Analysing methods of Workflow Execution on Data Mining

Lalita Kumari

Assistant Professor, Department of Computer Science, Amity University,
Patna, Bihar, India.
kumaril2003@yahoo.co.in

Niranjan Kumar
Assistant Professor, Department of Physics, Amity University,
Patna, Bihar, India.
Niranjan_sinha786@redifmail.com

Seemi Kumari

B.Tech Student, Department of Computer Science, Amity University,
Patna, Bihar, India.
seemisingh0@gmail.com

Abstract

In this paper we are analyzing the different methods of workflow execution. The paper provides an overview of the field of workflow mining, which involves the extraction of information from event logs to understand and improve business processes. The paper discusses the importance of workflow mining in the context of process improvement and optimization and discusses the various techniques and approaches that have been developed for this purpose. This paper discusses the various methods and techniques for optimizing workflow execution in data mining. The related work suggests that the choice of workflow execution method and platform should be carefully considered based on the specific requirements of the data mining project, including the size of the data set, the complexity of the workflow, and the available resources. Workflow management systems (WFMS) provide a framework for designing, executing, and monitoring data mining workflows, and can help to improve workflow efficiency and accuracy.

***Keywords*: Data mining; Mining technology; Workflow data.**

1. The Main Text

During the time of this rapid development if we bring the advantages of web database web technology and database technology then it will be profitable. The development of web pages from static to dynamic-database driven pages and its development to achieve the separation and application environments. To achieve the development in e-commerce web technology, distributed object technology and security technology have been advanced, major challenge was to achieve real-time transaction processing, scalability, security and client authentication and many more. By avoiding the bottleneck that is caused due to distribution and object technology it provides direct contact with server. It dynamically balances client requests and can be run from a single scalable server functionality. This technology effectively solves the real-time transaction processing, scalable web. Data access is the core problem in the entire data warehouse problem. Data warehouse from multiple information source to obtain the original data, after finishing store in internal database by providing access to tools to user this information environment support business global decision-making process in depth and comprehensive analysis of business management. Source data for data warehousing such as various production system, online transaction processing system operational data, external data source and so can be used as a data warehouse data source. Data extraction, transforming and loading tools, its job is to extract the data from the data source, inspecting data and it's sorting according to design requirements, the data is then reorganizing and processing and loaded into the data warehouse target database.

2. Data Modeling Tools

2.1 Core Storage

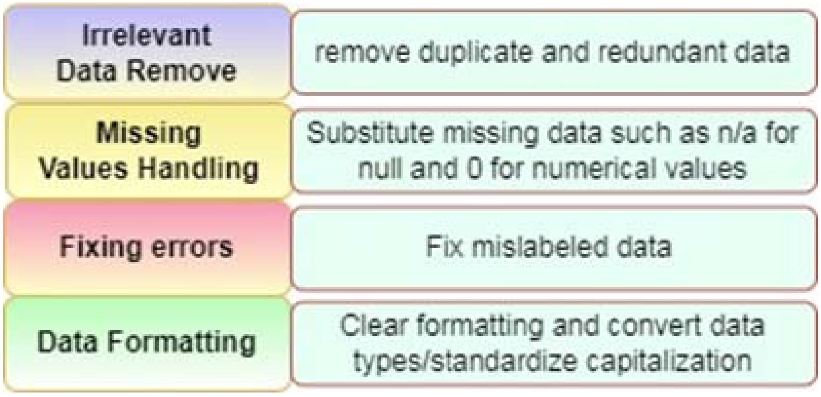
stores data models and metadata. Wherein the metadata describes the data warehouse information source and the target data itself, that is conversion from source to target data. Data warehouse target database, examination storage, sorting, preprocessing, and re-organizing the data after any among relational database management or special multidimensional management system can be used. Front - end data access and analysis tools, for business analysis and decision makers to access data in target database, and in- depth analysis. Such tolls that can be currently obtained are "relational query tool, standard client/ server tool, and decision support system(dss)/chief information system(cis)" software package.

