Ingénierie des Systèmes d'Information
Vol. 28, No. 5, October, 2023, pp. 1409-1421
Journal homepage: http://iieta.org/journals/isi

Employing Data and Process Mining Techniques for Redundancy Detection and Analytics in Business Processes

Fatima Zohra Trabelsi*, Amal Khtira, Bouchra El Asri

IT Architecture and Model Driven Systems Development Team, Advances Digital Entreprise Modeling and Informatique Retrieval Laboratory, Rabat IT Center, Ecole Nationale Superieure d'Informatique et d'Analyse des Systèmes, Mohammed V University, Rabat 10080, Morocco

Corresponding Author Email: fatimazohra.trabelsi@hotmail.fr

https://doi.org/10.18280/isi.280529

Received: 13 April 2023
Revised: 12 June 2023
Accepted: 13 July 2023
Available online: 31 October 2023

Keywords:
business process, modeling, process mining, similarity detection, data mining, process redundancy, business process optimization, cost reduction

ABSTRACT

The detection, quantification, and scrutiny of redundancies within business processes is pivotal in achieving cost reduction, enhancing efficiency, and ensuring compliance. Redundancies, often leading to inefficiencies, result in escalated costs and errors, thereby detrimentally influencing an organization's overall performance. To counter these issues, data mining and process mining techniques offer promising solutions by identifying and analyzing process redundancies. Data mining, an approach devoted to the analysis of large datasets in order to discern patterns, relationships, and anomalies, has been applied to business processes. It provides insights into redundancies by scrutinizing process-related data, such as event logs, thereby revealing patterns in task executions that may indicate redundancies. In contrast, process mining employs event logs to generate a process model mirroring the actual execution of a process. This actual process model is subsequently contrasted against an expected process model, facilitating the identification of redundancies such as unnecessary activities or loops. Cluster analysis, a technique employed in both data mining and process mining, is exemplified for its capacity to group similar process instances or models based on specified attributes or characteristics. The application of cluster analysis aids in the identification of redundant process models or similar process patterns, thereby enabling further comparison and optimization.

1. INTRODUCTION

Similarity search, the practice of comparing a particular object or process to a collection of objects to identify those exhibiting close resemblance, has evolved into a contemporary topic bolstering business process management [1]. The detection of similarities in business processes facilitates the identification of redundant activities, thereby optimizing resources [2].

In the field of recommendation systems, the similarity between users significantly influences the calculation of recommended items [3], enhancing the performance of the recommendation [4]. A common approach involves the calculation of similarity measurements between pairs of users or items [5]. However, the measurement of user similarity can vary when considering different classes of items, broadening our understanding of similarity measurement [6].

Similarity search algorithms streamline the search process by substantially reducing the number of comparison operations [7]. For instance, behavioral similarity can be measured using these algorithms, supporting decision-making and process improvement endeavors across various fields. Process mining can automate these similarity searches [8], providing valuable insights into how processes are executed within an organization. This technique finds utility in the analysis of a range of processes, from simple workflows to complex business processes involving multiple systems and actors, with applications spanning healthcare, logistics, finance, and manufacturing [9].

Business process modeling, which involves the creation of visual representations of the steps, activities, and resources necessary for the execution of a business process, constitutes a critical component of business process management (BPM) [10]. BPMN, a widely used graphical description language in process modeling, particularly in business process similarity measurement and process mining, offers a notation for the modeling and design of business processes [11]. These processes are subsequently analyzed and enhanced using process mining methods, allowing for a more comprehensive understanding and improvement of business processes [12].

This study presents a framework based on data mining and process mining to detect and analyze redundancy in business processes. Herein, process mining and BPMN (Business Process Model and Notation) have been integrated to enhance redundancy detection and analysis. Process mining techniques are employed to extract valuable insights from event logs, elucidating the actual execution of processes. These insights inform the creation of accurate business process models using BPMN, a standardized notation for representing process flows and dependencies.

Linking process mining with BPMN enables a seamless flow of information and analysis, thereby facilitating the effective identification of process redundancies. Process mining techniques unearth deviations and inefficiencies, which are subsequently reflected in the BPMN models. This integrated approach allows us to pinpoint areas of redundancy and make informed decisions for process optimization during business process reengineering.

The paper is organized as follows: Section 2 delves into the background, focusing on pivotal similarity detection techniques and similarity measurement methods. Section 3 outlines modeling, detecting, and measuring duplications in processes. Section 4 discusses the automation of similarity detection represented in a framework that includes the behavior method and the chosen algorithm. Section 5 presents a case study application of our algorithm. Finally, Section 6 provides the conclusion of the paper.

2. BACKGROUND

This section encompasses the foundational elements of this work, subdivided into two critical areas: detection techniques for similarity [13] and measurements of similarity [14].

2.1 Techniques for detecting similarity

Business Process (BP) management embraces a plethora of techniques for detecting and analyzing similarity [15]. These include but are not limited to process mining, business process reengineering, and business intelligence analytics.

2.1.1 Process mining

Process mining [5], a discipline within data science, utilizes facets of data mining and machine learning to analyze event logs, thus extracting information regarding the actual execution of processes within an organization [16]. The analysis of digital footprints such as event logs generated by information systems, offers insights into actual process flow, performance, and behavior [17].

The ultimate objective of process mining [18] entails the presentation of visual representation of process flow, pinpointing of bottlenecks, inefficiencies, and the proposal of improvement strategies to optimize the process. Redundancies within processes are identified through the analysis of event logs and process models, thus detecting repetitive or unnecessary steps. An analysis of activity sequences within event logs can expose a specific set of activities that, while frequently followed, could potentially be eliminated or abbreviated without detriment to the process outcome. This highlights a redundancy that warrants attention.

2.1.2 Business process reengineering

Business Process Reengineering [19], a management technique, is employed to redesign and enhance business processes. It involves a critical analysis of existing processes and the identification of areas requiring improvement, followed by a redesign of said processes [20]. The elimination of redundancies and inefficiencies is the key goal, leading to the design of new processes [21] that surpass the efficiency and effectiveness of pre-existing ones.

A critical examination of existing processes reveals inefficiencies and redundancies. Subsequent redesigning of these processes seeks to eradicate these issues [22]. Redundancies in the process are identified through a thorough analysis of existing processes, aiming to eliminate inefficiencies and streamline operations.

2.1.3 Business intelligence analytics

Business Intelligence Analytics [23] involves the use of data analytics tools to scrutinize process data, thereby identifying areas ripe for improvement. Data from a variety of sources, including process data, customer data, and financial data, can be analysed [24, 25], with the purpose of identifying redundancies and inefficiencies in the process. Insights into process performance can be obtained, highlighting opportunities for improvement [26].

Process redundancies are identified through the use of data analysis techniques and visualizations, which reveal patterns and anomalies in process data. By analyzing key performance indicators (KPIs) and metrics related to process efficiency and resource utilization, redundancies can be detected and targeted for process improvement initiatives.

Process mining provides an objective analysis and visualization of business processes based on actual event logs and data, identifying deviations and bottlenecks in real-time. However, it is reliant upon data quality and may not offer insights into the root causes of process similarities. Conversely, business process reengineering focusses on redesigning processes for heightened efficiency but may fall short in terms of data-driven analysis and real-time monitoring capabilities. Business intelligence analytics, while powerful for analyzing and reporting on historical data, may not provide the same level of process-specific insights as the aforementioned methods.

Each approach has its unique strengths and limitations in detecting similarities within business processes, hence the necessity for a comprehensive and integrated approach to process improvement. In essence, data-driven techniques such as process mining and Business Intelligence (BI) analytics can augment Business Process Reengineering (BPR) by providing valuable insights and supporting evidence-based decision-making.

The integration of process mining and BI analytics enhances the BPR approach by enabling organizations to accumulate comprehensive process data, visualize process performance, and pinpoint specific areas where redundancies are present. These techniques offer evidence-based insights that facilitate informed decision-making, aiding organizations in the optimization of their processes more effectively.

