

# DetECCIÓN DE CONCEPTOS Y RELACIONES PARA EVALUACIÓN DE RESPUESTAS

María Alejandra Paz Menvielle, Mario Alberto Groppo, Marcelo Martín Marciszack, Analía Guzmán, Karina Ligorría, Martín Cassatti, Seiyu Ricardo Higa Tamashiro, Juan Pablo Gimenez

*Departamento de Ingeniería en Sistemas de Información  
Facultad Regional Córdoba – Universidad Tecnológica Nacional  
Maestro Marcelo López esq. Cruz Roja Argentina – Córdoba  
0351 – 4686385*

*pazmalejandra@gmail.com, proyale@groppo.com.ar, marciszack@gmail.com, ,  
aguzman@sistemas.frc.utn.edu.ar, karinaligorría@hotmail.com, mcasatti@gmail.com,  
rickysst@gmail.com, gimenezjuan92@gmail.com*

## Resumen

*El presente documento presenta un método para analizar las respuestas escritas por alumnos en forma de texto redactado en lenguaje natural, a preguntas de un examen, con el fin de contrastar su grado de coincidencia con alternativas de respuestas suministrada por un docente.*

*Se muestran las técnicas que permitirán asignar valores a los conceptos y relaciones a fin de poder ponderar la respuesta suministrada por el alumno y compararla con la ponderación de la respuesta base elaborada por un docente.*

*Estas técnicas trabajarán con todos los casos, incluyendo los distintos grados de acierto que pueda tener la respuesta del alumno, exponiendo los mecanismos con los que se deben analizar los conceptos y las relaciones para obtener la ponderación de la respuesta provista.*

**Palabras clave:** *análisis de textos; grafos; detección de patrones; reconocimiento; detección de rutas*

## 1. Contexto

El presente trabajo forma parte del proyecto de investigación y desarrollo homologado por la Secretaría de Investigación, Desarrollo y Posgrado de la Universidad Tecnológica Nacional, reconocido con el código PIDEIUTNCO0003592 en el ámbito de la Universidad Tecnológica

Nacional, por un período de dos años y a partir del 1 de enero de 2015.

Debe dar cumplimiento en simultáneo a dos grandes requisitos. Por un lado los contenidos mínimos fijados para la asignatura Paradigmas de Programación, tal cual figuran en la ordenanza 1150 de la carrera, los cuales pertenecen al bloque de tecnologías básicas dentro del área programación, que están principalmente referidos a los paradigmas lógicos, funcional y de orientación a objetos. Y por otro lado cumplir con los descriptores y criterios de intensidad de formación práctica de la Resolución Ministerial 786/09, los que se encuentran definidos en el área de tecnologías básicas, sub-área programación que incluyen a los paradigmas y lenguajes de programación.

## 2. Introducción

El procesamiento de patrones es un ámbito de crucial importancia en todos los tópicos referidos a automatización de procesos. Tanto desde el punto de vista de detección de patrones desconocidos como desde el reconocimiento de patrones previamente detectados, los algoritmos asociados a ambos procesos han ido refinándose y ganando en complejidad y exactitud.

## 3. Estudios sobre representación del conocimiento. Grafos conceptuales

Un grafo conceptual [3] es un sistema de notación simbólica y de representación del conocimiento. Presentado por John F. Sowa, se basa en los gráficos existenciales [4] de Charles Sanders Pierce, en las estructuras de redes semánticas y en datos de la lingüística, la filosofía y la psicología.

El objetivo del presente proyecto es desarrollar un sistema que represente de manera adecuada y simple, las estructuras del lenguaje natural con el fin de determinar si la respuesta a una pregunta es correcta, dentro de un dominio especificado y aplicando las técnicas y reglas gramaticales relacionada con los lenguajes estructurados [8].

#### 4. Redes semánticas

Los estudios sobre redes semánticas sostienen que es posible representar un conocimiento en un dominio determinado mediante un grafo que modele, por un lado, los conceptos y por el otro las relaciones que conectan dichos conceptos. Se puede construir, de esta manera, un grafo que cumpla con la siguiente ecuación:

$$G = (V, A)$$

En donde  $V$  es el conjunto de todos los términos pertenecientes al dominio en cuestión, constituyendo los vértices que se denominarán  $t_1, t_2, \dots, t_n$ , y  $A$  es el conjunto de todas las relaciones válidas entre dos términos cualquiera.

