Búsqueda de patrones en un dominio representado en una base de datos de grafos dirigidos

María Alejandra Paz Menvielle (pazmalejandra@gmail.com), Cynthia Lorena Corso (corso.cynthia@gmail.com), Analía Guzmán (aguzman@frc.utn.edu.ar), Martín Gustavo Casatti (mcasatti@frc.utn.edu.ar), Karina Ligorria (karinaligorria@gmail.com)

CIDS-Centro de Investigación, Transferencia y Desarrollo de Sistemas de Información
Departamento de Ingeniería en Sistemas de Información
Facultad Regional Córdoba – Universidad Tecnológica Nacional
Maestro Marcelo López esq. Cruz Roja Argentina – Córdoba 0351 – 4686385

Resumen—El presente trabajo expone las actividades desarrolladas a fin de realizar la búsqueda de patrones en una base de datos de grafos dirigidos.

Se trabajará sobre un sistema de evaluación automática que permite calificar exámenes compuestos por preguntas que serán respondidas por alumnos de nivel universitario, con escritura de texto libre, en un dominio acotado. En este contexto se presenta la situación actual de la base de datos de grafos, que contiene las respuestas mencionadas, y su adaptación a fin de aplicar la búsqueda de patrones en la misma.

Para explicar claramente el proceso de aplicación de los patrones se presentan definiciones de conceptos y componentes relacionados a los grafos y a los patrones trabajados en grafos, específicamente su aplicación al análisis de textos.

Avanzando al detalle del trabajo se exponen las métricas existentes y las que han sido seleccionadas para ser aplicadas en el análisis de los patrones, destacando la importancia de las mismas en relación a las hipótesis planteadas en el proyecto de investigación. Posteriormente se presentan las fases que serán abordadas en el trabajo de investigación y los avances alcanzados en ellas, finalizando con las conclusiones a las que se ha arribado y las actividades de investigación con las que se continuará.

Palabras clave—análisis de texto; grafos; detección de patrones; diseño de datos; graph minning

1. CONTEXTO

El presente trabajo forma parte del proyecto de investigación y desarrollo que ha sido homologado por la Secretaría de Investigación, Desarrollo y Posgrado de la Universidad Tecnológica Nacional, desarrollado en el ámbito del CIDS – Centro de Investigación, Desarrollo y Transferencia en Sistemas de Información. Para su desarrollo se utiliza como caso testigo a la cátedra de Paradigmas de Programación, perteneciente a la carrera Ingeniería en Sistemas de Información, dictada en la Facultad Regional Córdoba , de la Universidad Tecnológica Nacional.

El trabajo que aquí se presenta es la continuación de los trabajos realizados durante el desarrollo del Proyecto "Metodología para determinar la exactitud de una respuesta, escrita en forma textual, a un interrogatorio sobre un tema específico", (PID EIUTNCO0003592). Durante el transcurso de dicho proyecto se generó una base de conocimiento modelada como grafo dirigido, que posee preguntas y respuestas de exámenes escritas en forma textual, así como cualquier otro concepto contenido en el programa de estudios de la materia Paradigmas de Programación. Dicho grafo es una base de conocimientos amplia y extensible, generada inicialmente por los docentes de la cátedra, utilizada para evaluar a los alumnos y que pueda ser alimentada con las respuestas que no hayan sido consideradas aún y que aporten mayor variedad a la base de conocimiento.

En el presente trabajo se propone complementar la funcionalidad del proyecto anteriormente mencionado, creando un segundo grafo, obtenido a partir de las respuestas de todos los alumnos que realizan los exámenes y que almacene el historial de las respuestas de los alumnos, para poder cotejarlo después con el grafo inicial y así poder realizar la búsqueda, el análisis y la propuesta de patrones frecuentes en el grafo conceptual.

Con el análisis de patrones se pretende descubrir algunas características importantes que se relacionen con las respuestas de los alumnos y que puedan reflejar la evaluación y el aprendizaje de los mismos. También que se pueda determinar el nivel de profundidad con el que se evalúa cada tema, si los distintos exámenes son consistentes en alcance y profundidad de evaluación de conceptos, entre otros. En relación a los contenidos de la materia, se pretende identificar si hay áreas de conocimiento evaluadas con mayor frecuencia, si hay temas que no se están evaluando, qué conceptos de la materia son mejor conocidos por los alumnos y cuáles lo son en menor medida. Esto permitirá, en última instancia, mejorar los materiales didácticos empleados como así también ajustar los instrumentos de evaluación.

2. INTRODUCCIÓN

Un patrón es una entidad a la que se le puede dar un nombre y que está representada por un conjunto de propiedades medidas y las relaciones entre ellas (vector de características)[6].