2.2 Techniques for measuring similarity

The principal bulk of similarity measurement methodologies rely on the labelling of activities [2]. It is critical to note that the selection of an appropriate similarity measure technique is contingent upon the unique characteristics of the process models in addition to the objectives of the similarity detection. To evaluate processes and ascertain their similarities, the following are the most commonly adopted techniques:

2.2.1 Graph edit distance

Graph edit distance, as expounded [1, 27, 28], serves as a prominent method for gauging the similarity between pairs of graphs in a tolerably erroneous manner. This method has been extensively utilized in pattern analysis and recognition. It calculates the minimum number of edit operations (such as insertion, deletion, and substitution) required to morph one process model into another. Despite the comprehensive measure of similarity between graphs it offers, considering both structural and attribute differences, it can be computationally demanding and sensitive to minor alterations in graphs.

2.2.2 Jaccard similarity

Jaccard similarity, as articulated [6, 14, 29], is a simple yet efficient technique that is applicable to a multitude of datasets, ranging from compact data sets to voluminous and complex datasets. It calculates the similarity between two sets of elements (in this case, process models) based on the ratio of their intersection to their union. Though simple and efficient, it solely emphasizes the elements shared between graphs, neglecting structural information.

2.2.3 Structural similarity

As suggested in studies [2, 30, 31], structural similarity pertains to the degree of similarity between two or more structures. This measure contrasts process models based on their structure, such as the number of tasks, the number of flows, and the task connectivity. While this method successfully captures the topological similarity between graphs, enabling the identification of similar patterns and structures, it may not take attribute differences into account.

2.2.4 Behavioral similarity

Behavioral similarity, as detailed in studies [5, 19, 32], involves the comparison of the sequences of events or activities that transpire during the execution of different processes to pinpoint patterns and similarities. It compares process models based on the process behavior, such as the sequence of tasks, the conditions for task execution, and the dependencies between tasks. By identifying similar behavior patterns, redundancies can be detected where unnecessary or repetitive activities occur.

In the present study, a fusion of process mining and behavioral similarity analysis is employed. Process mining techniques are deployed to extract behavior patterns from event logs and subsequently, behavioral similarity analysis is applied to measure the similarity between different process instances. This approach facilitates a comprehensive correlative link between the two methods for the identification of process redundancies.

3. MODELING, DETECTING AND MEASURING SIMILARITY IN PROCESSES

In this third section, we have shared it into three main parts, the different techniques and steps of the process mining, the BPMN model for enhancing the modelization and automating the detection of the duplication and the last part is about the method we have chosen for detecting similarities which is the behavior similarity technique.

3.1 Process mining techniques

Process mining techniques [8] are a set of methods that use event data from process execution to discover, analyze, and improve business processes. Some examples of process mining techniques [9] are:

Process Discovery [23]: This technique automatically extracts a process model from event data, such as logs of system events or process execution data. It can be done using different algorithms, like the Alpha Algorithm, the Heuristics Miner or the Inductive Miner.

Conformance Checking [16]: This technique compares an actual process, as it is executed, to a model of the process to detect deviations and identify potential issues.

Performance Analysis [22]: This technique analyzes process execution data to identify bottlenecks, delays and inefficiencies in the process.

Process Enhancement [18]: This technique uses process discovery and performance analysis results to improve the process by redesigning or restructuring it, or by identifying and implementing process improvements.

Process Simulation [33]: This technique uses process models to simulate process execution and evaluate the performance of different process variants.

Social Network Analysis [17]: This technique uses process execution data to analyze the collaboration and communication patterns among process participants.

Anomaly Detection [34]: This technique uses process execution data to detect abnormal or unusual behaviors in the process, which can indicate process deviations or potential issues.

Overall, process mining techniques are used to improve the efficiency and effectiveness of business processes by providing a detailed understanding of how the processes are actually executed, and by identifying potential process improvements.

The Figure 1 gives an overview of the process mining.

3.2 BPMN model

Process mining using BPMN (Business Process Model and Notation) [10] models is a popular approach for automating the detection of similar business processes. BPMN is a standard graphical notation, as shown in Figure 2, used to model business processes and is widely used in organizations. By using BPMN models [35] in process mining, organizations can leverage a widely adopted graphical notation to model business processes and perform process analysis. This approach allows for the identification of similar processes and the optimization of business processes to improve overall performance.

3.3 Similarity behaviour

Behavior similarity techniques [12] are used because they are specifically designed to compare and measure the similarity of behavior patterns within a process. Other techniques can also be used to analyze processes, but they may not be as effective at identifying similarities between processes.

In our work, we have chosen the Behavior similarity technique, because it is particularly useful when dealing with processes that have many steps or stages, where there may be different paths or variations in the process flow. By analyzing the behavior patterns of these processes, behavior similarity techniques can identify commonalities and differences between them, even if the processes are not identical.

Additionally, behavior similarity techniques are often used in conjunction with process mining techniques, which rely on event logs or other process data to create process models.

By analyzing the behavior patterns of these process models, behavior similarity techniques can identify similarities and differences between the actual process flows and the modeled process flows. Finally, in data mining, behavior similarity is used to identify patterns and similarities in the data, to compare the behavior of different process instances and to cluster similar data points together. This can be useful for identifying redundancies or inefficiencies in the processes, or for identifying groups of data points that exhibit similar behavior.

Integrating behavior similarity analysis with process mining and BPMN, helps in identifying redundant process patterns. Process mining extracts behavior patterns from event logs, which are then represented in BPMN models. Behavior similarity analysis compares these patterns to identify similarities across different process instances or models. If similar behavior patterns are found, it suggests potential redundancies in the process. This allows organizations to focus on optimizing and streamlining specific areas, resulting in improved process efficiency and effectiveness.

The proposed case study showcases how the combination of process discovery, BPMN modeling, conformance checking, and behavior similarity analysis in process mining can assist in identifying and eliminating redundant activities. These techniques help in optimizing processes, improving efficiency, and enhancing patient-centric care in real-world healthcare settings.

Scenario: A hospital wants to optimize their patient admission process to eliminate redundancies and improve efficiency.

Process Discovery and BPMN Modeling: The hospital collects event logs capturing the activities performed during the patient admission process. Process mining techniques are applied to these event logs, resulting in a visual representation of the process flow. The discovered process model is translated into a BPMN (Business Process Model and Notation) diagram, providing a standardized notation for representing the process flow, activities, decisions, and dependencies.

Conformance Checking: The hospital compares the BPMN model derived from the process discovery with the intended or reference BPMN model, which represents the ideal patient admission process without redundancies. By conducting conformance checking, the hospital identifies deviations between the actual BPMN model and the reference model. Redundant activities that are not part of the reference model are highlighted as potential areas for improvement.

Behavior Similarity Analysis: In addition to conformance checking, behavior similarity analysis is applied to the event logs and BPMN models. It compares the behavior patterns and sequences of activities across multiple process instances to detect similarities and redundancies. The analysis identifies repetitive or unnecessary steps in the patient admission process, such as duplicate data entry or redundant checks.

Identification of Redundancies: Through the combination of conformance checking and behavior similarity analysis, the hospital identifies redundant activities in the patient admission process. For example, it is discovered that multiple departments perform similar checks and verifications separately, leading to duplicated efforts and delays. By recognizing these redundancies in the BPMN model and event logs, the hospital decides to redesign the process, consolidating and streamlining these activities into a single step, eliminating redundancies, and improving overall process efficiency.

Outcome: By utilizing process discovery, BPMN modeling, conformance checking, and behavior similarity analysis, the hospital successfully identifies and addresses redundancies in the patient admission process. This optimization effort leads to streamlined operations, reduced delays, and improved patient experiences.

