Procesamiento Semántico Difuso Aplicado a un Modelo de Análisis de Textos basado en Grafos

Wenny Hojas-Mazo, Alfredo Simón-Cuevas
Universidad Tecnológica de La Habana "José Antonio Echeverría", Cujae
Ave. 114, No. 11901, CP: 19390, La Habana, Cuba
{whojas, asimon}@ceis.cujae.edu.cu

José A. Olivas, Francisco P. Romero
Universidad de Castilla La Mancha
Paseo de la Universidad, 4, Ciudad Real, España
{JoseAngel.Olivas, FranciscoP.Romero}@uclm.es

***Resumen*— La obtención de información relevante y conocimiento en textos aún constituye un gran desafío. En este trabajo, se presenta un enfoque de análisis de textos basado en grafos, soportado en el procesamiento semántico difuso del contenido. En esta propuesta se combinan procesos de consultas sobre los grafos obtenidos de los textos, la recomendación de conceptos relevantes y la recuperación de pasajes de texto, incluyendo mecanismos de integración semántica de los grafos como parte de las consultas. La propuesta fue evaluada en el análisis de artículos científicos con resultados prometedores.**

Palabras clave — análisis de textos; representación de textos en grafos; análisis semántico difuso; integración semántica de grafos

I. INTRODUCCIÓN

El gran volumen de información no estructurada actualmente disponible en textos constituye un recurso muy valioso. El procesamiento y análisis efectivo de esas fuentes textuales, para obtener información relevante y conocimiento, aún es una tarea desafiante, demandándose mayor atención a las soluciones de minería de texto [1]. Los textos han sido procesados usando diferentes tipos de representación, donde las bolsas de palabras y los modelos vectoriales son los más usados, fundamentalmente en la recuperación de información. Sin embargo, estas soluciones se enfocan más en facilitar el acceso a los textos, y no en el análisis de su contenido para descubrir patrones y conocimiento, siendo este uno de los objetivos principales del análisis de textos. En este sentido, el uso grafos surge como alternativa prometedora para el análisis y exploración de las estructuras conceptuales de los textos [13].

En este trabajo, se presenta un modelo de análisis de textos basado en grafos, en el que se propone un enfoque difuso para el tratamiento semántico del contenido. En este modelo, el contenido conceptual de los textos es representado en grafos, y se combina el uso de consultas sobre los grafos, la identificación de conceptos relevantes y la recuperación de pasaje, para obtener estructuras conceptuales interesantes e información textual útil asociada. Las consultas están concebidas sobre la base de la integración semántica de los grafos, para lo cual se propone un enfoque difuso mediante la aplicación de la agregación compensatoria [22] de medidas de similitud sintáctica y semántica. Este enfoque ofrece mayor diversidad, en cuanto al análisis del contenido textual, respecto a lo reportado en [5][21][12]. La aplicabilidad del modelo propuesto se evaluó en las tareas de extracción y síntesis de datos en la revisión sistemática de la literatura (SRL: *Systematic Literature Review*) reportada en [8], para soportar estás tareas al recuperar información relevante que apoye a los revisores en la extracción de datos de los artículos y la búsqueda de respuestas a las preguntas de investigación planteadas. El desarrollo de este caso de estudio ejemplifica como el modelo propuesto ofrece soporte computacional a las tareas de extracción y síntesis de datos en una SRL. Las principales contribuciones de este trabajo se resumen en: (1) la aplicación de la lógica difusa en la integración semántica de grafos, basada en la agregación compensatoria [22] de medidas de similitud sintáctica y semántica para evaluar la similitud entre conceptos en los grafos; y (2) la combinación de técnicas de análisis de grafo, tales como: consultas e identificación de conceptos relevantes, y la recuperación de pasajes de texto, para obtener estructuras conceptuales interesantes e información textual útil asociada.

El resto del trabajo está organizado de la siguiente manera: la Sección 2 expone algunos fundamentos teóricos del problema; la Sección 3 describe el enfoque propuesto; la Sección 4 presenta los resultados del caso de estudio desarrollado; y la Sección 5 expone las conclusiones arribadas e ideas de trabajo futuro.

II. ANÁLISIS DE TEXTOS MEDIANTE GRAFOS

La representación de textos basada en grafos evita la pérdida de información contextual y semántica del contenido, y reduce la dispersión de los conceptos y la información en su procesamiento. A través de un grafo, el texto se reduce a un número relativamente pequeño de conceptos y relaciones, por lo que una colección de textos puede ser fácilmente manejada y analizada computacionalmente. En este sentido, el empleo de técnicas y herramientas de procesamiento de grafos facilita la obtención de información y conocimiento a partir de esa representación, así como realizar análisis cualitativos y cuantitativos sobre los conceptos incluidos en un texto, para identificar conceptos fuertemente relacionados, estructuras conceptuales interesantes, conceptos relevantes, entre otras. Las operaciones sobre los grafos (ej. Unión) son una de esas herramientas útiles en este sentido [21]. Estos beneficios se evidencian en: detección de tópicos [7], la recuperación de información [4][14], y en varias revisiones [11][7][21]. Sin

embargo, solo se identificaron dos soluciones orientadas al análisis del contenido de textos representados en grafos y enfocada a conceptos [5][12].

