of their publications. Especially most of the research fields will show quantum leaps in the number of publications in 2022 compared to 2021 when the number of publications was significantly lower than in the previous years. In 2023, the number of publications will remain mostly the same as in 2022, with different topics showing non-significant increases or decreases. However, a few research fields will show fewer publications in 2022 and 2023, such as T11 (variant analysis) and T13 (process mining framework, guideline & tool). Due to the changes in trends from the emergence of new techniques and the connectivity with other disciplines, it is hard to get accurate values for the next few years. However, our work contributes to understanding future technology trends and development potential.

7.2. Further discussions

In this subsection, we further discuss the obtained research fields and their future trends in relation to the existing trends in process mining, both from academic and practical standpoints. First, to relate the existing academic trends, we analyze the workshop topics in the renowned conferences presented in Section 4 such as ICPM, BPM, and CAiSE. As of 2022, the workshops held with ICPM, BPM, and CAiSE conferences are as follows.

- Business Process Intelligence (BPI)
- DEClarative DECision and Hybrid approaches to processes (DEC2H)
- Streaming Analytics for Process Mining (SA4PM)
- Quantitative Aspects of Process Mining (QAPM)
- Social and Human Aspects of Business Process Management (BPMS2)
- Leveraging Machine Learning in Process Mining (ML4PM)
- Artificial Intelligence for Business Process Management (AI4BPM)
- Responsible Process Mining (RPM)
- Event Data and Behavioral Analytics (EdbA)
- Process Querying, Manipulation, and Intelligence (PQMI)
- Data Quality and Transformation in Process Mining (DQT-PM)
- Process-Oriented Data Science for Healthcare (PODS4H)
- Education meets Process Mining (EduPM)
- Process Management in Digital Production (PMDiPro)

- Business Processes Meet Internet-of-Things (BP-meet-IoT)

BPI has covered the entire research area of process mining, including process discovery, conformance checking, etc. The other workshops are mapped to specific research fields that are derived from this study. DEC2H, which covers declarative process mining, belongs to a subtopic of process discovery (T1). Articles on organizational mining (T4), streaming process mining (T8), and privacy-preserving process mining (T12) has been published through BPMS2, SA4PM, and RPM, respectively. In addition, the performance-oriented community, consisting of performance-related topics such as performance-driven process mining (T3), process monitoring (T5), and predictive process monitoring (T6), has published articles in QAPM. As performance-driven prediction has become essential, there has been in-depth attention to workshops covering machine learning and artificial intelligence, such as ML4PM and AI4BPM. Moreover, research on process model repair (T10) and event logs (T14) have been discussed in PQMI, EdbA, and DQT-PM, respectively. There have also been some workshops focused on specific applications domains, including PODS4H (healthcare (A1)), Edu4PM (education (A3)), PMDiPro (manufacturing (A4)), and BP-meet-IoT (IoT (A6)).

Second, to relate the existing practical trends in process mining, we analyze the functionalities supported by the renowned process mining tools that are evaluated as leaders and star performers in the process mining market by the *Everest group report* [45]: *Celonis*, *Software AG*, *Minit*, and *UiPath*. Moreover, we analyze the functionalities provided by the most representative open-source tool in process mining, *ProM*; The tool is contributed by process mining researchers, and most of the techniques developed so far have been implemented in the tool. Furthermore, we include *Fluxicon Disco*, which continuously received much attention for its compact and practical functionalities.

Table 4 presents whether or not the tools provide the functionalities corresponding to the research fields presented in this paper. It is noticeable that all tools cover process discovery (T1) and performance-driven process mining (T13). In other words, process discovery is the most fundamental topic and technique in process mining, and it is worthwhile to present performance measurements on the discovered process model. In addition, conformance checking (T2) and process model comparison (T16) are the functions all tools provide except for Disco. Although process model comparison is classified as deep-cut, it is a practically relevant technique. Moreover, organizational mining (T4) is only covered by Celonis and Minit, which indicates that the control-flow perspective is more influential than the resource perspective. Also, most tools provide process monitoring (T5), e.g., anomaly detection with statistical analysis, whereas they have not yet implemented predictive process monitoring (T6) that requires machine learning and deep learning. Also, relatively new algorithms, such as streaming process mining (T8), semantic process mining (T9), and process model repair (T10), are yet to be implemented in most tools except Celonis. Likewise, process mining commercial tools still do not cover privacy-preserving process mining (T12).