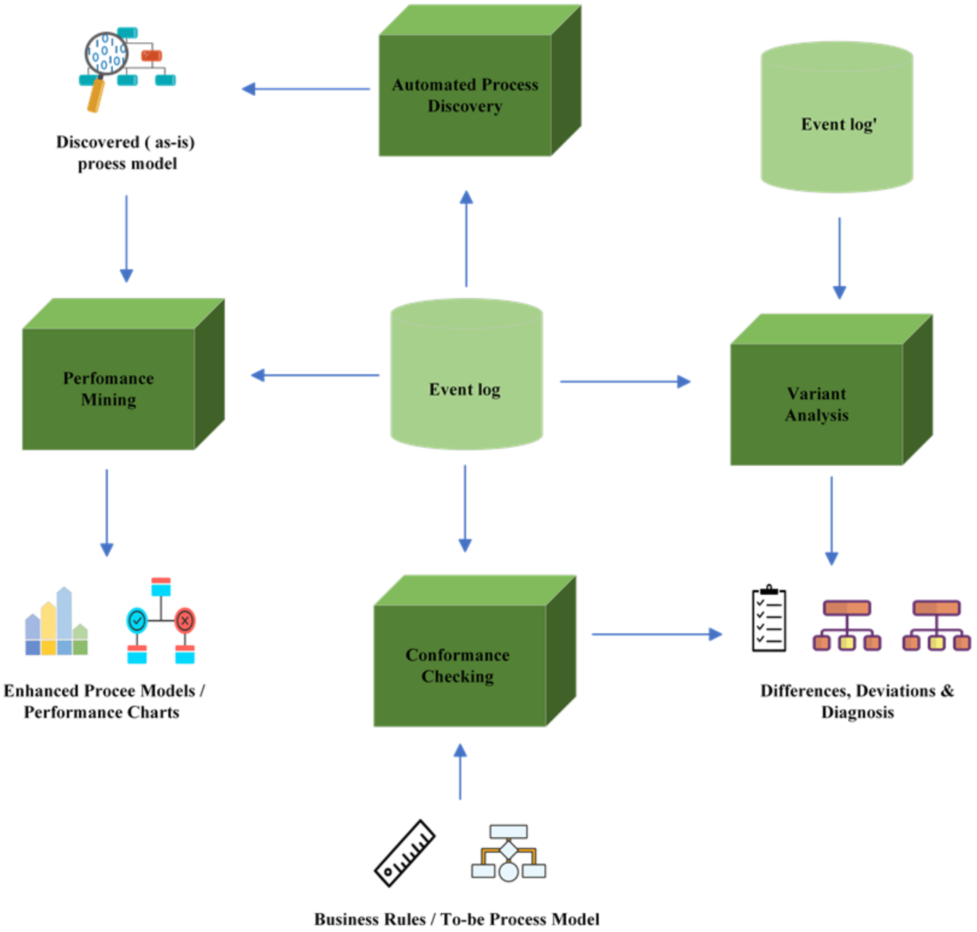

Figure 1. An overview of the process mining

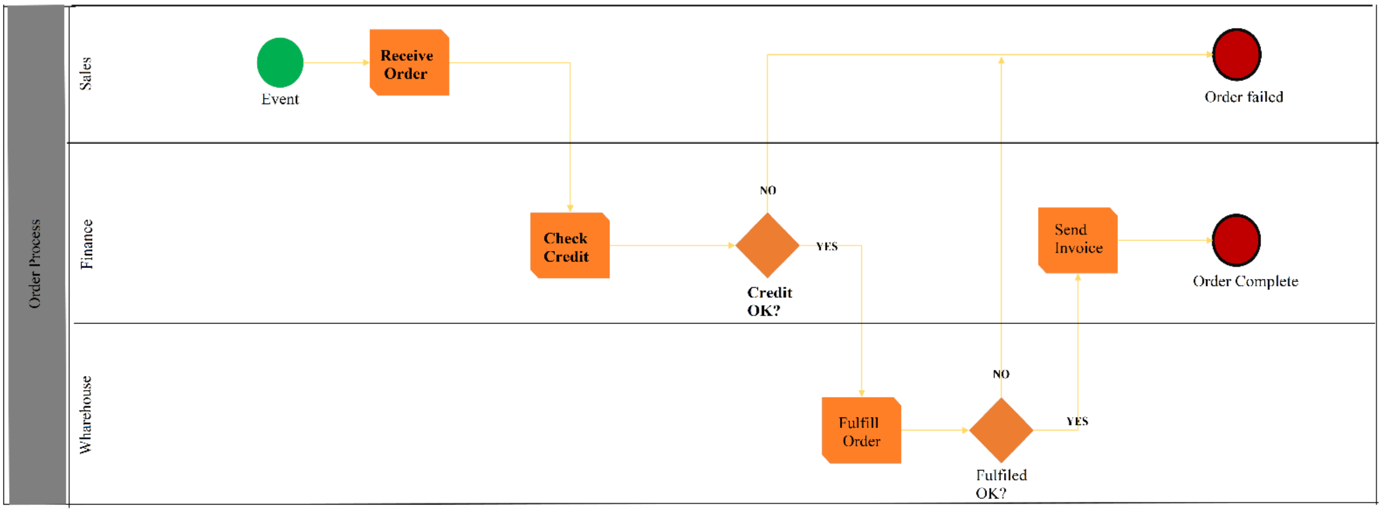

Figure 2. An example of a BPMN diagram

4. FRAMEWORK FOR AUTOMATING THE DETECTION OF SIMILARITY

By automating the detection of similar business processes, organizations can identify inefficiencies, optimize processes, and improve overall performance.

In this part, we have focused on applying one of the similarity measures techniques in a framework that we propose in the second part, which is the behavior method, in order to eliminate redundant and repeated processes.

In our paper, we proposed an efficient framework that helps in identifying and eliminating the redundant processes.

The proposed framework as shown in Figure 3 provides a structured approach to process optimization and can help organizations achieve greater efficiency, because by identifying and eliminating redundant processes, organizations can streamline their workflows and reduce unnecessary effort and resources, resulting in increased efficiency and productivity. It can also help in cost savings, because removing redundant processes can result in cost savings by reducing the need for additional staff, equipment, or materials. And improving performance like decision-making, due to a better insight into their processes and data, organizations can make more informed decisions that are based on reliable and accurate information. It can also help for a better customer satisfaction through streamlined processes that can result in faster and more accurate responses to customer requests, leading to higher levels of satisfaction. Improvement performances can also include Enhancement of competitiveness, by means of optimizing processes and reducing redundancies, organizations can become more agile and better equipped to compete in their market.

4.1 Data collection and preparation

Data collection is the process of gathering raw data from various sources, such as databases, files, and web services, and storing it in a central location. The collected data is usually in an unstructured or semi-structured format and may need to be transformed before it can be analyzed.

For transforming this data, we use the ETL process [36]. ETL stands for Extract, Transform, and Load. It is a process of integrating data from various sources, transforming it into a format that can be analyzed, and loading it into a data warehouse. The following steps are involved in the ETL process [37]:

Extract: In this step, data is extracted from various sources such as databases, files, APIs, etc. The data is collected and retrieved in its raw format Transform: The extracted data undergoes transformation to ensure its quality and consistency. This step involves cleaning the data, handling missing values, removing duplicates, standardizing formats, and performing calculations or derivations if needed.

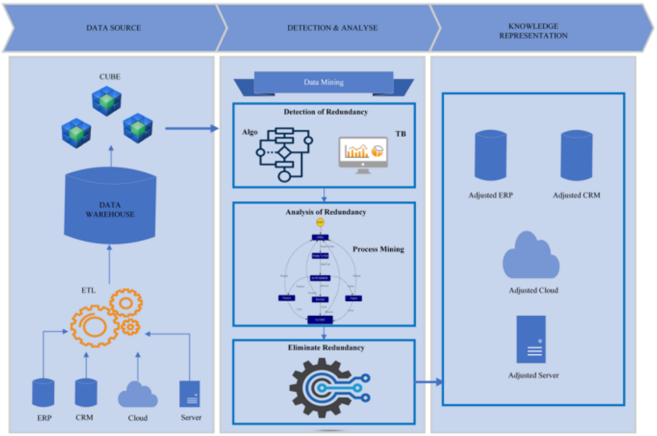

Figure 3. The proposed framework

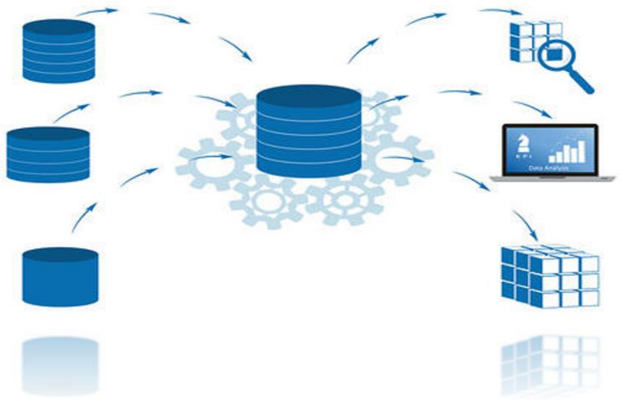

Figure 4. ETL process

Load: The transformed data is loaded into a target system, such as a data warehouse or a database, for further analysis and reporting. This step involves mapping the transformed data to the target schema and ensuring data integrity.