En [5] se propone un enfoque basado en la construcción automática de grafos etiquetados con n-gramas de tamaño variable a partir textos y en la aplicación de operaciones sobre los mismos para identificar y representar los tópicos principales de uno o más textos. En la construcción del grafo, el texto se segmenta en oraciones y se extraen los tokens, y estos son clasificados en: Ignorar (determinantes, pronombres y adverbios), Arco (conjunciones, preposiciones y verbos) y Nodo (secuencia de tokens no asociada a los anteriores), siendo estos dos últimos los utilizados para construir el grafo. Este método solo extrae relaciones explicitas del texto, por lo que es susceptible a que se representen elementos aislados en el grafo, lo cual impacta negativamente en los resultados del análisis a realizar a través de consultas y algoritmos de identificación de conceptos relevantes representados en grafos. En la modelación de los tópicos, inicialmente se generar un grafo por cada documento, luego se integran los grafos en uno solo aplicando la operación Unión. Posteriormente, se obtiene un ranking de relevancia de los nodos del grafo resultante y se aplica la operación de Proyección sobre los *k* nodos más relevantes, para obtener un sub-grafo que modela los tópicos. En esta propuesta, la integración de los grafos que se produce en la operación de Unión se basa en la equivalencia sintáctica de los nodos, sin considerar la semántica subyacente. Esto representa una debilidad, ya que se puede producir la integración de contenidos sin vínculos semánticos entre ellos, dado que los conceptos representados en los nodos pueden estar sujetos a ambigüedades. Este problema es solucionado de alguna manera en [12], donde se propone un enfoque similar de análisis a partir de grafos construidos automáticamente de los textos, pero más enfocado al análisis de conceptos y ofreciendo un mecanismo de operaciones de consulta sobre grafos más abarcador y flexible, que incluye operaciones de: Unión, Intersección y Proyección. En el caso de esta última, se ofrece mayor facilidad al usuario, ya que estos son los que seleccionan los conceptos de interés sobre los que se desea realizar el análisis. En esta solución, se incluye un método de análisis semántico para la búsqueda e integración de estructuras conceptuales de los grafos, soportado por un algoritmo de desambiguación y WordNet, el cual se aplica en el procesamiento de cada consulta. No obstante, también con la desventaja de que, como se plantea en [18], el tratamiento de la semántica basada en la desambiguación aumenta la complejidad de la solución y los recursos de cómputo requeridos, sugiriéndose el uso de técnicas de evaluación de similitud semántica, tales como las basadas en WordNet [20]. Además, estás soluciones no brinda una forma para conocer las fuentes textuales de las cuales provienen los conceptos y/o temas relevantes o de contexto, lo cual enriquecería y proporcionaría más fiabilidad al análisis de texto.

III. Modelo de Análisis Propuesto

El modelo incluye tres procesos fundamentales: representación de textos, análisis de grafos y recuperación de pasajes, e integra varios recursos de conocimiento, según se muestra en la Fig. 1. El primero, tiene el objetivo de estructurar el contenido textual mediante su representación en forma de grafo, a partir del cual se conforma un Repositorio de Grafos de Conocimiento (RGC), y usando n-gramas para la indexación de los textos. Los restantes procesos ofrecen la posibilidad al usuario de recuperar información relevante y obtener conocimiento desde esos contenidos estructurados, a través de consultas sobre el RGC, la identificación y recomendación de conceptos relevantes y la recuperación de pasajes de textos donde aparecen conceptos de interés. Este último da un valor agregado al análisis de conceptos con respecto al resto de las propuestas, pues permite identificar las fuentes textuales donde aparecen los conceptos, obteniendo un análisis más detallado.