Dados entonces dos términos,  $t_1$  y  $t_2$  existirá una línea  $a_{ij}$  que une dichos términos, si y sólo si, los términos están relacionados.

Si las relaciones no son bidireccionales se utilizan grafos dirigidos para representar el sentido de dicha relación. [2]

En este contexto y en base a lo investigado hasta el momento, se induce que el uso de grafos conceptuales podría ser una herramienta adecuada para el análisis de las expresiones y la detección de patrones de texto escritos en lenguaje natural.

En este contexto y en base a lo investigado hasta el momento, se induce que el uso de redes semánticas podría ser una herramienta adecuada para el análisis de las expresiones y la detección de patrones de texto escritos en lenguaje natural, por un lado por la mayor simplicidad que las mismas presentan como modelo de representación del conocimiento y por otra parte debido a la mayor facilidad de implementación de

mecanismos automatizados de evaluación que utilicen este modelo.

#### 5. Estudios sobre detección de patrones

Existen dos tipos de enfoque relacionados a la búsqueda de patrones dentro de información almacenada en forma de grafos.

El primer enfoque, es la detección de subgrafos [5], donde se buscan ciertas estructuras, contenidas en un grafo de mayor tamaño. Generalmente el resultado de dichas búsquedas es binario y simplemente se determina si la estructura se encuentra, o no, en el grafo destino.

El segundo enfoque es el de búsquedas inexactas, en donde se debe especificar al algoritmo, además de la estructura que se desea encontrar, un cierto umbral que debe ser tenido en cuenta si la estructura exacta no se encuentra. Teniendo en cuenta este umbral, el algoritmo busca los elementos de la estructura especificados con cierto grado de semejanza, por ejemplo con conceptos faltantes, con relaciones similares pero no idénticas, o con órdenes alterados o invertidos en las relaciones [6].

Debido a la naturaleza inherentemente variable de las respuestas registradas en exámenes al mismo conjunto de preguntas, tomaremos el enfoque de búsqueda inexacta, para la determinación de la validez, total o parcial, de las respuestas obtenidas.

#### 6. Desarrollo e Innovación

Para el presente trabajo se definió como dominio de aplicación a los contenidos curriculares de la materia Paradigmas de Programación, dictada en el segundo año de la carrera de Ingeniería en Sistemas de Información en la Universidad Tecnológica Nacional, Facultad Regional Córdoba.

Los conceptos impartidos en la materia se representaron a través de una estructura de grafos, en la forma de un grafo dirigido, con atributos que amplían el contenido semántico de las entidades involucradas.

#### 7. Modelado del conocimiento

Se partió del estudio de la modalidad académica de la materia para determinar los ejes

fundamentales y las unidades temáticas contenidas en la misma, así se definió el primer conjunto de nodos, de nivel cero, que comprendían los conceptos desde los cuales se derivarían el resto de los datos.

Una vez realizado esto, el resto de los nodos se fue encadenando de forma natural como parte de la evolución normal del dictado de la materia. En otras palabras, el grafo de conocimiento creció, en su primera etapa, en el mismo orden y progresión en el que la materia se dictaba como se puede ver en el ejemplo de la Figura 1.

Un camino similar se siguió con las relaciones que surgieron entre dichos nodos, las mismas se fueron definiendo a medida que las asociaciones entre conceptos se hacían necesarias.

La teoría de grafos en la que se basa este modelo [2] [3], establece que cada nodo contiene un conjunto de etiquetas diseñado para representar correctamente el concepto asociado y, a la vez, facilitar la posterior búsqueda.

Figura 1: Ejemplo de evolución de encadenamiento de nodos en momentos t1, t2, t3

Las etiquetas mínimas de los nodos son (véase Figura 2):

- Nombre: Nombre del concepto asociado, en la imagen ejemplo sería Paradigma, OOP, Lógico.
- Usuario: Usuario que agregó el concepto al grafo.
- Actualizado: Fecha y hora de la última modificación.

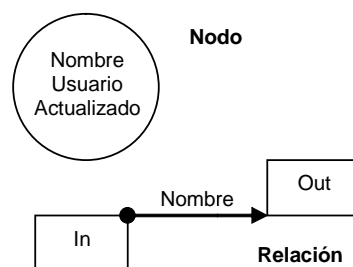


Figura 2: Nodos y relaciones

Se definieron distintos tipos de relaciones, para especificar cuál es la manera en la que dos conceptos pueden estar asociados.

