Ciencias de la Información Vol. 45, No. 3, septiembre-diciembre, pp. 37 - 44, 2014

Recuperación de conceptos similares de un Corpus de Mapas Conceptuales

Manuel de la Iglesia Campos
Alfredo Simón Cuevas
José A. Olivas

En este trabajo se propone un método que utiliza un corpus de mapas conceptuales para recuperar del mismo los conceptos similares a un concepto en específico, que se basa en la aplicación de un algoritmo de comparación entre grafos, y de otro de similitud entre mapas conceptuales. Este método se propone como solución para el tratamiento de la ambigüedad de aquellos conceptos que no pueden ser desambiguados cuando no se encuentran presentes en WordNet, o cuando la información aportada por este recurso no permite seleccionar un sentido como el más apropiado. En los experimentos realizados fue posible demostrar una correspondencia entre los resultados del método y aquellos obtenidos mediante una métrica de similitud semántica basada en WordNet. Esta propuesta puede ser útil para otras tareas como la construcción de mapas conceptuales, y la integración semántica entre mapas conceptuales.

Palabras clave: mapas conceptuales; similitud entre conceptos; procesamiento de lenguaje natural; desambiguación; algoritmo de Flujo de Similitud

RESUMEN

ABSTRACT

A method is proposed in this work that uses a corpus of concept maps to recover from it the concepts similar to some specific concept. It is based on the application of an algorithm for the comparison of graphs, and another algorithm that determines the similarity between concept maps. This method is proposed as solution for the treatment of the ambiguity of those concepts that cannot be disambiguated for not being present in WordNet, or when the information existing in this resource is not enough for allowing the selection of the most appropriate sense. A correspondence between the results of the method, and those obtained using a semantic similarity metric based in WordNet, was obtained in the experiments that were performed. This proposal can be useful in other tasks such as the concept maps construction, and the semantic integration between concept maps.

Keywords: concept maps; concept similarity; natural language processing; disambiguation; Similarity Flooding algorithm

Introducción

Los mapas conceptuales (MC) son "herramientas gráficas para organizar y representar el conocimiento" (J.D. Novak & Cañas, 2006), siendo considerados además como una técnica que representa, simultáneamente, una estrategia de aprendizaje, un método para captar lo más significativo de un tema y un recurso esquemático para representar un conjunto de significados conceptuales incluidos en una estructura de proposiciones (J. D. Novak & Gowin, 1984). Aunque los MCs se han usado fundamentalmente como apoyo a la enseñanza y el aprendizaje, en general constituyen una herramienta de gran utilidad para la gestión del conocimiento (García & Artiles, 2004; Simón Cuevas, 2008).

Los MCs están compuestos por conceptos y relaciones entre ellos, conformando proposiciones. Los conceptos son definidos

como regularidades o evidencias percibidas en eventos u objetos, hacen referencia a hechos, objetos, cualidades y animales, entre otros, y se designan mediante etiquetas. Las relaciones se etiquetan con una frase de enlace que especifica el tipo de relación que se establece entre los conceptos. Las proposiciones se forman por conceptos conectados por una frase de enlace, y pueden ser leídas independientemente del mapa conceptual y aún así tener sentido (A. J. Cañas & Carvalho, 2004). En el MC estos elementos se representan gráficamente: los conceptos como nodos; y las relaciones como enlaces dirigidos entre los nodos. Esta estructura es similar a la de un grafo, por lo cual puede considerarse al MC como un grafo dirigido con nodos y enlaces etiquetados.

La flexibilidad con que el conocimiento puede ser representado en el MC, ha favorecido su amplia utilización en varios entornos, pero al mismo tiempo hace más complejo el procesamiento computacional del conocimiento representado, ya que las computadoras no pueden "entender" ese contenido. Ante esta limitación se ha propuesto introducir cierta formalización en la representación de los MCs restringiendo en alguna medida el uso de frases de enlace y/o conceptos; o aplicar técnicas de la Inteligencia Artificial, esto último sin comprometer la flexibilidad de los MCs, ni introducirles formalismos (A. J. Cañas & Carvalho, 2004). Este trabajo se enfoca en esta última variante.