The previous Figure 4 presents a details diagram of an ETL process.

Data preparation involves transforming the collected data into a format that is ready for analysis. And to do this, we opt for the data warehousing [38], which is the process of storing data from various sources in a central repository, called a data warehouse [39]. A data warehouse is optimized for analytical processing and is designed to support complex queries and analysis [40]. Data warehousing involves the process of consolidating data from different sources into a centralized repository for analysis and reporting. Here are the main steps:

Designing the Data Warehouse Schema: This step involves designing the structure of the data warehouse, including defining dimensions and fact tables. Dimensions represent the various attributes for analysis, such as customer, product, time, and location, while fact tables contain the numerical measurements.

The data collection, preparation, and warehousing steps play a crucial role in enabling the algorithm to detect and eliminate redundancies. Here's how these steps contribute to the algorithm's effectiveness. Data collection involves gathering relevant data from various sources. This step ensures that the algorithm has access to a comprehensive dataset that represents the business processes or systems under analysis. The quality and completeness of the collected data directly impact the accuracy and reliability of the algorithm's results. Data preparation involves cleaning, transforming, and harmonizing the collected data to make it suitable for analysis. This step addresses issues such as missing values, inconsistent formats, outliers, and duplicates. By preparing the data, you improve the algorithm's ability to identify and handle redundancies effectively. Data warehousing involves consolidating the prepared data into a central repository optimized for analysis. The data warehouse provides a structured and organized environment for efficient data retrieval and processing. It allows the algorithm to access and analyze the data quickly, which is essential for detecting redundancies across different dimensions and perspectives.

By leveraging the data warehouse, the algorithm can efficiently explore relationships, patterns, and inconsistencies in the data. It can perform advanced analytics, such as data mining and process mining, to identify redundancies based on predefined rules, similarity measures, or behavioural patterns.

The algorithm utilizes the enriched and structured dataset stored in the data warehouse to compare and evaluate different processes, identify similarities, and detect redundancies across various dimensions. It can uncover redundant steps, duplicated workflows, or overlapping activities that can be optimized or eliminated to improve efficiency and effectiveness.

The data collection, preparation, and warehousing steps provide the necessary foundation for the algorithm to access comprehensive, clean, and structured data. This enables accurate analysis, pattern detection, and identification of redundancies, leading to effective process optimization and improvement.

Populating the Data Warehouse: The transformed data is loaded into the data warehouse tables. This involves extracting relevant data from the source systems and mapping it to the appropriate dimensions and fact tables.

And the final inputs are the cubes [41] that stores data in a multidimensional format, allowing for fast and efficient analysis of large datasets. Cubes are typically used in business intelligence and data analysis applications [42].

Creating a data cube allows for efficient multidimensional analysis and aggregation. Here's an overview of the steps:

Defining Dimensions: Identify the dimensions relevant to the analysis and define their hierarchies. Dimensions represent the different perspectives or attributes based on which analysis can be performed, such as customer, product, time, and location. Hierarchies define the levels of granularity within each dimension.

Defining Measures: Determine the numerical measures to be analysed and aggregated, such as sales revenue, quantity sold, or average customer rating.

Aggregating Data: Aggregating data involves pre-calculating summary statistics and aggregating the data at different levels of granularity based on dimensions and measures. This helps in efficient analysis and reporting by providing quick access to aggregated values.

4.2 Identification and detection of redundant processes

Using data mining algorithms is one of the best and most efficient way to identify and detect duplicated processes. There are several algorithms [43] that can be used like, Association rule mining [44] that identifies relationships between different processes, Clustering [45]: Which groups processes together based on their similarity and Decision tree [46] analysis that builds a tree-like model of decision rules that can be used to identify redundant processes.

In this step, we have proposed an algorithm with different steps to work with that involves the behavior similarity technique as a way of detecting redundancies and duplications.

In this algorithm, the focus is on identifying orders with similar behavior, rather than orders with identical values in certain fields.

This allows you to capture more complex patterns of duplication and reduce the risk of false positives or false negatives. To do so, we have chosen to associate the association rule mining to the clustering algorithm for having a new deduplication algorithm, in order the performance and the accuracy.

Firstly, a certain number of predicates must be defined. Let B(O) denote the set of behavior predicates for order O and C denote the set of clusters generated by the algorithm.

$P=\{p_1, p_2, ..., p_m\}$

$\forall p_i \in P \exists \forall B(O_i)$ where $B(O_i)=\{b(o_{ij})|j \in N\}$

Thus: $B(O)=\coprod_{i=1}^{m} B(O_i)$

$C=\{c_1, c_2, ..., c_m\}$

$\forall c_i \in C \exists RC_i$ where $RC_i=\{rc_{ij}|j \in N\}$

Thus: $RC=\coprod_{i=1}^{m} R(C_i)$

Let also denote D as a set of duplicated orders where:

$D=\{d_1, d_2, …., d_m\}$

The following steps are the main parts of the proposed algorithm:

(1) Calculate the behavior predicates for each order in the dataset using the set of predicates P. Let $B(O)$ denote the set of behavior predicates for order O.

(2) Create a feature table F where each row corresponds to an order in the dataset and each column corresponds to a behavior predicate.

(3) Apply a clustering algorithm, such as K-Means or DBSCAN, to the feature table F to group orders with similar behavior. Let C denote the set of clusters generated by the

algorithm.

(4) For each cluster *c* in *C*, choose a representative order *rc* that best captures the behavior of the group. Let *R* denote the set of representative orders.

(5) Use association rule mining to identify patterns in the set of representative orders *R*. This can help identify common sets of behavior predicates that are likely to occur together and can be used to further refine the definition of behavior similarity.

(6) Iterate through the dataset, comparing each order O to the representative orders *rc* in R based on their behavior features. Let *D* denote the set of duplicate orders.

(7) For each order *O* in the dataset, if there exists a representative order *rc* in R such that *B*(*O*) is similar to *B*(*rc*) according to the identified association rules, where similarity is defined using a threshold value or distance metric, then add *O* to the set *D* of duplicate orders.

(8) Remove all duplicate orders in the set *D* from the dataset.

By using clustering algorithms to group orders with similar behavior and association rule mining to identify patterns in the representative orders, we refine our definition of behavior similarity and better capture complex patterns of duplication. This also helps reduce false positives and false negatives and improve the overall accuracy of the deduplication algorithm. For having a de-duplicated dataset as an output, we need the following inputs:

- Dataset of orders *D*
- Set of behavior predicates *P*
- Clustering algorithm like k_means
- Association rule mining technique
- Similarity threshold (0.8 could be an example)
- F as an empty data frame with rows for each order in D and columns for each behavior predicate in P

Then, we move to the vizualisation part [47], according to the proposed framework, to communicate your findings and identify redundant processes that can be eliminated or optimized.

Algorithm: Detecting duplication of processes

```
\begin{algorithm}
  \caption{Deduplication Algorithm}
  \label{deduplication-algorithm}
  \begin{latin}
  \begin{algorithmic}[1]
      \REQUIRE Dataset of orders $D$ with behavior predicates $P$
      \STATE \textbf{Input:} Dataset of orders $D$, Behavior predicates set $P$
      \STATE \textbf{Output:} Deduplicated dataset $D'$
      \STATE \textbf{Initialization} $D' = \emptyset$
      \STATE \textbf{for each} order $O$ in $D$ \textbf{do}
      \STATE \quad Calculate behavior predicates $B(O)$ for order $O$ using set $P$
      \STATE \textbf{end for} \STATE Create a binary feature table $F$ where rows correspond to
          orders in $D$, and columns to behavior predicates in $P$
      \STATE \textbf{for each} order $O$ in $D$ \textbf{do}
      \STATE \quad \textbf{for each} predicate $p$ in $P$ \textbf{do}
      \STATE \quad \quad If $p$ is in $B(O)$ then $F(O, p) = 1$, else $F(O, p) = 0$
      \STATE \quad \textbf{end for}
      \STATE \textbf{end for}
      \STATE Apply clustering algorithm $A$ to feature table $F$ to group orders with similar behavior
          into clusters $C$
      \STATE Initialize set $R$ as an empty set
      \STATE \textbf{for each} cluster $c$ in $C$ \textbf{do}
      \STATE \quad Choose a representative order $rc$ in cluster $c$ with the highest silhouette score
      \STATE \quad Add $rc$ to $R$ \STATE \textbf{end for}
      \STATE Mine association patterns from the set of representative orders $R$ to obtain $patterns$
      \STATE Initialize an empty set $D'$
      \STATE \textbf{for each} order $O$ in $D$ \textbf{do}
      \STATE \quad Initialize $duplicate$ as False
      \STATE \quad \textbf{for each} representative order $rc$ in $R$ \textbf{do}
      \STATE \quad \quad Calculate similarity metric between $B(O)$ and $B(rc)$ based on association
          rules in $patterns$
      \STATE \quad \quad If similarity is greater than or equal to $0.8$ then set $duplicate$ to True and
          break
      \STATE \quad \textbf{end for} \STATE \quad If $duplicate$ is False, add $O$ to $D'$
      \STATE \textbf{end for}
      \RETURN Deduplicated dataset $D'$
    \end{algorithmic}
  \end{latin}
\end{algorithm}
```

