Modeling of the phases of the educational data mining through workflow networks

Emilcy J. Hernández-Leal[1], Néstor D. Duque-Méndez[1]

[1] Universidad Nacional de Colombia, Sede Manizales
ejhernandezl@unal.edu.co, ndduqueme@unal.edu.co

Abstract. Workflow networks are derived from Petri networks and are used for the modeling of systems, processes, and procedures. The objective of this work is to explore the application of workflow networks for the modeling of the phases of the data mining process and to apply it to a particular case for educational data. Given that in the data mining process it is possible to find some problems related to duplicate tasks, tasks that become invisible and presence of noise in the data, it is proposed that through the workflow networks these mining processes can be previously modeled and identify the problems mentioned in time. To make the simulation of the networks, the Workflow Petri net Designer (WoPeD) software version 3.5.1 is used. With the results achieved it can be demonstrated that it is possible to apply this approach to inspect the planning of data mining processes and for the particular case of educational data, to verify that they have the necessary conditions and resources to execute the techniques and analysis.

Keywords: Educational Data, Data Mining, Modeling, Workflow Networks.

1 Introducción

Los procesos de descubrimiento de patrones, que incluyen la minería de datos, son usados para el tratamiento y análisis de datos provenientes de diferentes dominios o campos de estudio, contienen una gran variedad de técnicas, pero en general se rigen por una serie de etapas que van desde la integración de las fuentes de datos, pasando por el tratamiento, limpieza, carga, almacenamiento, hasta los procesos de análisis como tal y la visualización de resultados [1].

La minería de datos se ha aplicado a datos educativos, en este caso toma el nombre de minería de datos educativa (EDM por sus siglas en inglés) [2]. Para este trabajo son empleadas Redes Workflow con el fin de simular un proceso de minería de datos aplicada al dominio de datos educativos y evaluar el funcionamiento de un modelo de análisis y minería de datos educativos y de interacciones en plataformas virtuales de aprendizaje para una institución de educación superior. Para lo anterior se ha planteado un

modelamiento del proceso de minería de datos y se ha especificado al caso particular de los datos educativos que alimentarán el proceso.

El documento se organiza de la siguiente forma: en la sección 2 se presenta el referente teórico de las Redes Workflow y se describe cómo funcionan los procesos de minería de datos y cada una de sus etapas. En la sección 3 se presenta el caso particular que se desea modelar por medio de las redes, dando un contexto general del dominio de datos educativo a trabajar. En la sección 4 se muestra el modelamiento y simulación del proceso y la discusión de los resultados obtenidos. Se finaliza con la sección 5 en la cual se traen a colación las conclusiones y se expone el trabajo futuro.

2 Referente teórico

A continuación, se describe brevemente algunos conceptos relacionados con la propuesta:

2.1 Redes Workflow

Las Redes de Petri, fueron introducidas por Carl Adam Petri en la década de los sesenta, de allí su nombre. En general, las redes Petri se usan como una herramienta matemática y gráfica para el estudio y modelado de diferentes sistemas. Con este tipo de redes se pueden analizar de forma completa diferentes fenómenos. Además, existen varias subclases de las Redes Petri habituales como: Red Petri Ordinaria, Red Petri Simple, Grafo Marcado, Máquina de Estados, Red de Libre Elección, Red Petri Lugar, entre otras [3].

Se han derivado también de las redes Petri otros tipos de redes como las Workflow que poseen diversas técnicas de análisis eficiente. En términos de ecuaciones se pueden definir de la siguiente forma [4]: Una Red Workflow $N= (P,T,F,\alpha,\Omega)$ es una red Petri (P,T,F) donde P es un conjunto finito de lugares, T es un conjunto finito de transiciones y F los arcos, con un lugar de inicio distinguido α que pertenece a P y una transición final distinguida $\emptyset \neq \Omega \subseteq T$, tal que

1. Para todo $p \in P$ se cumple $\bullet p=\emptyset$ implica $p=\alpha$,
2. para cada $t \in T$ se cumple $t\bullet=\emptyset$ implica $t\in\Omega$, y
3. cada nodo $x\in P\cup T$ está en una trayectoria desde el lugar inicial α hasta alguna transición final $\omega\in\Omega$.