La organización del trabajo es la siguiente: en la Sección 2 se analiza la problemática referente a la ambigüedad del lenguaje en los MC, y se describe de forma general el método que se presenta. La Sección 3 muestra los recursos y algoritmos que se utilizan en la propuesta. En la Sección 4 se presenta el método para la obtención de conceptos similares desde un corpus de mapas conceptuales. La sección 5 contiene la evaluación de la propuesta. Finalmente, en la Sección 6 se exponen las conclusiones del trabajo.

La ambigüedad en los Mapas Conceptuales

La ambigüedad es una característica inherente al lenguaje natural y ocurre cuando una palabra o frase puede tener más de un sentido o interpretación, hecho conocido por polisemia (Agirre & Edmonds, 2006). Los MCs son una representación simplificada de la estructura cognitiva de la persona (J. D. Novak & Gowin, 1984), en la que los conceptos, frases de enlace y proposiciones se expresan en lenguaje natural, por lo que están sujetos a ambigüedad (Costa, Rocha, & Favero, 2004). Esta ambigüedad presente en los MCs, y en particular la asociada a los conceptos, hace más complejo el análisis computacional de su contenido, principalmente cuando se requiere de información semántica; como por ejemplo para la integración automática de MCs y la construcción semi-automática de MCs.

El problema de determinardesde el punto de vista computacional el "sentido" o significado de una palabra por su uso en un contexto dado, a partir de un conjunto de sentidos posibles, es conocido como desambiguación semántica o del sentido de las palabras (WSD, según las siglas del inglés Word Sense Disambiguation) (Eneko Agirre & Edmonds, 2006).

A pesar de los rasgos de similitud estructural entre un MC y un texto, y el desarrollo alcanzado en la solución de la ambigüedad en textos (Agirre & Edmonds, 2006) (Navigli, 2009), el problema de la ambigüedad en MCs ha sido abordado solo en (A. Cañas, Valerio, Lalinde, Carvalho, & Arguedas, 2003), (Simón Cuevas, Ceccaroni, Rosete, Suárez, & de la Iglesia Campos, 2008) y (de la Iglesia Campos, 2012). En estos trabajos se proponen métodos que se basan en el conocimiento para obtener el sentido de los términos en un MC usando WordNet, recurso que será descrito en la siguiente sección de este trabajo. Estos métodos de desambiguación comparten el problema de la imposibilidad de desambiguar términos cuando estos no se encuentran en Word Net, o cuando la información aportada por este recurso no permite seleccionar un sentido como el más apropiado. Una posible solución puede ser el empleo de fuentes de información distintas a WordNet, como ontologías o corpus de MCs.

En este trabajo se presenta un método que emplea como fuente de conocimiento un corpus de MCs, sugiriendo del mismo aquellos conceptos con determinada similitud con respecto a un concepto de entrada o concepto a tratar. El método usa WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990; Fellbaum, 1998) como repositorio de sentidos. Al considerar a los MCs como grafos etiquetados, se utiliza el algoritmo Similarity Flooding (Melnik, Garcia-Molina, & Rahm, 2002) para identificar las correspondencias entre conceptos provenientes de dos MCs. En cierta medida, estas correspondencias son relativas a los grafos involucrados, por lo que la aplicación de este algoritmo se combina con el empleo de un algoritmo de comparación entre MCs (Cañas, Leake, & Maguitman, 2001; Leake, Maguitman, & Cañas, 2002) para poder determinar la similitud existente entre el MC de origen del concepto de entrada, y cada uno de los MCs integrantes del corpus. Al mismo tiempo, esto permite descartar aquellos MC muy poco similares, incrementando la eficiencia de la propuesta al evitar ejecuciones innecesarias del algoritmo Similarity Flooding.

Este método, dada su utilidad para sugerir conceptos sinónimos, permite tratar aquellos conceptos que, por alguna razón, no pueden ser desambiguados por los métodos de desambiguación existentes, además de que pudiera utilizarse como herramienta en otras tareas como la integración a nivel semántico entre mapas conceptuales, y la construcción semiautomática de mapas conceptuales (Kowata, Cury, & Silva Boeres, 2010).

Recursos y algoritmos empleados

WordNet

WordNet (WN) (Miller et al., 1990; Fellbaum, 1998) es una base de datos léxica, y también ha sido considerada como una ontología (Agirre & Martinez, 2001). Almacena conocimiento de dominio general. Su estructura se basa en conjuntos de sinónimos, denominados *synsets*, que definen el sentido de las palabras (en el caso de polisemia, existen varios *synsets* para una misma palabra). Los *synsets* estan conformados por: un identificador, el conjunto de palabras que comparten el mismo sentido, la categoría gramatical a la que pertenecen estas palabras, y una descripción textual del sentido (glosa).