En el dominio utilizado como caso testigo, por ejemplo, un patrón puede ser la ruta resultante de una respuesta de un alumno, de las cuales se extrae el vector de características formado por un conjunto de valores numéricos que pueden representar nivel de exactitud de la respuesta, la puntuación de la misma, la cantidad utilizada de conceptos y de relaciones, etc.

El reconocimiento automático, descripción, clasificación y agrupamiento de patrones son actividades importantes en una gran variedad de disciplinas científicas, como biología, sicología, medicina, visión por computador, inteligencia artificial, teledetección, etc. Lo importante de detectar patrones en los datos es que se pueden inferir causas para la agrupación de los mismos (en el caso de que estemos detectando patrones ya conocidos y ya estudiados).

Este trabajo, tiene paralelos con el llamado SNA (Social Network Analysis, Análisis de Redes Sociales) que es una disciplina cuyo objetivo es "Analizar la estructura de una red social para *inferir conocimiento* de un individuo, un grupo, o las relaciones entre ellos"[7]).

Un grafo conceptual[1] es un sistema de notación simbólica y de representación del conocimiento. Presentado por John F. Sowa, se basa en los gráficos existenciales[8][2] de Charles Sanders Pierce, en las estructuras de redes semánticas y en datos de la lingüística, la filosofía y la psicología.

Se describen a continuación conceptos fundamentales relacionados a la construcción de grafos: Nodo: Un nodo es un *ítem de conocimiento* (término, entidad) diferente de cualquier otro conocimiento en el modelo. Todos los conceptos que tienen su propio significado están representados como nodos. Un nodo representa la mínima cantidad de conocimiento que puede ser representada, la cual no puede ser dividida. Unidades de conocimiento mayores pueden representarse mediante grupos de nodos relacionados. Es importante mencionar que los nodos no representan grupos de entidades. Para los fines que se persiguen es necesario que cada nodo pueda ser identificado de manera unívoca y diferente de todos los demás nodos de la red.

Enlace: Un enlace se utiliza para establecer una relación entre dos nodos del modelo. Un enlace solo puede conectar dos nodos. Uno denominado *origen* desde el que sale el enlace y uno denominado *destino* al cual llega. Una relación entre un nodo de origen y dos nodos de destino requiere de dos enlaces, ambos con el mismo origen, pero con diferentes destinos. Un nodo puede ser origen de varios enlaces, así como puede ser destino de varios enlaces[9].

2.1. Patrones en Grafos

El reconocimiento o detección de patrones dentro de grafos busca detectar un subgrafo (patrón) en un grafo (objetivo). Debemos considerar que esta búsqueda de coincidencias se puede descomponer en dos partes:

1. Una concordancia estructural, en donde los nodos y relaciones del patrón conforman una estructura existente en el grafo objetivo.
2. Una concordancia a nivel de elementos, en donde los nodos y relaciones, a nivel de sus atributos particulares, tiene los mismos valores que en la estructura encontrada en el grafo objetivo.

Muchas veces la búsqueda de estas dos concordancias se ejecuta de forma separada para optimizar los algoritmos o reducir el espacio de búsqueda[10].

En el dominio bajo estudio, la detección de un subgrafo (patrón), se realizará en los grafos (objetivos) que representan los contenidos de la materia y las respuestas de los alumnos en las instancias de exámenes.

2.2. Métricas en grafos

Una herramienta ampliamente utilizada para describir grafos y que muchas veces se utiliza para iniciar el análisis de patrones existentes en los mismos, es el cálculo de métricas[11], locales o globales, que permiten caracterizar el grafo objetivo o el grafo patrón. Las métricas se pueden dividir en dos grandes grupos:

- Métricas estáticas: Cuando se calculan sobre un grafo estático en un punto en el tiempo determinado. Se enfocan principalmente en las características estructurales del mismo.
- Métricas dinámicas: Tienen en cuenta la dimensión temporal de los cambios que se producen sobre el grafo. Están más enfocadas en las variaciones entre dos instantes de tiempo, antes que en las características propias del grafo en cada uno de esos instantes.

Otro enfoque para el análisis de las métricas radica en analizar sobre qué componentes del grafo se realizan las mediciones. Desde este punto de vista se tienen diversas perspectivas, siendo las más comunes:

- Métricas de redes (o globales): Son las métricas que toman como referencia el grafo completo, con todos los nodos y arcos que lo conforman.
- Métricas nodos (o locales): Son aquellas que toman como referencia un nodo o subconjunto de nodos para realizar los cálculos.

A continuación, se detallan algunas métricas más comunes:

2.2.1. Métricas globales:

Centralidad: Esta métrica trata de determinar que nodo o nodos ocupan una ubicación central en la red, estando equidistante de los demás nodos.

Conexionado: Busca establecer el grado en el que los nodos de un grafo están conectados con todos los demás nodos del mismo. Se puede encontrar, aplicando esta métrica, componentes fuertemente conectados o débilmente conectados.