4.3 Analysis of redundant processes

Once the duplicated processes are identified, we have used process mining to analyze them. This involve identifying the root cause of the duplication, such as inefficient process design or lack of communication between teams. You also use process mining to identify opportunities for process optimization, such as streamlining the process or eliminating unnecessary steps. process mining is also used to visualize and communicate your findings. Like creating charts, graphs, or dashboards that show the impact of duplicated processes on performance metrics such as cycle time or throughput.

The outputs of the algorithm provide detailed information about the redundancies detected in the processes. This information can include the specific activities or steps that are duplicated, the frequency or occurrence of redundancies, and the impact on process performance metrics such as time, cost, or resource utilization.

By analyzing these outputs, you can gain insights into the root causes of redundancies, identify common patterns or trends, and assess the overall impact on process efficiency. This analysis helps you understand the current state of the processes and serves as a basis for making informed decisions during the redesign phase.

For example, the algorithm's outputs might reveal that certain activities are repeated unnecessarily across different processes, indicating an opportunity to consolidate or standardize those activities. The analysis might also highlight areas where redundant data inputs are causing delays or errors, pointing to the need for better data integration or validation processes.

4.4 Designing new processes

After eliminating the useless processes, the next step is to design new processes that are efficient and effective. Starting by identifying the scope of the process and mapping its flow, then we identify process improvement that eliminates redundancies. This involves simplifying the process, eliminating unnecessary tasks, or automating certain tasks. Finally, we monitor the new process to ensure that it is working as intended. This may involve using process mining to track performance metrics such as cycle time and throughput, and identify any areas for further optimization.

The outputs of the algorithm guide the redesign phase by pinpointing the specific areas where changes are needed to eliminate redundancies. They provide a roadmap for optimizing the processes and improving their efficiency.

Based on the algorithm's outputs, you can prioritize the identified redundancies and determine the most effective strategies for redesign. This might involve resequencing activities, removing unnecessary steps, merging similar processes, or introducing parallel processing to eliminate bottlenecks. The goal is to create streamlined, leaner processes that achieve the desired outcomes without unnecessary duplication.

For instance, if the algorithm identifies redundant approval steps in multiple processes, you can redesign the workflow to consolidate the approval process into a centralized system. This eliminates the need for redundant approvals and streamlines the overall process flow.

The following Figure 5 and Figure 6 show the main step of process mining that includes the analysis part managed by this step of the proposed framework.

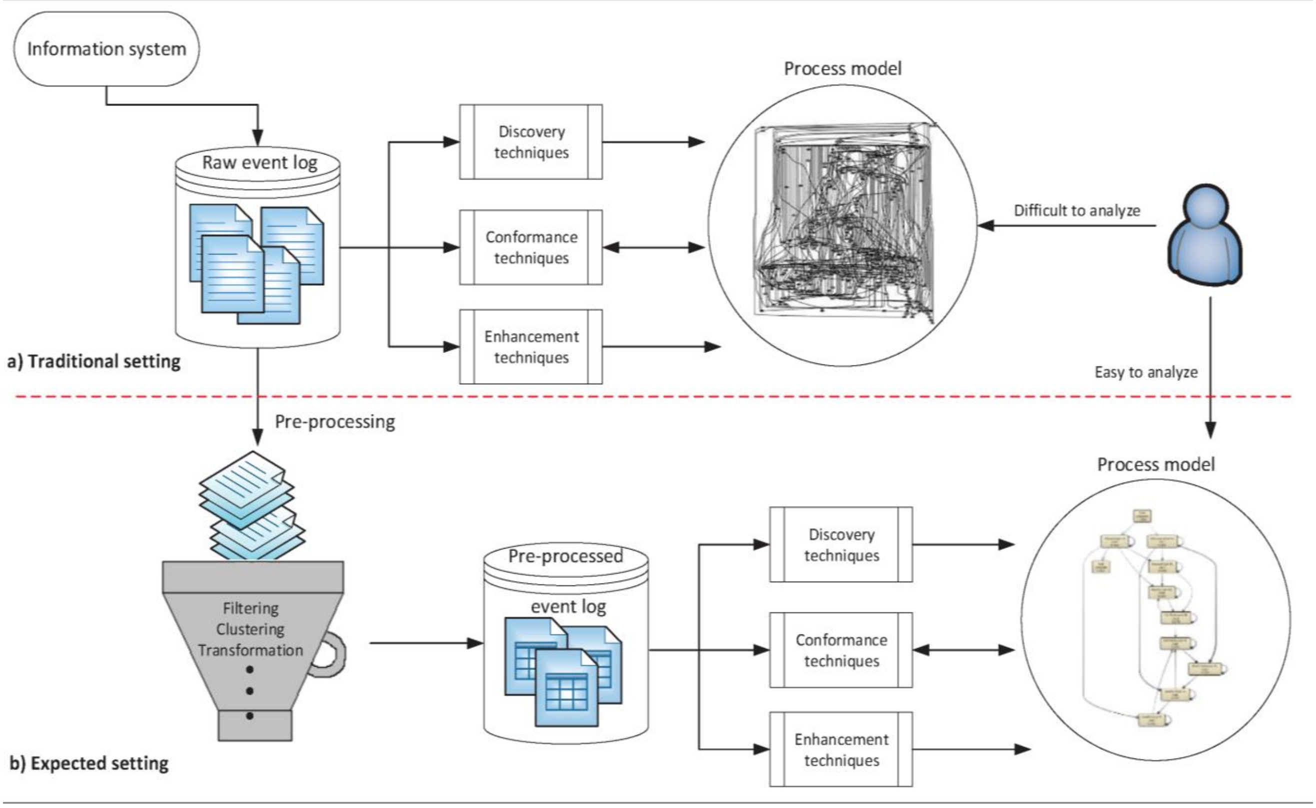

Figure 5. Analysis process

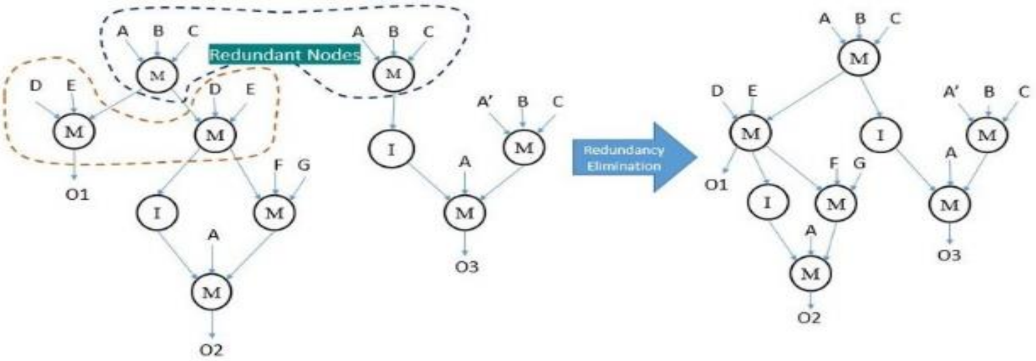

Figure 6. Elimination of duplication

4.5 Elimination of duplicated processes

To proceed in the elimination of the redundancy, we start by identifying the useless and redundant processes. By reviewing the workflow diagrams, we can identify tasks that are redundant or unnecessary. This involves looking for tasks that do not add value to the process, or that can be combined with other tasks to streamline the process. the RPA [48] is also used to identify tasks that are frequently repeated or that take up a lot of time. And finally, reviewing policies identifies areas where policies can be streamlined or consolidated to eliminate duplication.