Las redes Workflow están compuestas por tareas que son ejecutadas en un orden específico. Una tarea es una actividad o un evento y se puede asumir que cada tarea es atómica, cuando esta inicia existen solo dos posibilidades, que finalice satisfactoriamente o que falle. Una implicación importante de esta suposición es que todos los recursos requeridos para finalizar la tarea serán retenidos por la tarea hasta que ésta finalice o falle [5]. Las Redes Workflow pueden tener varios patrones de enrutamiento, es

decir, de paso de una transición a otra; hay cuatro patrones básicos: secuencial, iterativo, paralelo y selectivo [6], los cuales son representados en la Fig. 1 como (a), (b), (c) y (d) respectivamente.

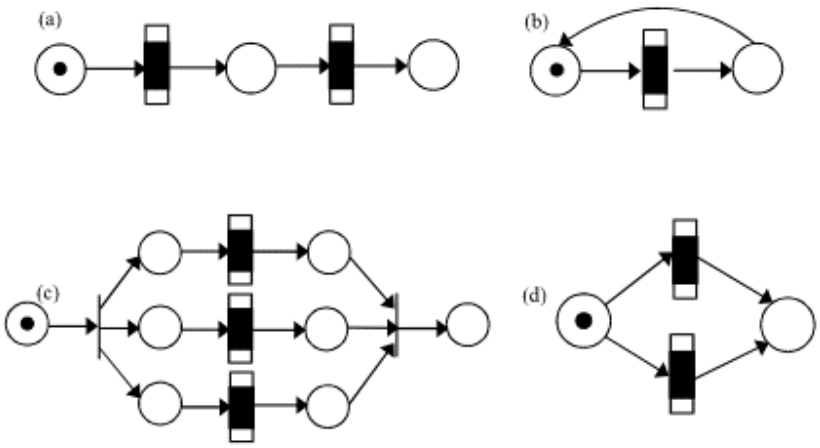

Fig. 1. Ejemplos de patrones de enrutamiento en las redes Workflow
Fuente: Tomado de [6]

2.2 Tareas o etapas de la minería de datos

Con el aumento general en la producción de datos y su disponibilidad, los procesos de minería de datos han tomado relevancia. Estos procesos tienen como fin descubrir, monitorear y mejorar procesos reales para extraer conocimiento desde registros de eventos.

La minería de datos suele estar enmarcada en el proceso de descubrir conocimiento a partir de una base de datos, KDD, el cual describe una secuencia de etapas, donde cada etapa es primordial y juega un papel en la transformación de los datos en conocimiento. El proceso incluye la selección de las fuentes de datos, el tratamiento y almacenamiento de los mismos, el análisis a través de métodos estadísticos, algoritmos de minería de datos u otras técnicas y la evaluación, interpretación y visualización de resultados. Al final de las etapas de un proceso de minería de datos se suele obtener un modelo descriptivo, que luego puede ser convertido en un modelo predictivo, de ser necesario [7]. Para efectos de este trabajo se denomina minería de datos al proceso completo de KDD, en la Fig. 2 se presentan las etapas generales de la minería.

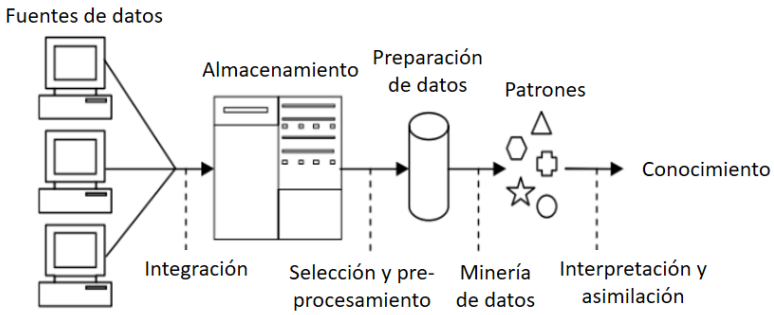

Fig. 2. Proceso de minería de datos
Fuente: Adaptado de [8]