2.1.1. Data warehouse management tools

It includes content security management, storage management, and other aspects.

2.1.2. Data cleaning

It is one of the most important and challenging tasks, which is a process of identifying and removing the errors to ensure quality and consistency of data set. For data cleaning various methodologies are available but the methodology suitable for one type of data may not be suitable for other kinds of data. Cleaning strategy depends on category of data and domain specific.



Irrelevant Data Remove	remove duplicate and redundant data
Missing Values Handling	Substitute missing data such as n/a for null and 0 for numerical values
Fixing errors	Fix mislabeled data
Data Formatting	Clear formatting and convert data types/standardize capitalization

Fig.1. Steps for Data Cleaning

3. Related Work

There have been numerous studies on workflow execution methods for data mining, with researchers exploring various approaches to optimize performance and accuracy. One study by [7] compared different workflow execution approaches, including sequential, parallel, and distributed execution, using the Mahout machine learning library. The study found that distributed execution was the most effective approach for large-scale data mining tasks, providing significant performance improvements compared to the other methods. This paper [11] evaluated the performance of different workflow execution platforms, including Apache Hadoop and Apache Spark, for text mining tasks. The study found that Apache Spark was more efficient than Hadoop for iterative data mining tasks, while Hadoop was more suitable for batch processing. Several studies have also explored the use of cloud computing platforms for workflow execution in data mining. For example, a study by [8] compared the performance of different cloud platforms, including Amazon Web Services and Microsoft Azure, for data mining tasks. The study found that the choice of platform could have a significant impact on performance and cost, highlighting the importance of careful platform selection. The related work suggests that the choice of workflow execution method and platform should be carefully considered based on the specific requirements of the data mining project, including the size of the data set, the complexity of the workflow, and the available resources. By selecting the right approach and platform, researchers can optimize their workflow for maximum efficiency and accuracy. There are also several studies that have explored specific techniques for optimizing workflow execution in data mining. For example, author has [10] proposed a task scheduling algorithm for distributed data mining, which aims to reduce communication costs between nodes and improve overall performance. This paper [9] proposed a workflow optimization method for big data processing, which involves

identifying and removing redundant tasks in the workflow to reduce processing time and improve efficiency. There have been several studies that have explored the use of machine learning techniques to optimize workflow execution. For example, author [11] proposed a deep reinforcement learning approach to optimize task scheduling in distributed data mining, which aims to learn an optimal policy for scheduling tasks based on the current system state. There have been several studies that have explored the use of workflow management systems (WFMS) for data mining. These systems provide a framework for designing, executing, and monitoring data mining workflows, and can help to improve workflow efficiency and accuracy. Some popular WFMS for data mining include KNIME, RapidMiner, and Apache Airflow. The related work suggests that there are many different approaches and techniques for optimizing workflow execution in data mining, and researchers should carefully evaluate the strengths and limitations of each approach to select the best method for their specific needs. The workflow life cycle consists of four phases: workflow design, workflow configuration, workflow enactment, and workflow diagnosis. The first phase, workflow design, involves the creation of a formal representation of a business process. This includes identifying the tasks involved, their order, the actors responsible for them, and any decision points or branching that may occur. The goal of this phase is to create a model that accurately reflects the business process and can be used as a blueprint for workflow implementation.
The second phase, workflow configuration, involves taking the workflow design and configuring it within a workflow management system. This includes defining the workflow tasks, actors, and their roles, as well as any business rules or conditions that may affect the workflow's execution. The goal of this phase is to create a functional workflow that can be executed within the system. The third phase, workflow enactment, involves the actual execution of the workflow within the workflow management system. This includes assigning tasks to actors, tracking progress, and managing any exceptions or errors that may occur. The goal of this phase is to ensure that the workflow executes as designed and produces the desired output. The fourth phase, workflow diagnosis, involves monitoring and analyzing the workflow's performance to identify any bottlenecks, inefficiencies, or areas for improvement. This includes collecting data on task completion times, resource utilization, and any errors or exceptions that may occur. The goal of this phase is to optimize the workflow's performance and improve overall business efficiency. The workflow life cycle provides a structured approach to designing, configuring, executing, and optimizing business processes, helping organizations to achieve their business objectives more efficiently and effectively. The goal of workflow mining is to extract useful information and knowledge from event logs or other data sources related to business processes, to improve and optimize those processes. Workflow mining aims to uncover the actual flow of activities and interactions within a business process, identify potential bottlenecks or inefficiencies, and suggest improvements or optimizations that can lead to better performance, reduced costs, and increased customer satisfaction. The insights gained from workflow mining can help organizations to make data-driven decisions, streamline their operations, and ultimately achieve their business objectives more efficiently and effectively.