El atributo fundamental de las relaciones es el Nombre, pero cada una tiene tres atributos definidos:

- Nombre: Indica el nombre de la relación según su definición académica
- In: Lista con todos los nodos de origen de la relación
- Out: Lista con todos los nodos de destino de la relación

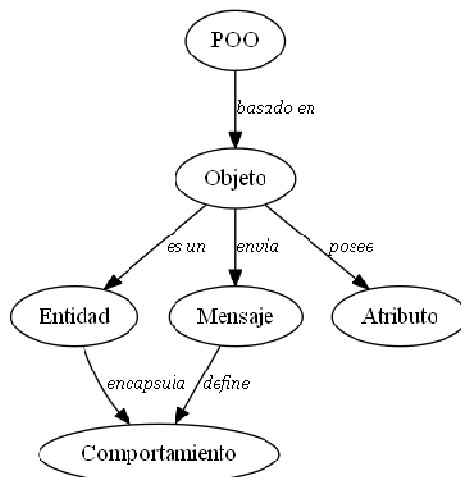
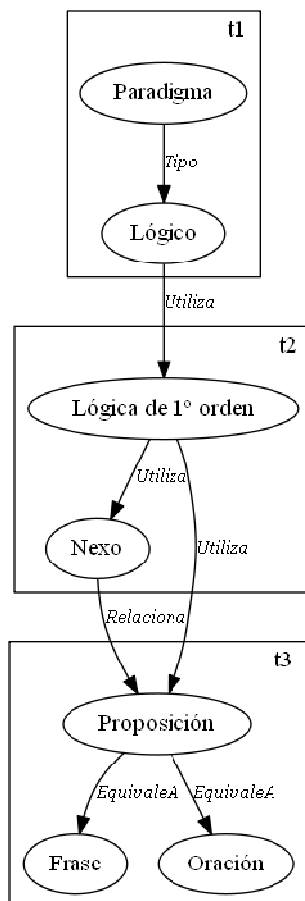


Figura 3: Tipos de relaciones

Tan valiosas como los conceptos son las relaciones que los unen y es por esto que se ha definido un mecanismo mediante el cual se pueden crear nuevos tipos de relación a medida que sea necesario. Hay que mencionar que cuando las relaciones pueden ser reutilizadas se promueve esta práctica ya que minimiza el tamaño de la base de conocimientos y hace que los contenidos de la misma sean más consistentes. En la figura 3 se pueden apreciar algunos de los tipos de relación que actualmente intervienen, en este caso, sobre conceptos relacionados con el Paradigma Orientado a Objetos (POO).

Las primeras pruebas se realizaron ingresando manualmente los datos en la base de datos. Actualmente se está desarrollando una herramienta GUI (Graphic User Interface) que se implementará en Java y hará posible simplificar el proceso de carga de datos; reducir los errores de tipo y tiempos de procesamiento.

## 8. Mecanismo de búsqueda

### 8.1. Corrección ortográfica y gramatical

El primer paso antes de iniciar el procedimiento de búsqueda propiamente dicho, es una revisión ortográfica y sintáctica del texto para evitar la sobrecarga innecesaria al buscar términos con errores o relaciones inválidas. Para ello se realiza una revisión por palabras, contra un diccionario de idioma español y se verifican las reglas gramaticales básicas, antes de suministrar el texto al motor de búsqueda.

### 8.2. Simplificación del texto

El texto, previamente validado en su ortografía y gramática, se divide en tokens o unidades atómicas, que se enviarán al motor de búsqueda para que el mismo determine si son conceptos pre-existentes en el grafo, o relaciones previamente cargadas. En todos los casos se trabajará con las raíces de las palabras y los verbos en infinitivo.

»	A	B	C	D	E	F	Vector inicial
«	A	<del>B</del>	C	<del>D</del>	<del>E</del>	F	B, D, E (no existen)
»	A	B[+]	C	H		F	Agregar B. Reemplazar D por H. Descartar E
«	A	B	C	H		F	B Agregado. H existe.
»	A	B	C	H		F	Vector final de búsqueda

Figura 4: Ajuste del vector de búsqueda

El motor de búsqueda devuelve la misma lista de tokens con marcadores que indican si cada uno de los términos existe, no existe o es una equivalencia de un concepto existente (véase Figura 4). El llamador, en este caso, decidirá qué acción tomar con respecto a los términos inexistentes. Puede descartarlos de la consulta, reformularlos o solicitar que se agreguen como conceptos nuevos en la próxima pasada. De esta forma, al ir realizando consultas sobre el grafo también se pueden ir realizando los pasos necesarios para que la base de conocimientos se mantenga actualizada.