3 Descripción del proceso de Minería de Datos Educativa a modelar

En [9] se desarrolló un modelo para el descubrimiento de patrones basado en el análisis de datos educativos y de las interacciones existentes entre los estudiantes y las plataformas virtuales de aprendizaje, con el uso de minería de datos educativa y analíticas de aprendizaje, que puede contribuir a la realización de algunas recomendaciones para fortalecer el proceso de enseñanza y aprendizaje, de manera que este se pueda adaptar y posiblemente personalizar de acuerdo a las características propias de los estudiantes y de sus interacciones. El modelo mencionado consta una serie de componentes, los cuales se pueden apreciar en la Fig. 3. A su vez estos componentes se particularizan en una serie de etapas y tareas que son mostradas en la Tabla 1.

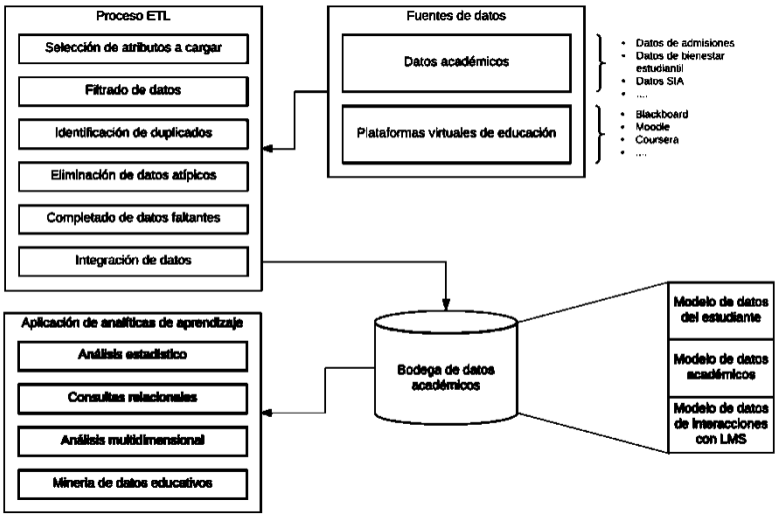

Fig. 3. Modelo propuesto para el descubrimiento de patrones en datos educativos.
Fuente: Tomado de [10].

Tabla 1. Fases y tareas contempladas en el modelo propuesto.

Fases	Tareas
1. Comprensión del dominio de datos	1.1 Identificación del dominio de datos 1.2 Revisión de conocimiento previo relevante en el dominio de datos 1.3 Identificación de los objetivos del usuario final
2. Recolección de fuentes de datos	2.1 Identificación de fuentes: datos históricos, datos en tiempo real, data stream, entre otras 2.2 Identificación de las estructuras de datos: datos estructurados, semi-estructurados y no estructurados 2.3 Caracterización de las fuentes de datos
3. Pre-procesamiento y limpieza de datos	3.1 Modelamiento de filtros para identificar datos atípicos y datos faltantes 3.2 Eliminación de ruido y datos atípicos 3.3 Identificación y estrategias de llenado de datos faltantes
4. Almacenamiento de datos	4.1 Revisión de la caracterización de las fuentes de datos 4.2 Selección de la estrategia (s) de almacenamiento 4.3 Construcción del esquema de almacenamiento y poblado de datos
5. Selección y adaptación del dataset	5.1 Selección de variables 5.2 Reducción de dimensionalidad y/o aplicación de métodos de transformación de datos
6. Aplicación de técnicas – algoritmos de Minería de Datos	6.1 Selección de la tarea de minería: predictiva o descriptiva 6.2 Selección del algoritmo para la tarea 6.3 Selección de la herramienta para la aplicación del algoritmo
7. Análisis del conocimiento descubierto	7.1 Interpretación de patrones extraídos 7.2 Consolidación del conocimiento descubierto 7.3 Evaluación del conocimiento descubierto

Fuente: Elaboración propia a partir de [10]

Para analizar el comportamiento del flujo de datos en el modelo propuesto e identificar si es posible llegar a los análisis esperados a través de las fases y tareas propuestas, se transfirió a una representación formal y procesable, para así poder identificar también el cumplimiento de las condiciones previas y si existen casos de ausencia de recursos. En la sección siguiente se describe el proceso de modelo y análisis que fue llevado a cabo con ayuda de un software.