Cantidad de Componentes: En un grafo que no es completamente conexto, indica la cantidad de subgrafos conexos que forman parte del grafo. Un componente es un conjunto de nodos conectados que forman parte del grafo principal.

Tamaño del componente gigante: Mide la cantidad de nodos que tiene el componente conectado que es mayor que todos los demás componentes del grafo. En un grafo conexo

el tamaño del componente gigante es igual a la cantidad total de nodos.

Ruta más corta/larga: Expresa la longitud (en arcos) mínima/máxima entre dos nodos dados.

2.2.2. *Métricas locales:*

Conectividad: Expresa la cantidad de conexiones que posee un nodo determinado. Se puede expresar como 'grado', si no tiene en cuenta la dirección de los arcos que inciden o salen del nodo, o como 'grado de entrada' o 'grado de salida' cuando solamente tiene en cuenta los arcos entrantes o salientes, respectivamente.

Centralidad: Es una métrica, asociada a un nodo en un grafo, que determina su importancia relativa dentro de éste, pudiendo dividirse en:

- Centralidad de grado: Cantidad de conexiones con otros nodos
- Centralidad de cercanía: Indica qué tan cerca se encuentra una unidad de la red de otras.
- Centralidad de intermediación: Indica si una unidad se encuentra dentro de algunas de las rutas más cortas que existen entre dos nodos de la red.

2.3. Análisis de Patrones

El análisis de patrones en el dominio bajo estudio, puede determinar si existen ciertos patrones que, aún, no siendo comunes en otras áreas de la teoría de grafos, si lo son recurrentes en este dominio. Se pueden determinar si son patrones temporales, es decir que tiendan a desaparecer en el tiempo a medida que la base de conocimientos va cambiando, o si son patrones permanentes y/o que se van reforzado con el tiempo.

Dicho análisis puede servir para descubrir algunas características importantes que se relacionan con el aprendizaje, entre ellas:

- los temas que revisten más dificultad de aprendizaje,
- la cantidad y tipos de errores más comunes y su relación con el tema o concepto evaluado,
- las tendencias de los alumnos al momento de responder las mismas preguntas, es decir, si lo hacen con los mismos conceptos o, por el contrario, tienen una riqueza expresiva alta.
- Se propone incluir una respuesta textual, utilizada como patrón, para poder determinar si las respuestas dadas por los alumnos tienen una correspondencia directa (literal) con respecto al material brindado para su estudio.

2.4. Diseño de reconocimiento de patrones

El objetivo principal de un sistema de reconocimiento automático de patrones es descubrir la naturaleza subyacente de un fenómeno u objeto, describiendo y seleccionado las características fundamentales que permitan clasificarlos en una categoría determinada[12][13].

Sistemas automáticos de reconocimiento de patrones permiten abordar problemas en informática, en ingeniería y en otras disciplinas científicas[14][15], por lo tanto el diseño de cada etapa requiere de criterios de análisis conjuntos para validar los resultados[16][17].

Luego de analizar diferentes formas de diseñar un sistema de reconocimiento de patrones, se consideran tres fases[18]:

1. Adquisición y preproceso de datos.
2. Extracción de características.
3. Toma de decisiones o agrupamiento.

En la fase de Adquisición y preproceso de datos, se preparará la infraestructura de la base de datos para poder continuar con las siguientes fases.

Para las dos siguientes fases, extracción de características y toma de decisiones o agrupamiento, se considerarán los objetivos y contenidos de la materia, las necesidades de los docentes de la cátedra, respecto a la información que les podría brindar el prototipo de calificación y en general, las características que son de interés para evaluar o estudiar los mecanismos de evaluación de la materia y el grado de aprendizaje de los alumnos. También se considerarán algunas métricas relacionadas a los patrones que se identifiquen.

3. DESARROLLO

3.1. Adquisición y preproceso de datos

En la introducción del presente trabajo, se mencionó que se cuenta con la base de conocimiento de la materia, donde están almacenados en un grafo, los datos adquiridos y procesados relacionados a los contenidos, las preguntas y las respuestas, tanto las provistas por los docentes como las respondidas por los alumnos.

Esta infraestructura surge del proyecto anterior, según se ven reflejados en los trabajos publicados [4, 5], donde se expone la arquitectura, las tecnologías utilizadas y los resultados obtenidos a través de una evaluación realizada por medio de un prototipo desarrollado para registrar los datos en una base de datos de grafos y realizar las consultas necesarias para la calificación.

La tecnología que se utilizó para crear el prototipo incluye herramientas de código abierto, ya que el uso principal se realiza en un ambiente universitario y por eso es de fundamental importancia no depender de ningún tipo de licenciamiento propietario.