### 8.3. Búsqueda de rutas

Una vez que el intercambio de información entre el cliente y el motor de búsqueda para simplificación del texto ha terminado, esto es, cuando el motor no tiene más sugerencias al respecto de los términos solicitados, y el cliente no desea realizar más cambios al conjunto enviado, se procederá a la búsqueda de las rutas válidas que interconectan los conceptos.

Para ellos se utilizarán mecanismos iterativos y recursivos para obtener las listas de relaciones que unen los conceptos. Hay que destacar que no es necesaria una búsqueda de nodos, porque ya se ha determinado que todos los términos que componen el texto a analizar se encuentran como nodos del grafo.

El objetivo de este paso no es obtener una única ruta, ni la mejor ruta, sino el conjunto de todas las rutas que contienen los conceptos indicados, sin discriminar los tipos de relaciones que los unen.

## 8.4. Valoración y ponderación de conceptos y relaciones

Considerando que la búsqueda de relaciones entre conceptos deberá ser utilizada para poder evaluar el grado de validez de una respuesta se planteó la necesidad de comparar las posibles respuestas con una respuesta base, definida como correcta, que debe ser suministrada por el sistema.

De esta forma, todas las posibles rutas serán comparadas con esta respuesta base y se computará su grado de validez en base a esa comparación.

También es importante tener en cuenta la expresividad de las respuestas dadas en forma de texto libre, en el que es posible utilizar sinónimos, equivalencias y la capacidad descriptiva del lenguaje escrito para brindar respuestas que son gramaticalmente distintas a la respuesta base pero conceptualmente equivalentes, y deberán ser evaluadas en consecuencia.

Los pasos necesarios para realizar la valoración de respuestas son los siguientes:

Para cada ruta encontrada, la llamaremos ruta candidata, se aplicará un algoritmo cuyo objetivo es contar la cantidad de conceptos y relaciones exactas que la componen. Se considerarán conceptos y relaciones exactas a aquellas que se hallan con el mismo nombre y en la misma posición que en la respuesta base.

Si la respuesta candidata tiene los mismos conceptos y relaciones, en las mismas ubicaciones y con los mismo nombres que en la respuesta base, se la denomina “respuesta perfecta” y tendrá un valor que se calcula como:

$$V_r = C + R$$

Siendo:

$V_r$  : Valor de la respuesta

$C$  : Cantidad de conceptos

$R$  : Cantidad de relaciones

Cualquier otro caso implicará que hay conceptos o relaciones inexactas (con valores ponderados  $< 1$ ) y se requerirá un análisis diferente para obtener la mejor respuesta.

A continuación se describen los distintos casos que se pueden presentar al momento de hacer las valoraciones de conceptos y de relaciones.

## 8.5. Valoración de los conceptos

**Caso 1:** Los conceptos son coincidentes pero las ubicaciones dentro de la respuesta difieren de la respuesta base.

En un grafo dirigido, por ejemplo el de la ruta de respuesta candidata, se considera que la distancia entre dos nodos es la cantidad de arcos que se deben atravesar para llegar desde el nodo de origen hasta el nodo de destino [7].

En el presente trabajo utilizaremos el concepto de *desplazamiento*, que definiremos mediante dos componentes. El primero es un valor numérico que expresa la diferencia entre las posiciones del concepto en la respuesta base y su ubicación en la respuesta candidata. El segundo componente es un signo, que indica la dirección del desplazamiento. Por convención tomaremos el signo positivo (+) para indicar conceptos que están desplazados hacia adelante con respecto a la respuesta inicial y un signo negativo (-) para indicar aquellos conceptos que aparecen antes en la respuesta candidata que en la base.