The outputs of the algorithm also play a role in the automation of processes. By identifying redundancies, you can identify opportunities for automation and leverage technology to streamline and optimize the execution of tasks.

The algorithm's outputs can help identify repetitive or manual tasks that are prone to errors or consume significant time and resources. These tasks can be automated using technologies such as robotic process automation (RPA), where software robots can perform routine tasks with speed and accuracy.

For example, if the algorithm identifies redundant data entry tasks across multiple processes, you can automate the data entry process by developing scripts or workflows that automatically extract and populate data from various sources, reducing the need for manual intervention.

4.6 Adjusted outputs

After eliminating redundancies in our processes, you have adjusted outputs that represent the new, optimized processes. These adjusted outputs are more efficient and effective than the previous outputs, as they have been designed to eliminate redundancies and streamline the process flow.

According to the framework, the behavior similarity is utilized as a measure to determine the level of similarity between different processes. It helps in comparing the behavior of processes based on predefined behavior predicates and identifying patterns or similarities that indicate potential redundancies.

The behavior similarity is applied in the process mining step, where the framework analyzes process data and identifies similar behavior patterns among different processes. By quantifying the level of similarity between processes, the framework can prioritize and target the redundant processes for further analysis and optimization.

Additionally, the behavior similarity also plays a crucial role in the redesign and automation of processes. By understanding the similarities and redundancies in the behavior of processes, the framework can propose optimized process designs and automation approaches to eliminate duplicate or unnecessary steps, streamline operations, and improve overall efficiency.

Therefore, the behavior similarity is a central component in the analysis layer of the framework, where it drives the identification of redundancies and informs the redesign and automation of processes.

5. CASE STUDY

Suppose we have a dataset of 10 customer orders, and we want to identify and remove any duplicate orders. Each order is represented by a set of behavior predicates, which are defined as follows:

- Order ID: a unique identifier for each order
- Customer ID: the ID of the customer who placed the order
- Order date: the date the order was placed
- Order total: the total amount of the order
- Product ID: the ID of the product(s) included in the order
- Quantity: the quantity of each product in the order

To begin, we calculate the behavior predicates for each order using the set of predicates P. Here's an example of the behavior predicates for one order:

- Order ID: 1
- Customer ID: 101
- Order date: 2022-01-01
- Order total: 50.00
- Product ID: [101, 102]
- Quantity: [2, 1]

We can represent this order as a set of behavior predicates: B(1) = {101, 2022-01-01, 50.00, [101, 102], [2, 1]}.

We do this for all orders in the dataset, and create a feature table F where each row corresponds to an order and each column corresponds to a behavior predicate. The following table represents F.

We then apply the clustering algorithm, to group orders with similar behavior.

After clustering, we have the following set of representative orders R:

R={r1, r2, r3}

where, r1 represents the group of orders [o1, o4], r2 represents the group [o2, o5], and r3 represents the group [o3].

Representative order selection:

After applying the K-Means clustering algorithm to the dataset, we obtain clusters based on the similarity of behavior features.

For each cluster, we calculate the centroid, which represents the average behavior of the orders within that cluster.

To select a representative order from each cluster, we choose the order that is closest to the centroid in terms of behavior feature values.

The representative order captures the central behavior pattern of the orders in the cluster.

We then apply association rule mining to identify patterns in the set of representative orders. Suppose we find the following association rule:

{behavior_predicate_1, behavior_predicate_3} => behavior_predicate_2

Let's define the behavior predicates based on the provided association rule and clarify their meanings:

Behavior Predicate 1: Product Type (e.g., electronic, clothing, accessories)

Behavior Predicate 2: Quantity (e.g., number of items in the order)

Behavior Predicate 3: Order Total (e.g., total cost of the order)

With these behavior predicates defined, the association rule can be expressed as:

{Product Type, Order Total} => Quantity

This association rule indicates that there is a relationship between the product type, order total, and the quantity of items ordered. For example, it suggests that certain combinations of product type and order total are likely to be associated with specific quantities.

This rule suggests that if orders have behavior predicates 1 and 3 in common, they are likely to also have behavior predicate 2 in common. We can use this rule to further refine our definition of behavior similarity.

Next, we iterate through the dataset, comparing each order to the representative orders in R based on their behavior features. For example, we compare order o1 to representative order r1:

B(o1) = {behavior_predicate_1, behavior_predicate_2,
behavior_predicate_3}
B(r1) = {behavior_predicate_1, behavior_predicate_2,
behavior_predicate_3}

Since the behaviour features of o1 and r1 are identical, we add o1 to the set of duplicate orders D.

Similarly, we compare o2 to r2.

B(o2) = {behavior_predicate_2, behavior_predicate_4}
B(r2) = {behavior_predicate_2, behavior_predicate_5}

In this case, the behavior features of o2 and r2 do not match exactly, but we can use the association rule we discovered earlier to determine if they are similar enough to be considered duplicates.

Since o2 has behavior predicate 2 in common with r2, and both also have behavior predicate 5, we can infer that they are likely to have behavior predicate 4 in common as well (according to the association rule). Therefore, we add o2 to the set D.

We repeat this process for all orders in the dataset, and end up with the following set of duplicate orders D:

D = {**o1, o2, o4**}

We then remove these orders from the dataset to obtain the set of unique orders:

unique_orders = {**o3, o5**}

Table 1 presents a dataset of 10 customers orders as an example of our case study.

The integration of clustering algorithms and association rule mining can enhance the deduplication process by capturing behavior similarity and identifying patterns within representative orders.

Clustering helps group similar orders together, improving the identification of duplicates within clusters.

Association rule mining provides insights into the relationships between behavior predicates, allowing for refined behavior similarity detection. Furthermore, Determining the optimal number of clusters in the clustering algorithm can be challenging. It requires domain knowledge and experimentation to find the appropriate balance between granularity and accuracy.

The quality and availability of data can greatly impact the effectiveness of the deduplication algorithm. Incomplete or inconsistent data can lead to inaccurate behavior similarity detection and representative order selection.

Table 1. The dataset of customer's orders

Order ID	Customer ID	Order Date	Order Total	Product ID	Quantity
1	101	2022-01-01	50.00	[101, 102]	[2, 1]
2	102	2022-01-02	35.00	[101, 103]	[1, 1]
3	101	2022-01-03	75.00	[101, 102, 103]	[2, 1, 1]
4	103	2022-01-04	40.00	[101, 102]	[1, 2]
5	104	2022-01-05	50.00	[102, 103]	[1, 1]
6	101	2022-01-06	65.00	[101, 103]	[2, 2]
7	103	2022-01-07	45.00	[101, 103]	[2, 1]
8	102	2022-01-08	30.00	[101]	[1]
9	104	2022-01-09	55.00	[102]	[2]
10	101	2022-01-10	50.00	[101, 102, 103]	[2, 2, 1]

6. CONCLUSIONS

The objective of eliminating redundant processes in business processes is to simplify them and make them more efficient. Redundant processes are steps or activities that are repeated unnecessarily or are not necessary for completing a given task. These processes can cause delays, errors, and additional costs.

By eliminating redundant processes, businesses can improve their productivity, quality, and responsiveness to their customers' demands. They can also reduce costs associated with process management by streamlining tasks and avoiding duplicate tasks. Furthermore, by simplifying processes, employees can focus on more important tasks and better understand their role in achieving the company's goals.

Ultimately, the elimination of redundant processes can help businesses be more competitive and improve customer satisfaction by offering faster and more reliable service.

The proposed approach focuses on the detection and elimination of redundancies in business processes. The techniques employed include ETL (Extract, Transform, Load) for data collection and preparation, process mining for analyzing process redundancies, and data mining for detecting similarities and patterns in process behavior.

In our paper, we have chosen to focus on usefulness of data mining and process mining in detection and analyzing duplications of processes. Data mining helps identify patterns and correlations in large datasets, which can be used to identify potential instances of duplicated processes. By analyzing historical data, data mining algorithms can uncover hidden relationships between different process steps, which can be used to identify where redundancies may exist. On the other hand, process mining algorithms can provide insights into how processes are actually executed, and can identify areas where redundancies may exist. This can be particularly useful in identifying bottlenecks or inefficiencies in a process, which may not be immediately apparent from a high-level view.