8. Conclusion

In this paper, we analyzed the literature on process mining using text mining and machine learning techniques. First, we collected 2,207 process mining articles and conference proceedings from 2002 to 2021. Next, we analyzed

Table 4: Key functionalities of process mining tools

	ProM	Celonis	Software AG	MINIT	UIPath	DISCO
T1	✓	✓	✓	✓	✓	✓
T2	✓	✓	✓	✓	✓	
T3	✓	✓	✓	✓	✓	✓
T4	✓	✓		✓		
T5	✓	✓	✓	✓		
T6	✓	✓				
T7	✓					
T8	✓	✓				
T9	✓	✓				
T10	✓	✓				
T11	✓	✓	✓	✓	✓	✓
T12	✓					
T13	✓	✓	✓	✓	✓	✓
T14	✓	✓	✓	✓	✓	✓
T15	✓	✓	✓	✓	✓	✓
T16	✓	✓	✓	✓	✓	

the with a pre-trained model-based topic modeling, exploratory data analysis with key features, network and community detection, and future prediction with a deep learning model. As a result, we provided knowledge on the main research topics in process mining and their temporal trends, the relationships between the research topics, and the predictions on the expansion of the topics. Moreover, we compared the provided knowledge with the trends in both academics and practices.

Our work has several limitations. First, the presented quantitative analysis based on text mining and machine learning techniques still require some manual steps. For instance, the pre-processing of collected documents is manually conducted, while the identification of research fields requires qualitative evaluations based on expert knowledge. Second, the presented research topics are by no means exhaustive. We only focus on the topics identified by text mining algorithms, and thus, some of the essential topics may not be a part of the further analysis, e.g., network detection, community detection, yearly patterns, etc. For instance, object-centric process mining receives much attention from the community as an essential research topic, but it is not part of our analysis since it is not identified as a research topic in the topic modeling algorithm. Finally, since our analysis relies on historical data, we cannot present topics that do not exist in history. The discovery and analysis of newly-emerging topics remain as the judgments of readers.

References

- [1] I. M. Sebastian, J. W. Ross, C. Beath, M. Mockler, K. G. Moloney, N. O. Fonstad, How big old companies navigate digital transformation, *MIS Quarterly Executive* 16 (3) (2017) 6.
- [2] C. Matt, T. Hess, A. Benlian, Digital transformation strategies, *Business & information systems engineering* 57 (5) (2015) 339–343.
- [3] P. C. Verhoef, T. Broekhuizen, Y. Bart, A. Bhattacharya, J. Q. Dong, N. Fabian, M. Haenlein, Digital transformation: A multidisciplinary reflection and research agenda, *Journal of Business Research* 122 (2021) 889–901.
- [4] J. vom Brocke, M. Jans, J. Mendling, H. A. Reijers, A five-level framework for research on process mining, *Business & Information Systems Engineering* 63 (5) (2021) 483–490.
- [5] T. Grisold, J. Mendling, M. Otto, J. vom Brocke, Adoption, use and management of process mining in practice, *Business Process Management Journal* (2020).
- [6] L. Reinkemeyer, *Process mining in action, Process Mining in Action Principles, Use Cases and Outlook* (2020).
- [7] W. M. P. van der Aalst, *Process Mining - Data Science in Action*, Second Edition, Springer, 2016.
- [8] A. Augusto, R. Conforti, M. Dumas, M. La Rosa, F. M. Maggi, A. Marrella, M. Mecella, A. Soo, Automated discovery of process models from event logs: review and benchmark, *IEEE transactions on knowledge and data engineering* 31 (4) (2018) 686–705.
- [9] A. Rozinat, W. M. P. van der Aalst, Conformance checking of processes based on monitoring real behavior, *Information Systems* 33 (1) (2008) 64–95.
- [10] W. M. P. van der Aalst, M. H. Schonenberg, M. Song, Time prediction based on process mining, *Information systems* 36 (2) (2011) 450–475.
- [11] I. Teinemaa, M. Dumas, M. L. Rosa, F. M. Maggi, Outcome-oriented predictive process monitoring: review and benchmark, *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13 (2) (2019) 1–57.
- [12] W. M. P. van der Aalst, H. A. Reijers, A. J. Weijters, B. F. van Dongen, A. A. De Medeiros, M. Song, H. Verbeek, Business process mining: An industrial application, *Information Systems* 32 (5) (2007) 713–732.
- [13] W. M. P. van der Aalst, J. Carmona, *Process mining handbook* (2022).
- [14] A. Rozinat, R. S. Mans, M. Song, W. M. P. van der Aalst, Discovering simulation models, *Information systems* 34 (3) (2009) 305–327.
- [15] M. Cho, M. Song, M. Comuzzi, S. Yoo, Evaluating the effect of best practices for business process redesign: An evidence-based approach based on process mining techniques, *Decision Support Systems* 104 (2017) 92–103.
- [16] V. Naderifar, S. Sahran, Z. Shukur, A review on conformance checking technique for the evaluation of process mining algorithms, *TEM Journal* 8 (4) (2019) 1232.
- [17] S. J. van Zelst, F. Mannhardt, M. de Leoni, A. Koschmider, Event abstraction in process mining: literature review and taxonomy, *Granular Computing* 6 (3) (2021) 719–736.
- [18] D. M. V. Sato, S. C. De Freitas, J. P. Barddal, E. E. Scalabrin, A survey on concept drift in process mining, *ACM Computing Surveys (CSUR)* 54 (9) (2021) 1–38.
- [19] E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, Process mining in healthcare: A literature review, *Journal of biomedical informatics* 61 (2016) 224–236.
- [20] A. Bogarín, R. Cerezo, C. Romero, A survey on educational process mining, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8 (1) (2018) e1230.
- [21] A. Corallo, M. Lazoi, F. Striani, Process mining and industrial applications: A systematic literature review, *Knowledge and Process Management* 27 (3) (2020) 225–233.
- [22] P. Zerbino, A. Stefanini, D. Aloini, Process science in action: A literature review on process mining in business management, *Technological Forecasting and Social Change* 172 (2021) 121021.
- [23] S. Kim, H. Park, J. Lee, Word2vec-based latent semantic analysis (w2v-lsa) for topic modeling: A study on blockchain technology trend analysis, *Expert Systems with Applications* 152 (2020) 113401.
- [24] T. K. Landauer, P. W. Foltz, D. Laham, An introduction to latent semantic analysis, *Discourse processes* 25 (2-3) (1998) 259–284.
- [25] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *Journal of machine Learning research* 3 (Jan) (2003) 993–1022.