4 Modelado y análisis de un proceso de KDD y Minería de Datos Educativa

En los procesos de minería de datos existen algunas deficiencias, que en ocasiones no se pueden evitar, como las tareas duplicadas, las tareas invisibles y el ruido en los datos; las redes Workflow se han utilizado para el modelado de procesos de KDD y en particular de minería de datos, mostrando que con el uso de estas, los problemas enunciados pueden llegar a ser identificados a tiempo [11]. Adicionalmente, se ha registrado que en la aplicación de las técnicas de minería de datos se pueden encontrar otro tipo de problemas al tratar grandes volúmenes de registros de eventos que hacen referencia a diferentes actividades y de allí se ha evidenciado, que es conveniente descomponer los procesos de minería para poderlos analizar con mayor facilidad [12].

De acuerdo a lo anterior, se decidió usar las redes Workflow para representar en primer lugar el proceso de minería de datos (ver Fig. 4) y luego el modelo de descubrimiento de patrones en datos educativos descrito en la sección anterior (ver Fig. 5), haciendo uso del software WoPeD en versión 3.5.1, este software es de código abierto y cuenta con licencia LGPL, su fin es proporcionar una herramienta de modelado sencilla de utilizar para hacer simulación y análisis de flujo de procesos y descripción de recursos utilizando redes de control; WoPeD está dirigido a investigadores, docentes y estudiantes que se encuentran trabajando con la aplicación de las redes de Petri y redes Workflow.

Se identificaron en primer lugar las condiciones, como por ejemplo la necesidad de comprender el dominio de datos y de recolectar las fuentes de datos; y los recursos, como por ejemplo los datos como tal o el dataset ya construido. Las fases se representaron en nodos tipo transición y se unieron con los nodos tipo lugar por medio de arcos orientados, para mostrar la secuencia lógica definida en el modelo estudiado.

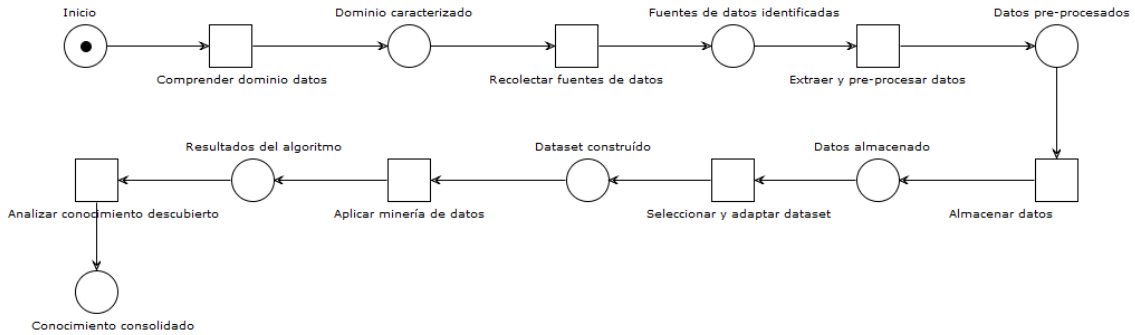

Fig. 4. Red Workflow para el proceso general de KDD y minería de datos
Fuente: Elaboración propia mediante el software WoPeD v. 3.5.1