4. Types of Workflow Mining

The execution data of a workflow can be analysed using various data mining techniques, including web mining. Web mining involves extracting useful information and knowledge from web-related data sources, such as web server logs, clickstream data, and user behaviour on web applications. In the context of workflow execution, web mining can be used to analyse the data generated by the workflow management system, such as event logs, task completion times, and resource utilization. This can help identify patterns, trends, and anomalies in the workflow's execution, as well as identify potential bottlenecks or inefficiencies. For example, web mining can be used to analyse the task completion times for each activity in the workflow, identifying tasks that are taking longer than expected and potentially causing delays in the overall process. It can also be used to analyse the resource utilization, identifying tasks that are overloading a particular resource or actor, and potentially leading to poor performance. Web mining can provide valuable insights into the performance of a workflow, helping organizations to optimize their business processes and achieve their objectives more efficiently and effectively. By analysing workflow execution data based on web mining, organizations can make data-driven decisions and improve their operations based on real-world data. Workflow logs are a common data format used to capture and store information about the execution of workflows. They provide a detailed record of each activity and event that occurs during the workflow's execution, including the task completion times, resource utilization, and any errors or exceptions that may occur. To facilitate the exchange and analysis of workflow logs, a common XML format has been developed. This format, known as the Workflow Log Data Specification (WLDS), provides a standardized way to represent workflow logs, making it easier to share and analyze the data across different systems and applications. The WLDS format defines a set of XML tags and attributes that can be used to represent different types of workflow events, such as task start and completion, resource allocation, and exception handling. It also defines a standard structure for organizing the workflow log data, including the workflow instance, the process model, and the event log. Using a common XML format for workflow logs can help organizations to more easily exchange and integrate data across different systems, making it easier to analyse and optimize business processes. It also facilitates the development of tools and techniques for workflow mining, allowing organizations to gain insights into their operations and make data-driven decisions to improve their performance.

```
<!ELEMENT workflow (processModel, eventLog)>
<!ELEMENT processModel (process)>
<!ELEMENT process (data, activities, transitions)>
<!ELEMENT data (attribute*)>
<!ELEMENT attribute (#PCDATA)>
<!ELEMENT activities (activity*)>
<!ELEMENT activity (data, event*, resource)>
<!ELEMENT event (data, type)>
<!ELEMENT resource (data, type)>
<!ELEMENT transitions (transition*)>
<!ELEMENT transition (data, source, target, event*)>
<!ELEMENT eventLog (trace*)>
<!ELEMENT trace (event*)>
<!ELEMENT data (attribute*)>
<!ELEMENT attribute (#PCDATA)>
<!ELEMENT type (#PCDATA)>
<!ELEMENT source (#PCDATA)>
<!ELEMENT target (#PCDATA)>
<!ELEMENT timestamp (#PCDATA)>
<!ELEMENT eventType (#PCDATA)>

<!ATTLIST workflow
id ID #REQUIRED>
<!ATTLIST process
id ID #REQUIRED>
<!ATTLIST data
id ID #REQUIRED>
<!ATTLIST activity
id ID #REQUIRED
name CDATA #IMPLIED>
<!ATTLIST event
id ID #REQUIRED
timestamp CDATA #REQUIRED
eventType CDATA #REQUIRED>
<!ATTLIST resource
id ID #REQUIRED
name CDATA #IMPLIED>
<!ATTLIST transition
id ID #REQUIRED>
<!ATTLIST trace
id ID #REQUIRED>
```