- [26] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).
- [27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [28] G. T. Reddy, M. P. K. Reddy, K. Lakshmana, R. Kaluri, D. S. Rajput, G. Srivastava, T. Baker, Analysis of dimensionality reduction techniques on big data, IEEE Access 8 (2020) 54776–54788.
- [29] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, C.-T. Lin, A review of clustering techniques and developments, Neurocomputing 267 (2017) 664–681.
- [30] J. Scott, What is social network analysis?, Bloomsbury Academic, 2012.
- [31] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! keyword extraction from single documents using multiple local features, Information Sciences 509 (2020) 257–289.
- [32] S. R. El-Beltagy, A. Rafea, Kp-miner: A keyphrase extraction system for english and arabic documents, Information systems 34 (1) (2009) 132–144.
- [33] B. Škrlj, A. Repar, S. Pollak, Rakun: Rank-based keyword extraction via unsupervised learning and meta vertex aggregation, in: International Conference on Statistical Language and Speech Processing, Springer, 2019, pp. 311–323.
- [34] G. Erkan, D. R. Radev, Lexrank: Graph-based lexical centrality as salience in text summarization, Journal of artificial intelligence research 22 (2004) 457–479.
- [35] Y. Sun, H. Qiu, Y. Zheng, Z. Wang, C. Zhang, Sifrank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model, IEEE Access 8 (2020) 10896–10906.
- [36] S. Bag, S. K. Kumar, M. K. Tiwari, An efficient recommendation generation using relevant jaccard similarity, Information Sciences 483 (2019) 53–64.
- [37] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of statistical mechanics: theory and experiment 2008 (10) (2008) P10008.
- [38] P. De Meo, E. Ferrara, G. Fiumara, A. Provetti, Generalized louvain method for community detection in large networks, in: 2011 11th international conference on intelligent systems design and applications, IEEE, 2011, pp. 88–93.
- [39] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: Lstm cells and network architectures, Neural computation 31 (7) (2019) 1235–1270.
- [40] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
- [41] C. J. Willmott, K. Matsuura, Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, Climate research 30 (1) (2005) 79–82.
- [42] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
- [43] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F. R. Bach, D. M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015, Vol. 37 of JMLR Workshop and Conference Proceedings, JMLR.org, 2015, pp. 448–456.
- [44] C. Bergmeir, J. M. Benítez, On the use of cross-validation for time series predictor evaluation, Inf. Sci. 191 (2012) 192–213.
- [45] A. Modi, H. K. Makan, S. Kumar, A. Nath, Process mining – technology provider landscape with products peak matrix assessment 2022, Tech. rep., Everest Group (2022).