La implementación se realizó en el lenguaje Java, ya que éste posibilita implementar la arquitectura en multicapa e intercambiar fácilmente los componentes que ofrece. La característica multiplataforma de Java, es muy importante en un proyecto de investigación relacionado con la educación universitaria, en donde es muy factible que distintas unidades académicas, posean distintas infraestructuras de hardware y software para implementar una solución de este tipo. A esto se debe agregar la buena integración que tiene con las librerías de corrección ortográficas, la base de datos de grafo y la librería de visualización, todas desarrolladas en este mismo lenguaje.

Para el manejo de la base de datos de grafos se utilizó el producto OrientDB Community Edition, es una aplicación de código abierto, con licencia Apache 2 y gratuita para todo tipo de uso. Esta base de datos de grafos implementa de forma nativa dos características que son centrales en el planteo del método de corrección: *nodos y arcos etiquetados* como así también *tipos de datos complejos* como atributos de los nodos.

Esas dos características hicieron posible que la implementación de los modelos teóricos planteados fuera directa, con el consiguiente ahorro en tiempos de desarrollo y simplicidad a la hora de utilizar dichos conceptos como parte del método de corrección.

Se utilizó, además, la librería gráfica GraphStream, para realizar la visualización de la base de grafos, que es una librería Open Source implementada en Java que provee toda la funcionalidad de visualización y trazado de rutas con una gran flexibilidad y facilidad de uso. El prototipo hace un uso intensivo de las propiedades de ruteo disponibles en la librería GraphView de forma tal que la visualización es clara y con la menor cantidad de cruces de líneas entre los nodos.

Sin embargo, para tener información relevante del dominio elegido y poder detectar patrones, se necesita agregar una base de grafos que contenga las respuestas de los alumnos y que se registre automáticamente cuando el sistema califique las respuestas de los alumnos en las distintas instancias de los exámenes.

Para poder implementar dichas actualizaciones, se ha realizado el siguiente estudio y análisis:

3.1.1. Factores para la representación del conocimiento

Actualmente se deben tener en cuenta al menos cuatro factores fundamentales a la hora de diseñar un sistema de representación del conocimiento en cualquier dominio dado[3]:

- **Adecuación Representacional:** Habilidad para representar todas las clases de conocimiento que son necesarias en el dominio.
- **Adecuación Inferencial:** Habilidad de manipular estructuras de representación de tal manera que devengan o generen nuevas estructuras que correspondan a nuevos conocimientos inferidos de los anteriores.
- **Eficiencia Inferencial:** Capacidad del sistema para incorporar información adicional a la estructura de representación, llamada metaconocimiento, que puede emplearse para focalizar la atención de los mecanismos de inferencia con el fin de optimizar los cómputos.
- **Eficiencia en la Adquisición:** Capacidad de incorporar fácilmente nueva información. Idealmente el sistema por sí mismo deberá ser capaz de controlar la adquisición de nueva información y su posterior representación.

Estos factores se tuvieron en cuenta durante el proceso de diseño de las estructuras de datos de la base de conocimientos, de manera tal que se puedan maximizar los resultados a la vez que se mantienen eficientes las operaciones de cómputo.

3.1.2. Atributos necesarios para el calculo de métricas

Los enfoques para calcular las métricas sobre un grafo dirigido son amplios y muy variados en cuanto a rapidez, eficiencia, uso de recursos de procesamiento y exactitud de los resultados.

Sin embargo el requisito fundamental que tiene que tener cualquier grafo para poder calcular dichas métricas es que los nodos y arcos deben poseer los atributos necesarios para que, como parte del análisis, se determinen si deben ser incluidos o no en una cierta métrica.

En el marco del presente trabajo de investigación se cuenta con una ventaja importante ya que el grafo ha sido diseñado específicamente para poder calcular las métricas necesarias. Esto no disminuye de ninguna manera la aplicabilidad de los métodos de búsqueda sino que más bien existe la posibilidad de implementar optimizaciones debido al mayor conocimiento de la información a procesar y los resultados buscados.

Las métricas estándar de grafos se pueden obtener de la estructura misma del almacenamiento, sin necesidad de adjuntar etiquetas o atributos adicionales. La simple existencia de nodos y aristas permite calcular métricas generales tales como la centralidad, el conexionado, las longitudes de ruta, y los componentes conexionados.

3.2. Extracción de características y agrupamiento

Para las dos siguientes fases, extracción de características y toma de decisiones o agrupamiento, se consideraron los objetivos y contenidos de la materia, las necesidades de los docentes de la cátedra, respecto a la información que les podría brindar el prototipo de calificación y en general, las características que son de interés para evaluar o estudiar los mecanismos de evaluación de la materia y el grado de aprendizaje de los alumnos.