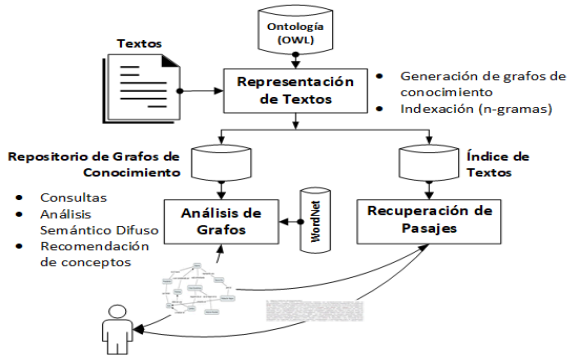

Fig. 1. Visión general del modelo propuesto

A. *Representación de los Textos*

En este proceso, el contenido de los textos es representado mediante dos esquemas: (1) grafos y (2) n-gramas, siendo el primero de ellos, la principal fuente de análisis del contenido textual. La generación del grafo está concebida de forma similar a lo reportado en [12], en tres fases: pre-procesamiento, extracción de conceptos y de relaciones. En el pre-procesamiento, se extrae la información sintáctica y gramatical del texto usando varias técnicas de procesamiento de lenguaje natural (NLP: *Natural Language Processing*) a través de FreeLing, en especial el análisis de dependencia, que aunque es más costoso computacionalmente que el análisis superficial ha dado mejores resultados [10]. Los conceptos se extraen empleando un conjunto de patrones léxico-sintácticos (representan mayoritariamente frases sustantivas) y a través del vocabulario representado en una ontología suministrada por el usuario (preferiblemente del mismo ámbito del texto). La extracción de relaciones abarca la identificación de relaciones explícitas e implícitas entre los conceptos identificados. Las explícitas se extraen de cada oración usando también patrones léxico-sintácticos y las implícitas (conectan conceptos que aparecen en diferentes partes del texto) se extraen mediante la técnica de string matching, el análisis de la proximidad entre conceptos en el texto y tenido en cuenta relaciones representadas en la ontología (si es suministrada). Este último tipo de relaciones permite incrementar la cobertura del texto, en cuanto a la representación de su contenido, aspecto clave para aumentar la eficacia en su análisis. Luego de la extracción de los conceptos y relaciones, se eliminan redundancias e inconsistencias, y se genera el grafo de cada texto. El uso de n-gramas está motivado por la necesidad de crear un índice de los textos, que permita la recuperación de contenidos textuales asociados a conceptos recuperados como resultados de las

consultas a los grafos, siendo esta una funcionalidad a través de la cual se extiende lo reportado en [12]. Los conceptos representados en los grafos pueden constituir secuencias de tokens dentro de las oraciones, por tanto, el uso de n-gramas facilita la recuperación de aquellas oraciones en las cuales estén presentes esos conceptos. La indexación se lleva a cabo segmentando los textos en oraciones, extrayendo de ellas los n-gramas (1-, 2-, y 3-gramas) y almacenándolas con esa estructura en un índice de Lucene. Ambos procesos de estructuración del contenido de los textos se ejecutan en paralelo.

B. *Análisis de Grafos*

El proceso de análisis de grafos se lleva a cabo mediante un mecanismo de consultas basado en [5][12], con la diferencia de que en esta nueva solución se propone un enfoque difuso para realizar el análisis semántico que se lleva a cabo en la integración de los grafos, como parte del procesamiento de las consultas. Adicionalmente, se incluye la identificación y recomendación de conceptos relevantes mediante la obtención de un ranking de conceptos, similar a lo reportado en [5], pero teniendo en cuenta como elementos: la frecuencia de aparición de los conceptos en los textos y la relevancia obtenida por el algoritmo PageRank [6], aplicado sobre el grafo que representa el contenido de la colección de textos.

1) Consultas

A través de estas consultas, se puede recuperar información relevante y obtener conocimiento desde un espacio de búsqueda especificado por el usuario y constituido por un conjunto de grafos almacenados en el RGC. El resultado de las consultas se puede expresar en conceptos individuales y/o estructuras conceptuales en forma de grafo. El procesamiento de las consultas incluye tareas de búsqueda e integración de información, por lo que, aunque los conceptos recuperados estén presentes en diferentes fuentes textuales, estos pueden ser integrados en el resultado final, propiciando así la generación de conocimiento. Los tipos de consultas incluidas en este proceso son: Unión, Intersección y Proyección.

La unión de dos grafos $G_1 = (N_1, E_1)$ y $G_2 = (N_2, E_2)$ se denota como $G_1 \cup G_2$ y de la misma manera $(N_1 \cup N_2, E_1 \cup E_2)$. El grafo unión se puede ver como una forma de mezclar dos grafos sin ninguna pérdida de información (es decir, sin excluir ningún nodo N_i, ni arco E_i). Por tanto, es un operador útil para mezclar información de múltiples documentos de texto [5]. En efecto, si se considera el corpus $D = \{d_1, \ldots, d_n\}$ y si G_i denota el modelo de grafo correspondiente para cada d_i, entonces la información combinada del corpus se puede representar por:

$$G_D = \bigcup_{i=1}^{n} G_i$$

Esta representación conlleva un rápido crecimiento del número de nodos y por consiguiente la obtención de una cantidad abrumadora de información. Por tanto, el uso de la unión de grafos para mezclar documentos textuales requiere típicamente post-procesamiento con un operador que extraiga la información relevante del grafo combinado [5].