5. Experiments and Result

Data mining tools are software applications that are used to extract useful information and insights from large sets of data. These tools use a variety of statistical and machine learning techniques to identify patterns, trends, and relationships in the data. Here are some common types of data mining tools: (i) Clustering tools: these tools group similar data points together based on their similarities and differences. Clustering is often used for customer segmentation and market research. Association rule mining tools: These tools identify correlations between variables and can help identify rules or patterns that exist between different variables in a dataset. (ii) Classification tools: these tools are used to classify data into predefined categories based on certain features or characteristics. For example, a classification tool could be used to predict which customers are likely to buy a certain product based on their demographic data. (iii) Regression analysis tools: these tools are used to identify the relationship between two or more variables and can be used to make predictions about future trends or behaviors. (iv) Neural network tools: these tools are designed to mimic the functioning of the human brain and can be used for complex pattern recognition tasks, such as image or speech recognition. (v) Decision tree tools: these tools create a visual representation of the decision-making process by breaking down a problem into smaller, simpler parts. They are often used for predictive modeling and risk analysis. (vi) Text mining tools: these tools are used to extract insights from unstructured data, such as text documents or social media posts. They can be used for sentiment analysis, topic modeling, and other text-based analysis. (vii) Visualization tools: These tools are used to create graphical representations of the data, which can make it easier to identify patterns and trends. Data mining tools are essential for businesses and organizations that deal with large sets of data. They can help identify valuable insights and improve decision-making processes, leading to improved performance and profitability. Dealing with noise and incomplete logs in web mining can be challenging, but there are several strategies that can be used to address these issues: *Data preprocessing*: One approach to dealing with noisy and incomplete data is to preprocess the data before mining it. This can involve techniques such as data cleaning, data integration, and data transformation. For example, missing data can be imputed using statistical techniques, or noisy data can be filtered out using outlier detection methods. *Data fusion:* Another approach is to combine data from multiple sources to improve the quality of the data. For example, web log data can be combined with data from other sources, such as user registration data or customer relationship management data, to create a completer and more accurate dataset. *Feature selection:* In some cases, it may be possible to select a subset of features that are less affected by noise and incomplete data. This can help to reduce the impact of noise. on the mining results. *Ensemble learning:* ensemble learning is a machine learning technique that combines multiple models to improve the accuracy and robustness of the mining results. This approach can be particularly useful when dealing with noisy and incomplete data. *Domain knowledge:* Finally, it is important to have domain knowledge of the data being mined. This can help to identify and filter out noisy data or to interpret the results in light of the underlying data quality issues. Dealing with noise and incomplete logs in web mining requires a combination of data preprocessing, data fusion, feature selection, ensemble learning, and domain knowledge. By using these strategies, it is possible to improve the quality and reliability of the mining results and to extract valuable insights from the data. In the context of web applications, workflow mining can be used to analyze the behavior of users as they interact with a website. There are several approaches to workflow mining in web applications, including: *Process discovery:* This approach involves extracting process models from event logs. Process models can be represented in various notations, such as Petri nets, BPMN, or EPC. Process discovery techniques include alpha algorithm, heuristics miner, fuzzy miner, genetic algorithm, and many others. *Conformance checking:* This approach involves comparing a process model (as designed) to an event log (as executed) to identify deviations or exceptions. Conformance checking techniques include token replay, alignments, and log-based metrics. *Performance analysis:* This approach involves analyzing the performance of a process model in terms of efficiency, effectiveness, and quality. Performance analysis techniques include process mining-based performance analysis, predictive analytics, and process simulation. *Resource analysis:* This approach involves analyzing the utilization of resources (such as people, machines, and systems) in a process. Resource analysis techniques include workload analysis, bottleneck analysis, and capacity planning. When comparing these approaches, each has its strengths and weaknesses. Process discovery is effective for identifying the actual process flows from event logs, but it may not be able to handle complex processes with many variants. Conformance checking is useful for identifying deviations, but it may not be able to detect certain types of deviations such as invisible tasks. Performance analysis is useful for identifying performance issues, but it may not be able to capture the context and reasons behind the performance issues. Resource analysis is useful for identifying resource utilization issues, but it may not be able to capture the overall process performance. Ultimately, the choice of approach will depend on the specific needs and objectives of the organization. In many cases, a combination of these approaches may be needed to fully understand and improve web-based processes.