También se consideraron algunas métricas mencionadas en el apartado anterior, relacionadas a los patrones inicialmente identificados. Se identificaron dos grandes clases de patrones, aquellos relacionados a la evaluación y aquellos relacionados al aprendizaje de los alumnos.

A continuación, se detalla cada clase de patrones, junto con sus patrones asociados.

3.2.1. Patrones asociados con la evaluación

Estos patrones buscan caracterizar los mecanismos de evaluación utilizados por los docentes y establecer criterios para su mejora tanto en cuanto a profundidad de los temas evaluados como a variedad y relevancia de los conceptos utilizados en dicha evaluación. Los patrones identificados son:

Conceptos centrales: Se consideran conceptos centrales a aquellos que tengan más relaciones que la media con conceptos relacionados. De esta manera se busca identificar a aquellos conceptos que son muy utilizados en el campo de conocimiento y cuya evaluación debe ser precisa y valiosa.

Conectores: Se trata de conceptos que, si bien pueden tener pocas relaciones, generalmente aparecen como nexos entre grupos de conceptos importantes. Son de particular importancia para analizar las relaciones que existen entre conjuntos de temas o unidades temáticas.

Estructuras conceptuales: Se considera que una estructura conceptual es la mínima unidad de conocimiento evaluable. Está formada por dos conceptos y una relación que los une. El análisis de la cantidad de estructuras conceptuales existente en el grafo puede dar una idea de la cantidad de conocimiento "evaluable" que hay en el mismo.

Conceptos autoreferenciados: Se considera que un concepto es autoreferenciado si siguiendo una determinada

cantidad de relaciones se puede volver sobre el mismo. Dichos conceptos deben ser evitados, sobre todo si la longitud del ciclo que vuelve sobre el mismo es muy corta ya que limita la posibilidad de que los alumnos se explayen en una respuesta. Se plantea en el presente proyecto establecer un tamaño de ciclo mínimo y detectar todos aquellos conceptos que se autoreferencien por medio de una cantidad de relaciones menor a la indicada.

Agujeros negros: Se definen de esa manera a las zonas del grafo que contienen conceptos o estructuras que no forman parte de las respuestas provistas por los docentes o que aparecen en una pequeña cantidad de las mismas, indicando contenidos que no son evaluados o lo son con poca frecuencia. La detección de este patrón es de suma importancia para lograr una evaluación equilibrada de todos los contenidos de la materia.

3.2.2. Patrones asociados con el aprendizaje

Los patrones asociados al aprendizaje tienen como finalidad analizar los contenidos asimilados por los alumnos, traducidos en las respuestas provistas a las preguntas de evaluación, detectar los conceptos recurrentes, los errores más comunes, los temas con dificultades de aprendizaje, las estructuras comunes existentes en las respuestas, entre otros. Se pretende con esto detectar problemas de aprendizaje que posibiliten mejorar tanto el dictado de la materia como la construcción de los instrumentos de evaluación.

Entre los patrones que se buscan se encuentran aquellos caracterizados de la siguiente forma:

Los temas que revisten mayor dificultad de aprendizaje: Aquellas estructuras de respuesta que tienen menor puntaje en la evaluación. Se puede analizar en este caso la respuesta total o ciertos conceptos particulares que pueden ser confusos para los alumnos.

Cantidad y tipo de los errores más comunes: Relacionada muy cercanamente con el patrón anterior, busca determinar de qué manera los alumnos suelen equivocarse y sobre qué conceptos, por ejemplo, utilizando sinónimos que no son válidos, ejemplificando de manera errónea, o estableciendo relaciones inválidas entre dos conceptos válidos.

Tendencias de los alumnos con respecto a las mismas preguntas: A través del análisis de diversas respuestas a una misma pregunta base, se busca determinar si los alumnos utilizan de forma uniforme siempre los mismos conceptos o si son capaces de utilizar conceptos relacionados, sinónimos y otras estructuras, para lograr una riqueza conceptual y una variedad expresiva alta.

Comparativa con definiciones del material de estudio: Se propone establecer una respuesta base exacta, que replique exactamente las definiciones existentes en el material de estudio, para verificar si ante la pregunta docente, el alumno elabora una respuesta o simplemente transcribe la definición aprendida de memoria.

3.3. El diseño físico de la base de datos

El diseño físico de la base de datos de grafos subyacente para el análisis y detección de patrones tiene un impacto directo y decisivo en el tipo y variedad de los algoritmos que se pueden utilizar y en el rendimiento general del sistema.

Este diseño debía balancear dos elementos importantes pero que muchas veces se contraponen. Por una parte el diseño debía ser lo suficientemente genérico y flexible como para permitir representar tanto el conocimiento actualmente disponible como permitir la incorporación futura de nuevos conceptos, relaciones, conectores y construcciones conceptuales complejas.