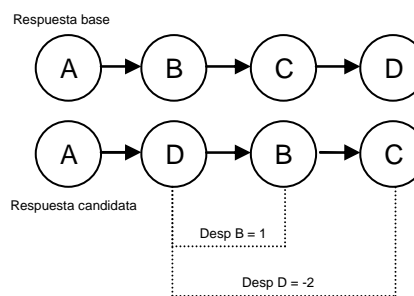


Figura 5: Desplazamiento entre conceptos

En este caso se calculará para cada concepto inexacto el desplazamiento con respecto a la respuesta base. Se considera que mientras más alejado esté el concepto de su ubicación correcta, su incidencia en la respuesta disminuye. Es por

ello que el valor del concepto se ponderará de la siguiente forma:

$$C_d = 1 - \left| \frac{P_c - P_b}{n} \right|$$

Siendo:

$C_d$  : Valor del concepto desplazado

$P_c$  : Posición del concepto candidato

$P_b$  : Posición del concepto en la respuesta base

$n$  : Cantidad de conceptos en la respuesta

**Caso 2:** Los conceptos no son exactos en el nombre, sino que son equivalentes a los conceptos de la respuesta base.

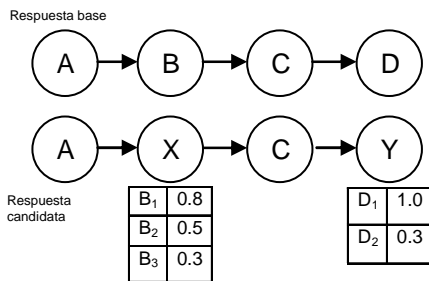


Figura 6. Peso de conceptos no-exactos

En este caso, cada concepto tendrá una tabla de equivalencias, definida por el docente, cuyo fin será ponderar cada una de las equivalencias a un valor numérico entre 0 y 1, donde 1 será una equivalencia completa, que indicará que ambos términos pueden usarse indistintamente. De esta forma el valor de  $C_e$  es el valor que se obtiene de las tablas de equivalencia, para el concepto dado tal como se observa en la figura 6.

Los casos 1 y 2 descriptos se podrán combinar cuando existan conceptos que no son exactos en el nombre, y además se encuentren en ubicaciones diferentes a las definidas en la respuesta base, ponderando el concepto candidato según:

$$C_{de} = C_d \cdot C_e$$

Siendo:

$C_{de}$  : Valor del concepto desplazado equivalente

$C_d$  : Valor del concepto desplazado

$C_e$  : Valor del concepto equivalente

## 8.6. Valoración de las relaciones

Las relaciones deben ser valoradas de una forma similar a los conceptos para obtener la valoración total de la respuesta. Se han detectado distintas situaciones o casos que serán descriptos a continuación.

**Caso 1:** La relación es exacta y los nodos de origen y destino son exactos. En este caso se considerará una relación exacta pura y su valor será igual a 1.

**Caso 2:** La relación es exacta pero un nodo o ambos, nodo origen y nodo destino, no son exactos. En este caso el peso de la relación se reducirá, debido a que no hay forma de asegurar que la relación original siga siendo válida al haber cambiado la exactitud de los nodos asociados. En este caso el valor de la relación se calculará como:

$$V_r = \frac{(1 + C_o + C_d)}{3}$$

Siendo:

$V_r$  : Valor de la relación

$C_o$  : Valor del concepto de origen

$C_d$  : Valor del concepto de destino

Como se puede observar a través de la fórmula, el caso de la relación exacta pura sería un caso particular en el que el valor del concepto origen y del concepto de destino son ambos igual a 1.

**Caso 3:** La relación no es exacta pero los nodos de origen y destino si lo son. En el caso en que la relación no es exacta se deberá analizar la lista de equivalencias suministrada por el sistema. Esta lista contendrá los términos equivalentes con su ponderación asociada para una relación dada en la respuesta base. Si la relación en la respuesta candidata se encuentra dentro de la tabla de equivalencias, el valor  $R_e$  de dicha relación, se obtendrá directamente de la tabla.

**Caso 4:** La relación no es exacta y los nodos de origen y/o destino tampoco lo son. En este caso se deberá utilizar también la tabla de equivalencias pero los pesos ponderados de los nodos de origen

y destino podrán ser alterados de acuerdo a sus propias equivalencias.

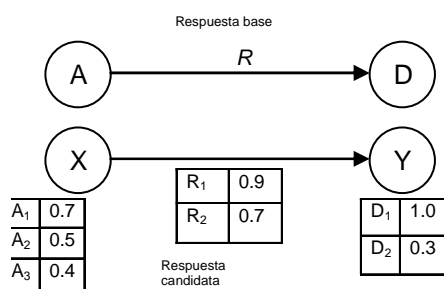


Figura 7. Relaciones y conceptos no-exactos

Aquí el valor de la relación se definirá como la media matemática entre los valores de los nodos de origen y de destino y el valor ponderado de la relación equivalente.