6. Conclusion

Optimizing workflow execution is a critical task in data mining, and researchers have explored numerous approaches to improve the performance and accuracy of data mining workflows. The related work suggests that the choice of workflow execution method and platform should be carefully considered based on the specific requirements of the data mining project, including the size of the data set, the complexity of the workflow, and the available resources. Various techniques have been proposed to optimize workflow execution in data mining, including distributed execution, task scheduling algorithms, workflow optimization, and machine learning approaches. In addition, workflow management systems (WFMS) provide a useful framework for designing, executing, and monitoring data mining workflows. Selecting the right approach and platform for workflow execution can have a significant impact on the efficiency and accuracy of data mining tasks. By carefully evaluating the strengths and limitations of each approach, researchers can optimize their workflows to achieve the best possible results.

References

[1] Amy H.L. Lim; Chien-Sing Lee,; Murali Raman, (2012): Hybrid genetic algorithm and association rules for mining workflow best practices, Expert Systems with Applications, Volume 39, Issue 12, Pages 10544-10551, ISSN 0957-4174.

[2] Choudhary, A.; Singh, P.; & Kundra, H. (2015): Comparative study of Hadoop MapReduce, Mahout and Spark for data mining. International Journal of Computer Science and Mobile Computing, 4(1), 25-33.

[3] Garg, S., Sharma, M., & Kumar, A. (2018): Performance analysis of cloud platforms for data mining applications. International Journal of Computer Science and Mobile Computing, 7(3), 116-124.

[4] Janez Kranjc, Roman Orač, Vid Podpečan, Nada Lavrač, Marko Robnik-Šikonja, ClowdFlows: Online workflows for distributed big data mining, Future Generation Computer Systems, Volume 68, 2017, Pages 38-58, ISSN 0167-739X.

[5] Jasper P. J, (2021): Trace reconstruction in system logs for processing with process mining, Procedia Computer Science, Volume 180, Pages 352-357, ISSN 1877-0509.

[6] Markus Hammori, Joachim Herbst, Niko Kleiner, Interactive workflow mining—requirements, concepts and implementation, Data & Knowledge Engineering, Volume 56, Issue 1, 2006, Pages 41-63, ISSN 0169-023X.

Authors Profile

Dr. Lalita Kumari is currently working as Assistant Professor at Amity School of Engineering and Technology (ASET), Amity University, Patna. She has 14+ years of experience in teaching years of teaching experience. Her area of research includes Machine Learning, Digital Image Processing, Natural Language Processing, Advanced Algorithm.

Dr. Niranjan Kumar is currently working as Assistant Professor at Amity School of Engineering and Technology (ASET), Amity University, Patna. He has more than 25 years of Industrial as well as Teaching Experience, around 10 years he has worked with a diverse range of systems,

Miss Seemi Kumari is currently studying in B.Tech third year at Amity School of Engineering and Technology (ASET), Amity University, Patna.