Por otra parte las estructuras debían permitir la utilización de diversos algoritmos, establecidos y probados, de una forma eficiente, así como debía dar la posibilidad de desarrollar, probar e implementar algoritmos desarrollados para los usos específicos mencionados en el presente trabajo.

Estas dos necesidades contrapuestas obligaron a plantear algunas soluciones de compromiso y a utilizar todos los mecanismos específicos presentes en el motor de base de datos elegido (OrientDB) para mantener un rendimiento aceptable sin sacrificar flexibilidad.

Analizaremos a continuación algunas de las decisiones de diseño más relevantes:

3.3.1. Minimización de las redundancias

Desde el momento en que se dispone a analizar las respuestas provistas por alumnos a exámenes, redactadas en forma de texto libre, es esperable encontrar un gran número de repeticiones en los conceptos expresados ya que no existe una cantidad infinita de respuestas válidas a una pregunta dada.

Desde el punto de vista del análisis de los patrones subyacentes en las respuestas una gran cantidad de conceptos repetidos no ayuda en si a la detección de patrones sino que, al contrario, simplemente aumenta el espacio de búsqueda de los algoritmos, reduciendo su eficiencia, pero sin obtener una mejor calidad en los patrones detectados.

Es por eso que uno de los principales desafíos fué la minimización de los conceptos utilizados para representar las respuestas de los alumnos pero manteniendo además el valor asociado a saber cuáles conceptos se utilizan con mayor frecuencia y en última instancia cuál es esa frecuencia.

En la figura 1 vemos un ejemplo con tres respuestas posibles a la pregunta identificada como R_1:

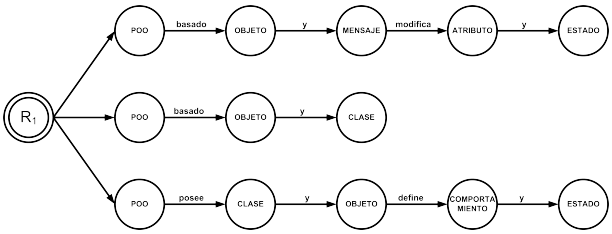

Figura 1: Respuestas con redundancia

La solución implementada hace uso del mecanismo de etiquetado existente en el motor de base de datos, para minimizar la cantidad de nodos y relaciones a la vez que es posible reconstruir todas las posibles respuestas proporcionadas por los alumnos. Esto se logra etiquetando los arcos con una etiqueta relacionada a la respuesta, en este caso A_1, A_2 y A_3. Siguiendo todos los arcos etiquetados A_1 se reconstruye la respuesta 1, siguiendo A_2 la respuesta 2 y así sucesivamente, sin tener que repetir ni los nodos ni los arcos.

La base de datos minimizada, de acuerdo al ejemplo precedente, se puede ver en la figura 2

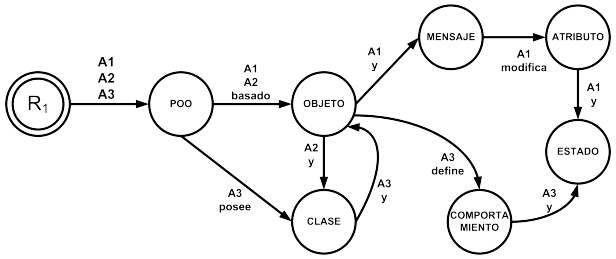

Figura 2: Respuestas minimizadas

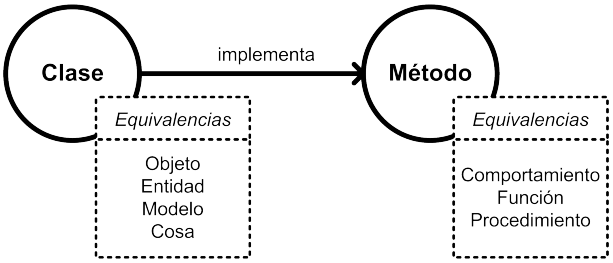

Figura 3: Diseño de nodos. Equivalencias

Este método es ampliable a *n* respuestas, sin ningún tipo de limitaciones.

3.3.2. Diseño físico de los nodos conceptuales

Tal como se mencionara anteriormente uno de los criterios guía para el diseño de una base de conocimiento efectiva es el de *Adecuación Representacional*. Este criterio busca que la representación del conocimiento sea útil para la inferencia, que es lo que en última instancia es el objetivo fundamental de cualquier base de conocimiento.

El desafío, en este sentido, radicaba en cómo almacenar conocimiento compacto pero que a la vez sea representativo y amplio en su expresividad. Para ello se modeló el almacenamiento de forma tal que incluya dos tipos de información en el mismo nodo físico. Por un lado se registra el *Concepto Puro*, como el concepto más general y abarcativo de todos los posibles que transmiten un significado coherente dentro del dominio.