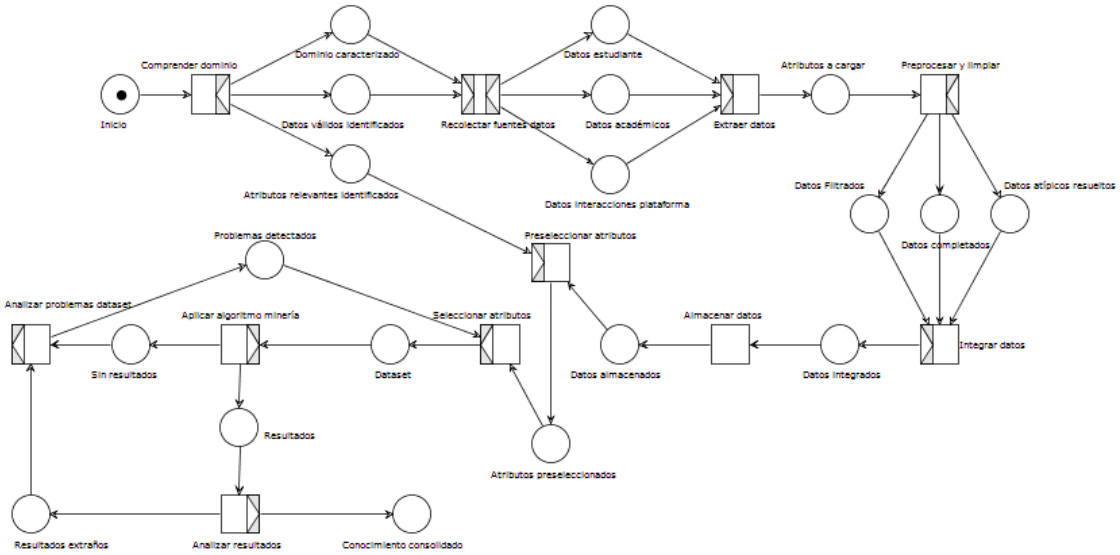

Fig. 5. Red Workflow correspondiente al modelo analizado
Fuente: Elaboración propia mediante el software WoPeD v. 3.5.1

Después de construir la red de la Fig. 5 se verificó que las tareas de las fases del modelo propuesto se vieran reflejadas en este, luego se dio inicio a la ejecución de la simulación y se logró cotejar que era posible llegar al nodo de lugar final a partir del nodo de partida y pasando por las transiciones correspondientes con la secuencia de disparos. En la Tabla 2 se presenta el análisis estructural y de robustez para las propiedades de la red construida.

Tabla 2. Análisis de las propiedades de la red Workflow que representa el modelo

Tipo de análisis	Elementos analizados	Resultado
Análisis estructural	Nodos tipo lugar	20
	Nodos tipo transición	15
	Operadores	10
	Arcos	42
	Operadores usados erróneamente	0
	Violaciones de libre elección	0
Robustez (soundness)	Lugar inicial	01
	Lugar final	01
	Componentes conectados	31
	Componentes fuertemente conectados	31
	Lugares no acotados (boundness)	0
	Transiciones muertas (dead-lock)	0
	Transiciones no vivas (non-live-transitions)	0

Fuente: Elaboración propia a partir del software WoPeD v. 3.5.1

Para el análisis estructural se muestran, en la Tabla 2, elementos como los nodos tipo lugar, transición, operadores y arcos; así mismo no se encuentran operadores usados erróneamente ni hay violaciones de libre elección. Para el análisis de robustez, que refleja el punto de vista funcional, se dan a conocer algunos aspectos básicos como el número de lugar inicial, lugar final, componentes conectados y fuertemente conectados; también otros indicadores como los lugares no acotados, las transiciones muertas y las transiciones no vivas, los cuales al estar en cero demuestran que la red que representa el modelo de descubrimiento de patrones de datos educativos no tiene bloqueos de ejecución, por lo cual, se pueden llevar a cabo las fases y tareas derivadas del proceso general de minería de datos.

5 Conclusiones y trabajo futuro

Se logró representar las fases del modelo de descubrimiento de patrones en datos educativos mediante el uso de Redes Workflow y se comprobó que es posible su aplicación para llegar a los resultados esperados representados en el conocimiento consolidado, que permitirá luego la realización de algunas recomendaciones para fortalecer el proceso de enseñanza y aprendizaje y la posible adaptación y personalización de acuerdo a las características propias de los estudiantes y de sus interacciones.