$$V_r = \frac{(R_e + C_o + C_d)}{3}$$

Siendo

$V_r$ : Valor de la relación

$R_e$ : Valor de la relación equivalente

$C_o$ : Valor del concepto de origen

$C_d$ : Valor del concepto de destino

En el ejemplo de la figura 7, la relación

$$A_1 \rightarrow R_1 \rightarrow D_2$$

Tendría un valor ponderado de:

$$V = \frac{(A_1 + R_1 + D_2)}{3}$$

$$V = \frac{(0.7 + 0.9 + 0.3)}{3} = 0.633$$

## 8.7. Evaluación

Una vez realizado el cálculo del valor de la respuesta, el valor obtenido variará entre 0 y el valor de la respuesta base. Por ejemplo:

$$V = V_{base}$$

Implicaría una respuesta correcta “perfecta” con un grado de coincidencia conceptual del 100%.

Los valores intermedios indicarán el grado de aproximación de la respuesta suministrada a la respuesta base y podrán ser útiles como indicadores para calcular el porcentaje de acierto de la evaluación en que la respuesta analizada estará incluida. En tal sentido será posible fijar dos umbrales ( $V_{min}$  y  $V_{max}$ ) a partir de los cuales se considerará a las respuestas dentro de los siguientes parámetros:

$$V \geq V_{max} \rightarrow \text{correcta}$$

$$V \leq V_{min} \rightarrow \text{incorrecta}$$

$$V_{min} < V < V_{max} \rightarrow \text{evaluar}$$

Si el valor obtenido no se encontrara dentro de los extremos establecidos, la respuesta debería ser evaluada por un docente a fin de colocar a la misma una calificación correcta en base a una interpretación o un contexto que exceden aquellos representados por la base de conocimientos disponible, y por ende, al mecanismo de evaluación automatizado expuesto.

## 9. Conclusiones y trabajos futuros

Se ha presentado un método alternativo que permitirá calificar exámenes formados por preguntas que se responderán como ensayo escrito en forma de texto libre, por alumnos de nivel universitario.

La metodología descrita se apoya en la teoría de grafos para construir el dominio de trabajo, que se corresponde con el área de conocimiento de la cátedra Paradigmas de Programación dictada en la Universidad Tecnológica Nacional Facultad Regional Córdoba, y en el estudio de las gramáticas y lenguajes formales [1].

Se continuará con el análisis de textos desde esta perspectiva ampliando el aprendizaje sobre la materia y otras técnicas y metodologías existentes, pudiendo de esa manera acercar, a la comunidad académica, un poco más a la solución de la problemática permanente que significa que un computador pueda analizar textos correctamente.

Se prevé la realización de pruebas de concepto con casos reales de evaluación y la participación de docentes de la cátedra, para validar la premisa básica del trabajo con respecto a la calidad de la evaluación realizada automáticamente por el sistema así como la ponderación realizada por el mismo, sobre los conceptos, para la obtención de la nota.

También, dentro de dicho estudio, se incluirá una validación del modelo representado por el grafo para obtener una retroalimentación de los docentes con respecto a la representatividad que la herramienta propuesta brindaría para su trabajo diario.

Está previsto continuar con el desarrollo del sistema que aplique en forma on-line esta metodología, para que el aplicativo expanda las posibilidades de potenciales usuarios, pudiendo generar el intercambio de ideas con personas relacionadas al tema, desde otras cátedras.

## 10. Referencias

[1] J Glenn Brookshear, *Teoría de la Computación*, Addison Wesley, 1989.

[2] Sowa, John F, "Semantic networks" en *Encyclopedia of Cognitive Science*, 2006.

[3] Sowa, John F. "Conceptual graph summary", en *Conceptual Structures: Current Research and Practice*. Ellis Horwood, New York London Toronto, 1992, pp. 3-66.

[4] Peirce, Charles S. "Existential graphs", en *Collected Papers of Charles Sanders Peirce* 4, 1909, pp. 1-7.

[5] Pavlidis, Theodosios, *Structural pattern recognition*. Vol. 2. New York: Springer-verlag, 1977.

[6] Olmos, Ivan, Jesus A. Gonzalez, and Mauricio Osorio. "Inexact Graph Matching: A Case of Study", en *FLAIRS Conference*, 2006.

[7] Buckley, Fred, and Frank Harary. *Distance in graphs*. Addison-Wesley Longman, 1990.

[8] Hopcroft, Motwani, Ullman. *Teoría de autómatas, lenguajes y computación*. Pearson, Addison-Wesley, 2008.