La expresividad y amplitud a la hora de representar el conocimiento se logra por medio de la implementación de listas de *Conceptos equivalentes*, los cuales se almacenan de forma indivisible del Concepto Puro asociado.

OrientDB incluye la posibilidad de utilizar tipos de datos complejos como parte de los atributos de los nodos, y en este caso se utiliza un tipo Dictionary (una tabla clave/valor) para almacenar las equivalencias. En la clave se almacena el texto correspondiente a la equivalencia, mientras que en el campo valor se almacena el peso relativo de la equivalencia con respecto al Concepto Puro.

Esto permite mantener una gran expresividad en las respuestas permitidas a los alumnos sin aumentar exponencialmente la cantidad de nodos que se requieren en la base de conocimientos. De esta manera, suponiendo que se deba almacenar el concepto asociado a *Clase*, el mismo permite utilizar *Clase*, *Objeto*, *Entidad*, *Modelo* o *Cosa*, de forma indistinta, todos ellos relacionados al mismo Concepto Puro, sin necesidad de crear un nodo distinto para cada uno de ellos. Tal como se ve en el ejemplo de la figura 3, este esquema permite una expresividad importante ya que con solamente 2 nodos y sus correspondientes equivalencias se pueden llegar a 20 expresiones relacionadas (5 en el primer nodo y 4 en el segundo, con todas sus posibles combinaciones).

Los nodos conceptuales además cuentan con una lista de etiquetas que indican en qué respuestas de los alumnos se ha mencionado el concepto, tal como se vió en el ejemplo de la figura 2. Esta implementación tiene dos ventajas fundamentales:

1. Consultando todos los nodos para una etiqueta dada se puede reconstruir cualquiera de las respuestas provistas por los alumnos.
2. Para el cálculo de metricas, simplemente contando la cantidad de etiquetas por cada nodo conceptual se tiene una medida rápida y eficiente de la utilización de ese concepto en las respuestas (tal como se vé en el ejemplo de la figura 4)

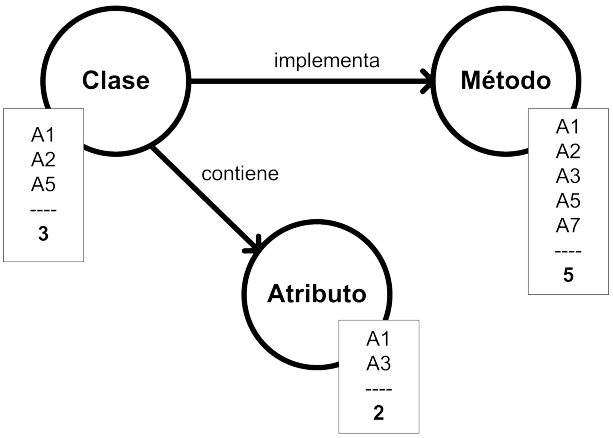

Figura 4: Etiquetas. Uso de conceptos

3.3.3. Diseño físico de las relaciones entre nodos

El modelo de almacenamiento de la base de datos utilizada para el presente proyecto cuenta con la posibilidad de etiquetar relaciones entre nodos, una característica que no es omnipresente entre los sistemas de bases de datos de grafos, pero que en el marco del presente proyecto de investigación ha demostrado ser de gran utilidad.

Esta característica permite utilizar el mismo enfoque utilizado para los nodos y hacerlo extensivo a las relaciones. De esta manera tanto nodos como relaciones se pueden etiquetar de acuerdo al uso de un concepto o relación en una respuesta dada, lo que permite reutilizar las relaciones y no tener que crear gran cantidad de relaciones idénticas, cada una de ellas para una respuesta distintas.

OrientDB también cuenta con una característica útil a la hora de modelar relaciones entre conceptos producto de respuestas textuales, y esta es la unidireccionalidad de las relaciones. OrientDB fuerza el esquema de Grafo Dirigido,

en el cual todas las relaciones entre dos nodos tienen una dirección, dada por un cierto nodo de origen y uno de destino.

Este esquema es consistente con las reglas de la lengua castellana donde las construcciones tienen un cierto orden y la inversión de ese orden no necesariamente es válido y respeta el sentido de la construcción original. El esquema además es flexible ya que ante los casos en que esa *inversión* del orden de los conceptos es válida y mantiene el sentido, es posible establecer dos relaciones con órdenes opuestos, para reflejar esa condición.

4. Conclusiones y trabajos futuros

El reconocimiento de patrones es un ámbito de gran auge en la actualidad, si bien la mayoría de las aplicaciones se orientan al reconocimiento de patrones sobre imágenes u otro tipo de información bidimensional, el reconocimiento sobre grafos está cobrando cada vez mayor importancia debido en gran medida al auge de las redes sociales, naturalmente modeladas como grafos, principalmente en la búsqueda de información valiosa que ya existe en la estructura de datos y que se pueda aprovechar.