La *Intersección* (*Inter*) permite recuperar conceptos y estructuras conceptuales comunes a un % (cota mínima) de los grafos que conforman el espacio de búsqueda, el cual requiere ser especificado por el usuario como parámetro, y es definido como Valor de Soporte (VS) [12]. La *Proyección* ($Proj^R$), en sus diferentes variantes, permite recuperar las estructuras conceptuales asociadas a determinados conceptos de interés especificados por el usuario, considerando diferentes niveles de vecindad (denotado por *R*); el cual también es especificado por el usuario. Se utilizan tres tipos de consultas *Proyección*, para ofrecer diferentes perspectivas de recuperación de información, considerando sobre el concepto de interés: (1) solo enlaces de entrada ($Proj^{R, IN}$); (2) solo enlaces de salida ($Proj^{R, OUT}$); y (3) todos los tipos de enlaces ($Proj^R$). Las primeras dos, son útiles para analizar los niveles de autoridad o centralidad de los conceptos de interés, respecto a otros conceptos con los que está relacionado en el texto; inspirado en los conceptos de Kleinberg [16]. Notar que esta consulta puede conducir a un subgrafo que formado por varios componentes (mutuamente desconectados). La *Intersección* y *Proyección* también ofrecen la posibilidad de obtener automáticamente resúmenes de la colección de textos, desde diferentes perspectivas, ya sea general (*Intersección*) o enfocados a conceptos de interés (*Proyección*).

2) Análisis Semántico Difuso

En el modelo propuesto, el proceso análisis semántico se lleva a cabo con el objetivo de integrar la información contenida en la colección de textos a través de la integración de los grafos que representan sus estructuras conceptuales. Este proceso se ejecuta como parte del procesamiento de las consultas definidas para la obtención de resultados sobre un contenido semánticamente integrado, aunque provenga de fuentes textuales diferentes. La integración de los grafos está basada en la integración de conceptos semánticamente similares. En [12] este proceso se realiza a partir de la identificación del sentido en WordNet de los conceptos (usando un algoritmo de desambiguación) y la unificación de aquellos que tengan el mismo significado. Sin embargo, el análisis semántico de la información textual, a nivel del significado de las palabras, está usualmente sujeto a la subjetividad, vaguedad y problemas de imprecisión, dada la inherente ambigüedad del lenguaje natural. Debido a esto y a las limitaciones mencionadas sobre el uso de algoritmos de desambiguación en este tipo de enfoques, es que se decide tratar este problema desde la perspectiva de la lógica difusa.

En esta nueva propuesta, el análisis de la similitud semántica difusa entre los conceptos se lleva a cabo mediante la combinación de la medida de similitud sintáctica de Levenshtein [17], con otras tres medidas de similitud semántica definidas sobre la base de lo reportado en [18] para medir la similitud sentencia-a-sentencia y varias de las medidas disponibles en el paquete WordNet::Similarity [20] (Resnik - R , Lin - LIN, and Jiang & Conrath – J & C)); un enfoque similar se aplica en [2]. Esta combinación se realiza a través de la función de agregación compensatoria reportada en [22] (ecuaciones 1 y 2) y la *t-norma* algebraica descrita en la ecuación (3), obteniendo un solo valor de similitud entre conceptos a partir de los valores numéricos (s_i) resultantes de

cada medida. En el campo de la medición de similitud semántica, las funciones de agregación son generalmente definidas y usadas para combinar varios valores numéricos, a partir de la agregación de los resultados de diferentes medidas de similitud semántica para obtener un único valor resultante [19]. El flujo de esta propuesta se muestra en Fig. 2, y la misma permite reducir los efectos negativos derivados de la incertidumbre que se produce en la decisión sobre qué medidas son más relevantes y el peso que ha dicha relevancia se asigna.

$$Z_\gamma(s_1, s_2, \dots, s_n) = \left(\prod_{i=1}^{n} s_i\right)^{1-\gamma} * \left(1 - \prod_{i=1}^{n}(1 - s_i)\right)^{\gamma} \quad (1)$$

$$\gamma = \frac{T(s_1, s_2, \dots, s_n)}{T(s_1, \dots, s_n) + T(1 - s_1, \dots, 1 - s_n)} \quad (2)$$

$$T(s_1, s_2, \dots, s_n) = \prod_{i=1}^{n} s_i \quad (3)$$

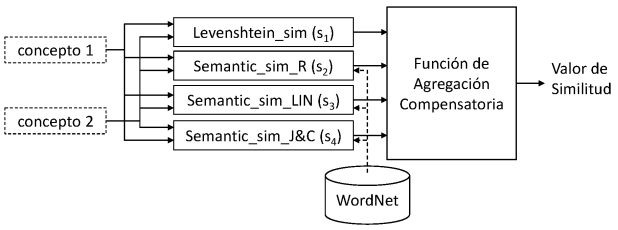

Fig. 2. Flujo de trabajo para la evaluación de la similitud entre conceptos.