The framework begins with data collection and preparation, where data is extracted from various sources and transformed into a suitable format. It then moves to the analysis phase, utilizing process mining techniques to identify redundancies and inefficiencies in the processes. This analysis involves visualizing process flows, identifying bottlenecks, and highlighting areas where redundancies exist.

The next step involves applying data mining techniques to detect similarities and patterns in process behavior. This helps in identifying redundant process steps or subprocesses that can be eliminated or optimized. The framework further facilitates the elimination of redundancies through process adjustments, which may involve streamlining process steps, optimizing process flows, or automating certain tasks.

The outcomes of implementing this framework are improved process efficiency, reduced redundancies, and streamlined operations. By eliminating redundancies, organizations can achieve cost savings, improve productivity, and enhance overall process performance. The framework provides valuable insights into process analysis, redesign, and automation, leading to more efficient and effective business processes.

The proposed framework is designed to help organizations improve their business processes by identifying and eliminating redundancies. The behavior similarity measure in the algorithm compares the behavior of different processes to identify similarities and redundancies. The algorithm takes as input a set of processes and then calculates the behavior similarity between each pair of processes and identifies redundancies based on a user-defined threshold. An algorithm is also proposed as a part of data mining tool to strongly detect redundancy.

The case study conducted using this framework provides concrete results and examples of how the approach has been successfully applied. It demonstrates the practicality and effectiveness of the techniques employed, highlighting the value of adopting this framework in real-world business scenarios.

The benefits of the proposed framework and algorithm are numerous. By identifying and eliminating redundancies, organizations can reduce costs, increase efficiency, and improve customer satisfaction. The framework provides a structured approach to analyzing and improving business processes, while the algorithm provides a powerful tool for detecting redundancies. Together, they can help organizations achieve their process improvement goals and stay ahead of the competition.

The proposed approach has several strengths that make it valuable for addressing redundancies in business processes. Firstly, it combines multiple techniques such as ETL, process mining, and data mining, allowing for a comprehensive analysis of the processes. This multi-faceted approach provides a holistic view of the redundancies and enables effective decision-making in process optimization.

Another strength is the emphasis on behavior similarity analysis, which enables the identification of similar process patterns and the detection of redundant steps or subprocesses. By leveraging behavior similarity, the approach can pinpoint areas where redundancies exist and suggest targeted adjustments to eliminate them.

Additionally, the framework promotes data-driven decision-making by utilizing data collection, preparation, and analysis. By leveraging the power of data, organizations can make informed decisions and validate the effectiveness of process adjustments.

Despite its strengths, the proposed approach also has some limitations. Firstly, it requires access to comprehensive and accurate data for meaningful analysis. In situations where data quality is poor or limited, the effectiveness of the approach may be compromised.

7. LIMITATIONS

The most remarkable limitations of our proposed algorithm are the followings:

The deduplication algorithm's performance heavily relies on the quality and relevance of the behavior predicates used. Selecting the most informative and relevant behavior predicates is crucial for accurate deduplication.

The algorithm assumes that behavior similarity is solely determined by the selected behavior predicates. However, other factors, such as contextual information or temporal patterns, may also influence behavior similarity.

But, in this part we can propose some potential improvements that might be benefits for these limitations. For example:

Exploring different clustering algorithms, such as hierarchical clustering or density-based clustering, could provide alternative approaches for grouping orders with similar behavior.

Incorporating additional data preprocessing techniques, such as data cleaning or feature engineering, can improve the quality and consistency of the behavior predicates, leading to more accurate deduplication results.

Considering more advanced machine learning algorithms, such as deep learning or ensemble methods, could potentially enhance the deduplication process by capturing complex patterns and interactions among behavior predicates.

8. FUTURE WORKS

Exploring potential areas for future research is essential for the continuous development and improvement of the proposed framework. There are several directions that can be considered:

Scaling up to larger datasets: Conducting experiments and testing the proposed framework on larger and more complex datasets would provide valuable insights into its scalability and performance. This could involve exploring techniques for handling big data, optimizing algorithms for efficiency, and evaluating the framework's effectiveness on a larger scale.

Investigating alternative algorithms and techniques: While the current framework incorporates well-established algorithms and techniques, there is room for exploring alternative approaches. Researchers can investigate different data mining algorithms, behavior similarity measures, or process mining techniques to compare their effectiveness in identifying and eliminating redundancies.

Adapting the framework to different process domains: The framework's applicability across various process domains can be explored. Conducting case studies in different industries or domains, such as healthcare, manufacturing, or finance, would provide insights into the framework's generalizability and its ability to address redundancies specific to those domains.

Enhancing automation capabilities: Automation plays a crucial role in process optimization. Future research can focus on developing intelligent automation techniques that can automatically identify and eliminate redundancies in processes, reducing the manual effort required in the analysis and redesign phases of the framework.

Incorporating real-time analysis: As organizations strive for real-time decision-making, integrating real-time data collection and analysis capabilities into the framework would be beneficial. This would enable the identification and elimination of redundancies in near real-time, allowing for proactive process optimization.

By addressing these areas for future research, researchers can further advance the proposed framework, expand its applicability, and enhance its capabilities for addressing redundancies in business processes.

REFERENCES

[1] Dijkman, R., Dumas, M., Van Dongen, B., Käärik, R., Mendling, J. (2011). Similarity of business process models: Metrics and evaluation. Information Systems, 36(2): 498-516. https://doi.org/10.1016/j.is.2010.09.006

[2] Amiri, M.J., Koupaee, M. (2017). Data-driven business process similarity. IET Software, 11(6): 309-318. https://doi.org/10.1049/iet-sen.2016.0256

[3] Aygün, S., Okyay, S. (2015). Improving the pearson similarity equation for recommender systems by age parameter. In 2015 IEEE 3rd Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), pp. 1-6. https://doi.org/10.1109/AIEEE.2015.7367282

[4] Wu, X., Huang, Y., Wang, S. (2017). A new similarity computation method in collaborative filtering based recommendation system. In 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), pp. 1-5. https://doi.org/10.1109/VTCFall.2017.8288359

[5] Ayub, M., Ghazanfar, M.A., Maqsood, M., Saleem, A. (2018). A Jaccard base similarity measure to improve performance of CF based recommender systems. In 2018 International Conference on Information Networking (ICOIN), pp. 1-6. https://doi.org/10.1109/ICOIN.2018.8343073

[6] Su, Z., Zheng, X., Ai, J., Shen, Y., Zhang, X. (2020). Link prediction in recommender systems based on vector similarity. Physica A: Statistical Mechanics and Its Applications, 560: 125154. https://doi.org/10.1016/j.physa.2020.125154

[7] Kunze, M., Weidlich, M., Weske, M. (2011). Behavioral similarity–A proper metric. In Business Process Management: 9th International Conference, BPM 2011, Clermont-Ferrand, France, August 30-September 2, 2011. Proceedings 9, pp. 166-181. https://doi.org/10.1007/978-3-642-23059-2_15

[8] Verbeek, H.M.W., Buijs, J.C.A.M., Van Dongen, B.F., van der Aalst, W.M. (2010). Prom 6: The process mining toolkit. Proc. of BPM Demonstration Track, 615: 34-39.

[9] Taymouri, F., La Rosa, M., Dumas, M., Maggi, F.M. (2021). Business process variant analysis: Survey and classification. Knowledge-Based Systems, 211: 106557. https://doi.org/10.1016/j.knosys.2020.106557

[10] Entringer, T.C., de Oliveira Nascimento, D.C., da Silva Ferreira, A., Siqueira, P.M.T., de Souza Boechat, A., Cerchiaro, I.B., Ramos, R.R. (2019). Comparative analysis main methods business process modeling: Literature review, applications and examples. International Journal of Advanced Engineering Research and Science, 6(5).