Los *synsets* están organizados básicamente por relaciones semánticas y léxicas (estas

últimas entre palabras y no entre sentidos), conformando una gran red. Entre estas relaciones se encuentran: hiperonimia/ hiponimia, meronimia/holonimia, antonimia, causa, instancia, roles y sinonimia, cada una de ellas con sus acepciones y especificaciones (Climent, Rodríguez, & Gonzalo, 2008). Cada *synset* se etiqueta con uno o varios de los dominios de una taxonomía de dominios de ámbito general denominada WordNet Domains (Bentivogli, Forner, Magnini, & Pianta, 2004; Magnini & Cavaglia, 2000; Magnini, Strapparava, Pezzulo, & Gliozzo, 2002), identificando dónde pueden ser empleadas las palabras agrupadas por el *synset*.

Algoritmos de comparación entre mapas conceptuales

En la literatura consultada se encuentran dos propuestas que abordan la temática de la comparación entre varios MCs: la presentada por Cañas y colaboradores en (Alberto Cañas et al., 2001), y como continuación de esta, la presentada en (Leake et al., 2002). En estos trabajos se presentan métodos para determinar la similitud existente entre dos MCs, desarrollados a partir del trabajo reportado en (Kleinberg, 1998) sobre análisis topológico de grafos, en el que se le asignaban pesos a los nodos según sus roles como "concentradores" o "autoridades".

En la propuesta de (Alberto Cañas et al., 2001) se categorizan los conceptos a partir de cuatro tipos, categorías o roles de nodos: los nodos autoridades; que son conceptos hacia los cuales convergen otros conceptos, y son los conceptos con mayor cantidad de enlaces provenientes de nodos concentradores; los nodos concentradores (o centros de actividad), que son los conceptos con mayor cantidad de enlaces apuntando hacia nodos autoridades; los nodos superiores son conceptos que generalmente aparecen en la parte superior del MC en su representación gráfica; y los inferiores son conceptos que generalmente aparecen en la parte inferior del MC cuando se representa gráficamente. A cada concepto se le asignan cuatro pesos (denominados respectivamente *a-weight, h-weight, u-weight* y *l-weight*) que representan el grado en que el concepto pertenece a cada una de las cuatro categorías expuestas. El rol de cada concepto en un MC se caracteriza por medio de sus pesos asociados, por lo que los roles de conceptos de MCs diferentes pueden confrontarse mediante la comparación de estos pesos.

Para comparar dos mapas conceptuales, se determina cuánto corresponden entre sí los conceptos individuales de ambos mapas conceptuales, teniendo en cuenta cada categoría de pesos y la cantidad de palabras en común entre las etiquetas de los conceptos. Tanto en (Cañas et al., 2001) como en (Leake et al., 2002), la similitud total entre mapas conceptuales se determina a partir de computar, de diferentes formas, el conjunto de pesos obtenidos en cada una de las cuatro categorías que se han mencionado.

Algoritmo Similarity Flooding

El algoritmo de Flujo de Similitud, o Similarity Flooding (Melnik et al., 2002) es un algoritmo para la identificación de correspondencias entre elementos de diversas estructuras de datos, denominadas en la propuesta como modelos. Se basa en el planteamiento de que elementos de dos modelos distintos son similares cuando sus elementos adyacentes son similares entre sí (Melnik et al., 2002). Entre sus ventajas se encuentra que permite, mediante el empleo de estructuras basadas en grafos, determinar correspondencias entre elementos para los cuales no se puede determinar una semejanza de forma directa, y puede ser empleado en modelos de diferente tamaño y estructura. El algoritmo se resume a continuación:

1. Se crea un grafo para cada uno de los modelos. Cada nodo del grafo se asocia a un elemento del modelo.

2. Se construye un "grafo (dirigido) de propagación de similitud", en el cual cada nodo (*a*,*b*) representa un mapeo entre un nodo *a* de uno de los grafos y un nodo *b* del otro grafo.

3. Se crea una arista desde cada nodo (*a, b*) hasta cada nodo (*a1, b1*) si y solo si existen aristas desde *a* hasta *a1* y desde *b* hasta *b1* en los respectivos grafos de origen.