En la integración de información entre dos grafos G_1 y G_2 se define un umbral U, que representa el valor mínimo de similitud existente entre dos nodos-conceptos para ser integrados. Cada nodo del grafo puede agrupar uno o varios conceptos en dependencia de las integraciones anteriores. En este proceso, se mide la similitud entre cada nodo N_i de G_1 y todos los nodos de G_2 y luego se integran aquellos nodos que tengan una similitud igual o superior a U. Este proceso se ejecuta iterativamente hasta que se compare el último nodo N_i con todos los nodos de G_2. La ecuación (4) describe el cálculo de la similitud entre dos nodos N y M, donde N está formada por el conjunto de conceptos $\{cn_1, cn_2, \dots, cn_n\}$ y M por otro conjunto de conceptos $\{cm_1, cm_2, \dots, cm_m\}$. Donde *simSD* es la función de agregación descrita anteriormente y *ln* y *lm* las cantidades de conceptos en los nodos N y M. Los nodos con etiqueta l_r, representan el primer concepto del grupo y estos conceptos se representan a su vez por conceptos separados por una coma y encerrados entre [ ] (ej. $[c_1, c_2]$).

$$sim(N, M) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} simSD(cn_i, cm_j)}{\ln * lm} \quad (4)$$

Los conceptos representados en un mismo nodo se consideran que tienen una fuente relación de sinonimia y por tanto, de esa manera son tratados en el modelo de recuperación concebido como parte de las consultas de *Proyección* (*Q*). En este tipo de consultas el usuario especifica uno o varios conceptos de interés (*CQ*), y selecciona un conjunto de grafos del RGC para formar el espacio de búsqueda (G_D). Se definieron varias reglas para identificar si un nodo $n_j / n_j \in N^{Gd}$ es recuperado o no, a partir del concepto $c_i \in CQ$. Estas reglas son ejecutadas en el mismo orden en que aparecen. Siendo, los conceptos $a / a \in CQ$, y $b / b \in N^{Gd}$, $T(c_i)$ el conjunto de palabras que conforman c_i, $l_r(b)$ la etiqueta representativa del nodo n_j y $ST(n_i)$ el grupo de conceptos de n_i. El nodo b es recuperado si: (R_1) $a \equiv l_r(b)$*(sintácticamente);* o (R_2) $a \in T(l_r(b))$*;* o (R_3) $a \in ST(b)$. Este tipo de consultas requiere que el usuario conozca los CQ que serán objeto de análisis. En tal sentido, se incorpora al modelo un mecanismo de identificación y recomendación de conceptos relevantes, el cual se describe a continuación.

3) Identificación y Recomendación de conceptos

En este proceso se obtiene un ranking de los *k* conceptos más relevantes de la colección de textos, teniendo en cuenta como criterios: frecuencia de aparición de los conceptos en los textos y la relevancia de los conceptos en el grafo que representa el contenido de la colección de textos. La frecuencia de aparición de los conceptos se computa en el mismo proceso de integración de los grafos generados que se ejecuta como parte de la *Unión*, a partir de esta consulta se extraen y ordenan los *k* conceptos más frecuentes de forma descendentemente para ser mostrados al usuario. La identificación de conceptos relevantes de acuerdo al segundo criterio se lleva a cabo mediante la aplicación del algoritmo PageRank [6] sobre el grafo resultante de la operación de *Unión*, en el cual está representada la conceptualización de la colección de textos. De la misma forma, como resultado se muestra al usuario de forma ordenada descendentemente los *k* conceptos con mayor valor de relevancia, según el PageRank. Esta funcionalidad potencia el uso de las consultas de *Proyección*, ya que el resultado de estos procesos sirve de guía al usuario en cuanto al análisis de los textos sobre la base de disponer de información sobre los conceptos más relevantes y representativos de la colección.

C. Recuperación de Pasajes

La inclusión de este proceso tiene el objetivo de ofrecer al usuario la posibilidad de recuperar fragmentos de textos de las fuentes primarias de información, a partir de conceptos representados en el grafo resultante de las consultas definidas para el análisis de los grafos, mediante el uso de Lucene. La forma en que se utiliza la recuperación de pasajes en este trabajo también puede considerarse como un tipo de análisis de grafos, pues se está analizando el grafo con respecto a los textos. En este caso, las consultas se construyen a partir de la selección por el usuario, de conceptos y/o proposiciones de interés representados en el grafo y se utiliza el conjunto de conceptos agrupados en un mismo nodo del grafo, para ejecutar un proceso de expansión de las consultas. Además, se emplea el operador *OR* para los conceptos de la consulta y *AND* para las proposiciones. La construcción de los pasajes se lleva a cabo a partir de la identificación de oraciones centrales en el Índice de Textos, las cuales pueden ser expandidas según el tamaño del pasaje (*PS*) especificado por el usuario. La oración central constituye la oración en la que aparecen los conceptos incluidos en la consulta, según el operador utilizado. *PS* se define como la cantidad máxima de oraciones que conforma el pasaje, y como este valor numérico puede ser par o impar, se

definieron reglas para determinar la cantidad real de oraciones (adyacentes a la oración central) a ser incluidas en el pasaje: *(R1) Si PS = 1 entonces se devuelve la oración central; (R2) Si PS es par entonces se incluyen las PS/2 oraciones anteriores a la oración central y las (PS/2)-1 oraciones contiguas; y (R3) Si PS es impar entonces se incluyen PS/2 oraciones anteriores y contiguas a la oración central.* Como resultado, por cada concepto y/o proposición, se muestran los pasajes de texto recuperados, los identificadores de los textos fuentes y la cantidad de pasajes recuperados por cada uno, resaltando los conceptos incluidos en la consulta.