[11] Li, Z., Wu, J., Zhang, X., He, J., Chen, P., He, K. (2020). Using metadata for recommending business process. The Journal of Supercomputing, 76: 3729-3748. https://doi.org/10.1007/s11227-018-2601-5

[12] Nuritha, I., Mahendrawathi, E.R. (2019). Behavioural similarity measurement of business process model to compare process discovery algorithms performance in dealing with noisy event log. Procedia Computer Science, 161: 984-993. https://doi.org/10.1016/j.procs.2019.11.208

[13] Van Dongen, B., Dijkman, R., Mendling, J. (2013). Measuring similarity between business process models. Seminal Contributions to Information Systems Engineering: 25 Years of CAiSE, 405-419. https://doi.org/10.1007/978-3-642-36926-1_33

[14] Gazdar, A., Hidri, L. (2020). A new similarity measure for collaborative filtering based recommender systems. Knowledge-Based Systems, 188: 105058. https://doi.org/10.1016/j.knosys.2019.105058

[15] Wegener, D., Rüping, S. (2010). On integrating data mining into business processes. In Business Information Systems: 13th International Conference, BIS 2010, Berlin, Germany, May 3-5, 2010. Proceedings 13, pp. 183-194. https://doi.org/10.1007/978-3-642-12814-1_16

[16] Der Van Aalst, P.W.M. (2012). Process mining: Overview and opportunities. ACM Transactions on Management Information Systems, 3(2): 71-717. https://doi.org/10.1145/2229156.2229157

[17] Poncin, W., Serebrenik, A., Van Den Brand, M. (2011). Process mining software repositories. In 2011 15th European Conference on Software Maintenance and Reengineering, pp. 5-14. https://doi.org/10.1109/CSMR.2011.5

[18] Reinkemeyer, L. (2020). Process mining in action. Process Mining in Action Principles, Use Cases and Outloook. https://doi.org/10.1007/978-3-030-40172-6

[19] Bhaskar, L.H. (2018). Business process reengineering: A process based management tool. Serbian Journal of Management, 13(1): 63-87.

[20] Anand, A., Fosso Wamba, S., Gnanzou, D. (2013). A literature review on business process management, business process reengineering, and business process innovation. In Enterprise and Organizational Modeling and Simulation: 9th International Workshop, EOMAS 2013, Held at CAiSE 2013, Valencia, Spain, June 17, 2013, pp. 1-23. https://doi.org/10.1007/978-3-642-41638-5_1

[21] Grover, V., Malhotra, M.K. (1997). Business process reengineering: A tutorial on the concept, evolution, method, technology and application. Journal of Operations Management, 15(3): 193-213. https://doi.org/10.1016/S0272-6963(96)00104-0

[22] Attaran, M. (2004). Exploring the relationship between information technology and business process reengineering. Information & Management, 41(5): 585-596. https://doi.org/10.1016/S0378-7206(03)00098-3

[23] Terragni, A., Hassani, M. (2018). Analyzing customer journey with process mining: From discovery to recommendations. In 2018 IEEE 6th International

Conference on Future Internet of Things and Cloud (FiCloud), pp. 224-229. https://doi.org/10.1109/FiCloud.2018.00040

[24] Larson, D., Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. International Journal of Information Management, 36(5): 700-710. https://doi.org/10.1016/j.ijinfomgt.2016.04.013

[25] Lim, E.P., Chen, H., Chen, G. (2013). Business intelligence and analytics: Research directions. ACM Transactions on Management Information Systems (TMIS), 3(4): 1-10. http://dx.doi.org/10.1145/2407740.2407741

[26] Sharda, R., Delen, D., Turban, E., Aronson, J., Liang, T. (2014). Business intelligence and analytics. System for Decesion Support, 398: 2014.

[27] Gao, X., Xiao, B., Tao, D., Li, X. (2010). A survey of graph edit distance. Pattern Analysis and Applications, 13: 113-129. http://dx.doi.org/10.1007/s10044-008-0141-y

[28] Myers, R., Wison, R.C., Hancock, E.R. (2000). Bayesian graph edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(6): 628-635. https://doi.org/10.1109/34.862201

[29] Hawashin, B., Lafi, M., Kanan, T., Mansour, A. (2020). An efficient hybrid similarity measure based on user interests for recommender systems. Expert Systems, 37(5): e12471. https://doi.org/10.1111/exsy.12471

[30] Dumas, M., Garcıa-Banuelos, L., Dijkman, R. (2009). Similarity search of business process models. Data Engineering, 25.

[31] Nuritha, I., Mahendrawathi, E.R. (2017). Structural similarity measurement of business process model to compare heuristic and inductive miner algorithms performance in dealing with noise. Procedia Computer Science, 124: 255-263. https://doi.org/10.1016/j.procs.2017.12.154

[32] Haydar, C.A. (2014). Les systèmes de recommandation à base de confiance. Université de Lorraine.

[33] Bae, J., Caverlee, J., Liu, L., Yan, H. (2006). Process mining by measuring process block similarity. In Business Process Management Workshops: BPM 2006 International Workshops, BPD, BPI, ENEI, GPWW, DPM, semantics4ws, Vienna, Austria, September 4-7, 2006. Proceedings 4, pp. 141-152. https://doi.org/10.1007/11837862_15

[34] Bezerra, F., Wainer, J., van der Aalst, W.M. (2009). Anomaly detection using process mining. In International Workshop on Business Process Modeling, Development and Support, pp. 149-161. https://doi.org/10.1007/978-3-642-01862-6_13

[35] Costa, M.B., Tamzalit, D. (2017). Recommendation patterns for business process imperative modeling. In Proceedings of the Symposium on Applied Computing, pp. 735-742. https://doi.org/10.1145/3019612.3019619

[36] Vyas, S., Vaishnav, P. (2017). A comparative study of various ETL process and their testing techniques in data warehouse. Journal of Statistics and Management Systems, 20(4): 753-763. https://doi.org/10.1080/09720510.2017.1395194

[37] Muñoz, L., Mazón, J.N., Pardillo, J., Trujillo, J. (2008). Modelling ETL processes of data warehouses with UML activity diagrams. In OTM Confederated International Conferences on the Move to Meaningful Internet Systems, pp. 44-53. https://doi.org/10.1007/978-3-540-88875-8_21

[38] Chaudhuri, S., Dayal, U. (1997). An overview of data warehousing and OLAP technology. ACM Sigmod Record, 26(1): 65-74. https://doi.org/10.1145/248603.248616

[39] Garani, G., Chernov, A., Savvas, I., Butakova, M. (2019). A data warehouse approach for business intelligence. In 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 70-75. https://doi.org/10.1109/WETICE.2019.00022

[40] Kumbhare, T.A., Chobe, S.V. (2014). An overview of association rule mining algorithms. International Journal of Computer Science and Information Technologies, 5(1): 927-930.

[41] Dehdouh, K., Boussaid, O., Bentayeb, F. (2020). Big data warehouse: Building columnar NOSQL OLAP cubes. International Journal of Decision Support System Technology (IJDSST), 12(1): 1-24. https://doi.org/10.4018/IJDSST.2020010101

[42] Etcheverry, L., Vaisman, A., Zimányi, E. (2014). Modeling and querying data warehouses on the semantic web using QB4OLAP. In International Conference on Data Warehousing and Knowledge Discovery, pp. 45-56. https://doi.org/10.1007/978-3-319-10160-6_5

[43] Aher, S.B., Lobo, L. (2012). Applicability of data mining algorithms for recommendation system in e-learning. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics, pp. 1034-1040. https://doi.org/10.1145/2345396.2345562

[44] Venkatkumar, I.A., Shardaben, S.J.K. (2016). Comparative study of data mining clustering algorithms. In 2016 International Conference on Data Science and Engineering (ICDSE), pp. 1-7. https://doi.org/10.1109/ICDSE.2016.7823946

[45] Alasadi, S.A., Bhaya, W.S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16): 4102-4107.

[46] Nikam, S.S. (2015). A comparative study of classification techniques in data mining algorithms. Oriental Journal of Computer Science and Technology, 8(1): 13-19.

[47] Ragavi, R., Srinithi, B., Sofia, V.A. (2018). Data mining issues and challenges: A review. International Journal of Advanced Research in Computer and Communication Engineering, 7(11): 118-121.

[48] Risques et Sécurité associés à la Robotic Process Automation. (2019). Réflexions des établissements financiers du Forum des Compétences, Janvier 2019. https://www.forum-des-competences.org/assets/files/risques-et-securite-lies-a-la-rpa-livrable.pdf