En el caso particular de este proyecto de investigación se busca detectar patrones subyacentes tanto en la información del programa de la materia como de los exámenes elaborados por los docentes o en los conocimientos aprehendidos por los estudiantes, de forma tal que se pueda mejorar el dictado de la materia y mejorar el aprendizaje, objetivo este central en cualquier proceso educativo.

Se estima que la infraestrutura sugerida para almacenar la base de datos y los patrones previamente mencionados brindarán a la cátedra de Paradigmas de Programación un sustento fáctico sobre el cual basar las posibles modificaciones a los materiales de estudio a los métodos de dictado y a los instrumentos de evaluación, así como una herramienta tecnológica eficaz para evaluar, en un tiempo relativamente breve, el impacto de esas modificaciones sobre el cursado y la evaluación.

Se proseguirá la investigación de forma tal que los docentes cuenten con herramientas informáticas para realizar consultas sobre la base de datos de grafos, que ayuden a detectar patrones diferentes a los actualmente planteados. El objetivo final es que los propios docentes puedan indicarle al sistema, a modo de aprendizaje supervisado, qué se debe buscar de acuerdo a los objetivos de la cátedra y tener una realimentación inmediata que posibilite el análisis de diversos escenarios a la hora de plantear herramientas de evaluación y planificar el dictado.

Referencias

[1] John F. Sowa. "Conceptual Structures". En: editado por Timothy E. Nagle y col. Upper Saddle River, NJ, USA: Ellis Horwood, 1992. Capítulo Conceptual Graphs Summary, páginas 3-66. ISBN: 0-13-175878-0. URL: http://dl.acm.org/citation.cfm?id=168857.168864.

[2] W Hartshorne. *Burks, editors. Collected Papers of Charles Sanders Peirce, Cambridge, Massachusetts, 1931–1935*.

[3] Frank Van Harmelen, Vladimir Lifschitz y Bruce Porter. *Handbook of knowledge representation*. Volumen 1. Elsevier, 2008.

[4] María Alejandra Paz Menvielle y col. "Arquitectura y operatoria de un sistema de corrección de exámenes automatizado, utilizando grafos dirigidos". En: *IV Congreso Nacional de Ingeniería Informática y Sistemas de Información, CONAIISI, Universidad Católica de Salta, Facultad de Ingeniería, Argentina (2016)*. 2016.

[5] María Alejandra Paz Menvielle y col. "Teoría y práctica de un sistema de corrección automatizada, utilizando grafos dirigidos como base de conocimiento". En: *V Congreso Nacional de Ingeniería Informática y Sistemas de Información, CONAIISI, Universidad Tecnológica Nacional Facultad Regional Santa Fe, Argentina*. 2017.

[6] Satosi Watanabe. *Pattern recognition: human and mechanical*. John Wiley & Sons, Inc., 1985.

[7] John Scott y Peter J Carrington. *The SAGE handbook of social network analysis*. SAGE publications, 2011.

[8] Charles Sanders Peirce. *The essential Peirce: selected philosophical writings*. Volumen 2. Indiana University Press, 1992.

[9] Mile Pavlić, Ana Meštrović y Alen Jakupović. "Graph-based formalisms for knowledge representation". En: *Proceedings of the 17th world multi-conference on systemics cybernetics and informatics (WMSCI 2013)*. Volumen 2. 2013, páginas 200-204.

[10] Wenfei Fan. "Graph pattern matching revised for social network analysis". En: *Proceedings of the 15th International Conference on Database Theory*. ACM. 2012, páginas 8-21.

[11] Maarten Van Steen. "Graph theory and complex networks". En: *An introduction* 144 (2010).

[12] Batagelj V., Bock H y Ferligoj A. *Data Science and Classification*. Springer, 2006.

[13] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[14] Pierre A Devijver y Josef Kittler. *Pattern recognition theory and applications*. Volumen 30. Springer Science & Business Media, 2012.

[15] Anke Meyer-Baese y Anke Meyer-Baese. *Pattern recognition for medical imaging*. Academic Press, 2004.

[16] Hae Yong Kim, Javier Giacomantone y Zang Hee Cho. "Robust anisotropic diffusion to produce enhanced statistical parametric map from noisy fMRI". En: *Computer Vision and Image Understanding* 99.3 (2005), páginas 435-452.

[17] Hae Yong Kim y Javier Oscar Giacomantone. "A new technique to obtain clear statistical parametric map by applying anisotropic diffusion to fMRI". En: *Image Processing, 2005. ICIP 2005. IEEE International Conference on*. Volumen 3. IEEE. 2005, páginas III-724.

[18] Luis Alonso Romero y Teodoro Calonge Cano. "Redes neuronales y reconocimiento de patrones". En: (2001).