IV. Evaluación del Modelo Propuesto: Caso de Estudio

La evaluación experimental de esta propuesta resulta compleja debido, fundamentalmente, a la ausencia de métodos y colecciones de evaluación estandarizados para este tipo de soluciones, en especial para la representación del texto [10], de la cual depende en gran medida el resto de los procesos. La evaluación de la recuperación de información suele enfocarse en la medición de resultados a partir de los documentos recuperados dado una consulta, y no a estructuras conceptuales incluidas en ellos, como se pretende con esta propuesta. En este sentido, se decidió aplicar un enfoque de evaluación similar a [12], en el contexto de las SLR.

La SLR es un medio de identificación, interpretación y evaluación de la evidencia científica relevante sobre una pregunta de investigación, tópico o fenómeno de interés [15]. Muchas de las tareas de la SLR, como la extracción y síntesis de datos, requieren alto consumo de tiempo, y mucho trabajo manual, implicando un gran esfuerzo [3]. En la extracción de datos se extraen de cada uno de los estudios un conjunto de datos (ej. bibliográficos, cuantitativos y cualitativos) que se definen en la fase de planificación en forma de un formulario. En la síntesis se combinan esto datos para darle respuesta a las preguntas de investigación iniciales. Este caso de estudio se enfoca en las tareas de extracción y síntesis de datos de la SLR reportada en [8], donde se analizan 11 artículos científicos. El objetivo del caso de estudio es ejemplificar la aplicabilidad del modelo propuesto para las tareas mencionadas. A partir de esos artículos, se conformaron 11 textos, con 822 palabras y 33 oraciones cada uno (como promedio) usando las secciones de resumen, introducción y conclusiones de cada artículo. Según la propuesta, inicialmente se indexan los textos y se generan los grafos a partir de ellos, constituyéndose el Índice y el RGC.

El modelo propuesto permite soportar y combinar la extracción y síntesis de datos mediante la utilización y combinación de sus principales procesos. La aplicación de la consulta de intersección (*Inter*) (similar a [12]) y el uso de la recomendación de conceptos, obtienen los conceptos que tienen una frecuencia igual o superior al *SV* en la colección de textos y las relaciones entre ellos representadas en los grafos. Estos conceptos pueden considerarse palabras claves o terminología que caracteriza ese contenido, y permitan responder a las preguntas de investigación. En la Fig. 3 se ejemplifica el resultado de la consulta $Inter^{50}$(RGC), donde el tamaño de los nodos representa el nivel de frecuencia de los conceptos. En esta Fig. se identifican [*software*], [*dependability*, *reliability*] e [*ISO*] como conceptos más relevantes, lo que se corrobora con el título del artículo [8], donde tres de ellos aparecen. En este ejemplo, también se ilustran resultados del análisis semántico difuso en la unificación de '*dependability*' y '*reliability*', lo cual contribuye a conocer rápidamente la terminología usada para referirse al mismo concepto. La relación '*strong relation*' indica que los conceptos [*software*] e [*ISO*] tienen una fuerte relación contextual, sugiriendo un análisis adicional, que puede soportarse con el uso de la *Proyección* y/o la recuperación de pasajes. El uso de la recuperación de pasajes en este trabajo brinda, a diferencia de las otras propuestas, la posibilidad de conocer en que fuentes se encuentran los conceptos y su contexto, permitiendo extraer datos que se requieran para llenar los formularios de los estudios, y así poder registrar esta información. Además, de las etiquetas de los campos del formulario pueden obtenerse conceptos de referencia que sirvan de base al modelo para obtener información, principalmente cualitativa, que sirva tanto para la extracción como para la síntesis de datos.

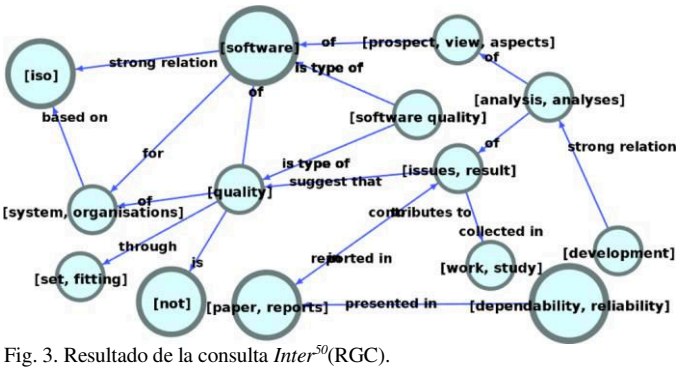

Fig. 3. Resultado de la consulta $Inter^{50}$(RGC).