Esta contribución corresponde a un primer paso para corroborar que este tipo de procesos de minería de datos pueden simulados con anterioridad a su ejecución en un entorno real. Con lo cual se contribuye a evitar posibles fracasos en procesos experimentales y reducir la incertidumbre inicial del proceso. Además, se determinó que los nodos tipo lugar y las transiciones simuladas corresponden a una aproximación considerable de lo que se puede observar en la ejecución en un entorno real de este tipo de procesos.

Se ratifica que las Redes Workflow son adecuadas para la representación de procesos de minería de datos a partir de un conjunto de datos educativos, ya que permiten ejecutar los flujos entre las etapas y comprobar la capacidad de cumplimiento de los objetivos del modelo de análisis de datos respectivo. A pesar de que estos procesos de minería de datos se suelen ver como flujos lineales, se encuentra que al revisar cada fase y tareas específicas, se encuentran múltiples condiciones de entrada y salida de algunas de las fases y se refleja su complejidad; así mismo, se demuestra la importancia de una fase inicial de comprensión del dominio de datos, la cual produce salidas que alimentan no solo la fase contigua sino también otras fases posteriores en el proceso.

Como trabajo futuro se plantea hacer una revisión más exhaustiva de cada una de las fases y tareas, explorando la inclusión de otros aspectos como fuentes de datos y detallando la transición correspondiente a la aplicación del algoritmo de minería. Con lo anterior se espera poder proponer un modelo de minería de datos educativa no lineal, que posteriormente pueda ayudar para la automatización de las fases del mismo.

Teniendo como base está simulación del proceso de minería de datos aplicado al dominio de datos educativo, se planea validar con el caso de estudio que se propone en [10], el cual considera un entorno real.

Agradecimientos

Al programa de Formación de Capital Humano de Alto Nivel para el Departamento de Norte de Santander en el marco de la Convocatoria N°753 de Colciencias.

Referencias

[1] S. H. Begum, "Data Mining Tools and Trends – An Overview," *Int. J. Emerg. Res. Manag. &Technology*, pp. 6–12, 2013.

[2] R. S. Baker and P. S. Inventado, "Educational Data Mining and Learning Analytics," in *Learning Analytics*, New York, NY: Springer New York, 2014, pp. 61–75.

[3] M. L. Llorens Agost, "Redes Reconfigurables. Modelizacíon y Verificacíon," Universidad Politécnica de Valencia, 2003.

[4] C. Favre, D. Fahland, and H. Völzer, "The relationship between workflow graphs and free-choice workflow nets," *Inf. Syst.*, vol. 47, pp. 197–219, 2015.

[5] J. Wang and D. Li, "Resource oriented workflow nets and workflow resource requirement analysis," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 23, no. 05, pp. 677–693, Jun. 2013.

[6] S. Jing and Y. Du, "An Approach of Data Mining Process Based on Stochastic Well-formed Workflows," *Inf. Technol. J.*, vol. 13, no. 13, pp. 2224–2228, 2014.

[7] N. L. Quiroz Gil and C. A. Valencia, "Aplicación del proceso de KDD en el contexto de bibliomining: El caso Elogim," *Rev. Interam. Bibl.*, vol. 35, no. 1, pp. 97–108, 2012.

[8] M. A. Bramer, *Principles of data mining*. Springer, 2013.

[9] N. D. D. Méndez, M. G. Ocampo, and F. Moreira, "Storage Scheme for Analysis of Academic Data and Interaction of Students With Virtual Education Platforms," in *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality - TEEM 2017*, 2017, pp. 1–7.

[10] M. Giraldo Ocampo, "Descubrimiento de patrones en interacciones entre estudiantes y plataformas virtuales de educación mediante el uso de analíticas de aprendizaje," Universidad Nacional de Colombia, 2017.

[11] J. Wang, S. Yu, and Y. Du, "The Equivalency between Logic Petri Workflow Nets and Workflow Nets," *Sci. World J.*, pp. 1–7, 2015.

[12] W. M. P. van der Aalst, "Decomposing Petri nets for process mining: A generic approach," *Distrib. Parallel Databases*, vol. 31, no. 4, pp. 471–507, Jul. 2013.