Al tener identificados conceptos de interés se puede aplicar la consulta de *Proyección*, así como la recuperación de pasajes para profundizar en el análisis de los mimos. En este caso, se ejecutó esta consulta usando el RGC como espacio de búsqueda y '*standard*' como concepto de interés ($Proj^{1}$*(RGC, standard)*), respondiendo a la pregunta: *"Which software reliability models have been developed by following the recommendations in International Standards?"* [8], donde '*standard*' es uno de los términos relevantes. El resultado se muestra en la Fig. 4, donde se representan varios conceptos relacionados con '*standard*' y asociados a '*International Standards*', tales como: *ISO*, *IEEE*, *SQuaRE*, y *COSMIC*. En la Fig. 4 se representan también conceptos presentes en diferentes fuentes textales, lo cual es posible a partir del análisis semántico difuso concebido. Este análisis se complementó con la recuperación de los pasajes, donde aparecen algunos de los conceptos representados en la Fig. 4 (se muestra en la misma figura), y puede ser utilizado en la extracción de datos para conocer en estudios aparecen los distintos estándares. Los resultados expuestos reflejan beneficios de la propuesta como soporte a las SLR para las tareas de extracción y síntesis de datos a partir del análisis del contexto en el que están siendo usados los conceptos, la obtención de vistas resumidas y sintetizadas del contenido, que facilitan el análisis cualitativo y cuantitativo sobre los conceptos, entre otros aspectos.

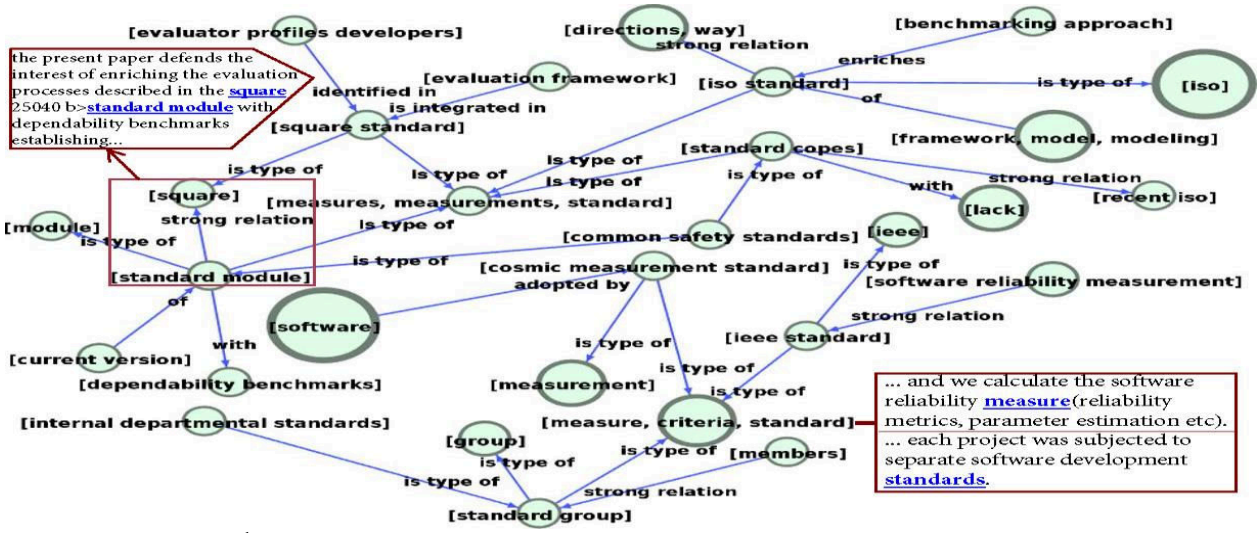

Fig. 4. Resultado de la consulta *Proj*^{I}*(RGC, standard)*.

CONCLUSIONES Y TRABAJO FUTURO

Este trabajo presentó un modelo de análisis de textos basado en grafos, soportado en el procesamiento semántico difuso del contenido. En el modelo se combinan técnicas de análisis de grafo y recuperación de pasajes, para recuperar información relevante de una colección de textos. El enfoque difuso propuesto para el análisis semántico, donde se combinan medidas de similitud entre conceptos a través de una función de agregación, permitió incrementar los resultados obtenidos en las consultas. El caso de estudio desarrollado evidenció beneficios de esta propuesta para soportar computacionalmente las tareas de extracción y síntesis de datos en el contexto de una SLR. En trabajos futuros, se considerarán otras operaciones sobre los grafos como la detección de patrones frecuentes, así como otros operadores de agregación y métricas de similitud en el análisis semántico sobre los conceptos.

AGRADECIMINETOS

Este trabajo ha sido parcialmente financiado por el proyecto METODOS RIGUROSOS PARA EL INTERNET DEL FUTURO (MERINET), financiado por el Fondo Europeo de Desarrollo Regional (FEDER) y el Ministerio de Economía y Competitividad (MINECO), Ref. TIN2016-76843-C4-2-R.

REFERENCIAS

[1] Aggarwal, C. C., y Zhai, C. X.: “Mining Text Data”. Springer (2012).

[2] Al-Amin, S. T., M. Hasan, y M. M. A. Hashem. “A Modified Fuzzy Relational Clustering Approach for Sentence-Level Text”. In Proceedings of EICT 2015, 2015, pp. 1-5.

[3] Al-Zubidy, A., J. C. Carver, D. P. Hale, y E. E. Hassler. “Vision for SLR tooling infrastruc-ture: Prioritizing value-added requirements”. Information and Software Technology. 91, 2017, pp. 72-81.

[4] Benkoussas, C., y P. Bellot. “Information Retrieval and Graph Analysis Approaches for Book Recommendation”. The Scientific World Journal. 2015, pp. 1-8.

[5] Bronselaer, A. y G. Pasi. “An approach to graph-based analysis of textual documents”. In Proceedings of EUSFLAT. 2013, pp. 634-641.

[6] Brin, S. y L. Page. “The anatomy of a large-scale hypertextual Web search engine”. Comput. Netw. ISDN Syst. 30(1-7), 1998. pp. 107–117.

[7] Chang, J. Y., y I. M. Kim. “Research Trends on Graph-Based Text Mining”. Int. J. of Software Engineering and Its Applications. 8(4). 2014, pp. 147-156.

[8] Febrero, F., C. Calero, y M. A. Moraga. “Software reliability modeling based on ISO/IEC SQuaRE”. Information and Software Technology. 70. 2016, pp. 18-29.

[9] Felizardo, B. K., G. Andery, F. V. Paulovich, R. Minghim, y J. C. Maldonado. “A Visual Analysis Approach to Validate the Selection Review of Primary Studies in Systematic Reviews”. Information and Software Technology. 54. 2012, pp. 1079-1091.

[10] Glauber, R., y D. Barreiro-Claro. “A systematic mapping study on open information extraction”, Expert Syst. Appl. 112. 2018, pp.372-387.

[11] Hassan, G. S., A. K. Abdulsahib, y S. S. Kamaruddin. “Graph-Based Text Representation: A Survey of Current Approaches”. Research J. of Applied Sciences, Engineering and Technology. 14(9), 2017, pp. 334-340.

[12] Hojas-Mazo, W., A. Simón, M. de la Iglesia, F. P. Romero, J. A. Olivas. “A concept-based text analysis approach using knowledge graph”. In Proceedings of IPMU 2018, CCIS 854, 2018, pp. 696–708.

[13] Indurkhya, N. “Emerging directions in predictive text mining”. WIREs Data Mining and Knowledge Discovery. 5, 2015, pp. 155-164.

[14] Karim, G., T. K. Mouna, T. Lynda, y B. J. Maher. “Graph-based Methods for Significant Concept Selection”. Procedia Computer Science. 60, 2015, pp. 488–497.

[15] Kitchenham, B. A., y S. Charters. “Guidelines for performing systematic literature reviews in software engineering”. Tech. Rep.EBSE 2007-001. Keele University and Durham University Joint Report. 2007.

[16] Kleinberg, J. M. “Authoritative sources in a hyperlinked environment”. Journal of the ACM. 46(5), 1999, pp. 604–632.

[17] Levenshtein, A. “Binary Codes Capable of Correcting Deletions, Insertions and Reversals”. Soviet Physics Doklady. 10(8), 1966, pp. 707-710.

[18] Li, Y., D. McLean, Z. Bandar, J. O’Shea, y K. Crockett. “Sentence similarity based on semantic nets and corpus statistics”. IEEE Transactions on Knowledge and Data Engineering. 2006, pp.1138–1150.

[19] Martinez-Gil, J. “CoTO: A Novel Approach for Fuzzy Aggregation of Semantic Similarity Measures”. Cognitive Systems Research, 40. 2016, pp. 8-17.

[20] Pedersen, T., S. Patwardhan, y J. Michelizzi. “WordNet::Similarity - Measuring the Relatedness of Concepts”. In: Proc. of the 19th National Conference on Artificial Intelligence (AAAI-04), 2004, pp. 1024-1025.

[21] Sonawane, S. S., y P. A. Kulkarni. “Graph based Representation and Analysis of Text Document: A Survey of Techniques”. International Journal of Computer Applications. 96(19), 2014, pp. 1-8.

[22] Zimmermann, H.-J., y P., Zysno. “Latent connectives in human decision making”. Fuzzy Sets and Systems. 4(1), 1980, pp. 37-51.