4. A cada nodo (*a*,*b*) se le asocia el valor de similitud existente entre *a* y *b*.

5. Los valores de similitud de cada nodo del grafo de propagación se incrementan a partir de los valores de similitud de los nodos vecinos, multiplicados por los coeficientes de propagación. Este "flujo de similitud" se realiza de forma iterativa, por lo que la similitud inicial se propaga a través de los enlaces hacia los nodos adyacentes.

6. El flujo se termina cuando se logra una convergencia de los valores de similitud, o luego de un número máximo definido de iteraciones.

7. Se selecciona el subconjunto de pares de nodos que constituyen las correspondencias más aceptables, mediante un proceso de filtrado.

En (Marshall, Chen, & Madhusudan, 2006) se presenta la aplicación del *Similarity Flooding* para determinar correspondencias entre los conceptos de pares de MCs, construyendo un grafo en el cual cada concepto y su etiqueta se representan como nodos independientes, asignándose un enlace del nodo correspondiente a la etiqueta hacia el nodo correspondiente al concepto. Se plantea que las frases de enlace también pueden representarse como nodos en el grafo. En esta propuesta se obtienen buenos resultados, en un entorno de experimentación muy controlado.

Método para la obtención de conceptos similares desde un Corpus de Mapas Conceptuales

A continuación se describe el método para obtener, dado un concepto específico de un mapa conceptual, aquellos conceptos de un corpus de MCs que sean similares al mismo, y el valor de similitud que existe entre ellos. Para ello deben considerarse las siguientes convenciones:

• *CMC*: Corpus de mapas conceptuales $\{mc_1, mc_2, mc_3, \ldots, mc_n\}$ que conforman el corpus .

• c_e: Es el concepto de entrada, para el cual se recuperarán los conceptos similares dentro de *CMC*.

• *mc*: Es el mapa conceptual donde se encuentra c_e.

• *Ctx* (*m*, *c*, *r)*: Es un contexto de *c*, conformado por el conjunto de conceptos y frases de enlace del mapa conceptual *m* en un radio *r* alrededor del concepto *c*. A su

vez también constituye un mapa conceptual.

• *Sim* (mc_i, mc_j): Medida de similitud entre dos mapas conceptuales mc_i y mc_j, obtenida mediante la aplicación de un algoritmo de comparación entre mapas conceptuales como los reportados en (Cañas et al., 2001) y (Leake et al., 2002), el último de los cuales es el empleado en la propuesta.

• *SF* (mc_i, mc_j, c): Medida de similitud entre el concepto c de mc_i y cada concepto c_k de mc_j, como un conjunto de pares (c_k, sim_k), siendo sim_k el valor de similitud entre c y c_k. Esta medida se obtiene aplicando el método Similarity Flooding (Marshall et al., 2006; Melnik et al., 2002), filtrando sus resultados para el concepto c.

Primeramente se selecciona de *CMC* el conjunto de aquellos *MC* más similares a un contexto en el que se encuentra el concepto de entrada c_e, dentro de *mc*, empleando para ello un algoritmo de comparación entre *MC*. Luego se aplica el algoritmo *Similarity Flooding* entre *mc* y cada $mc_i \in CMC$ seleccionado, y el resultado se combina con la similitud entre contextos de los mapas conceptuales de origen, obteniéndose finalmente un conjunto conformado por conceptos del corpus y sus valores de similitud global con respecto a c_e, que puede emplearse para reducir la ambigüedad de c_e. Los pasos del método se detallan a continuación:

Paso 1: Obtener el conjunto MCS de mapas conceptuales de CMC más similares a un contexto.

En este primer paso, a partir de un contexto de *mc* en el que se encuentra el concepto de entrada c_e, se obtiene el conjunto *MCS* conformado por parejas (mc_i, sim_i), donde mc_i es cada mapa conceptual de *CMC*, y sim_i su correspondiente valor de similitud con respecto al contexto de c_e. Mientras *MCS* esté vacío y se pueda ampliar el contexto inicial, se amplía el radio del contexto en una unidad. Se toma como radio inicial del contexto el valor 1, pues fue apreciado empíricamente que, para contextos pequeños, se obtuvieron resultados más precisos en la posterior aplicación del *Similarity Flooding*. A continuación se muestra el seudocódigo de este paso:

$r = 1$
$MCS = \{\}$
hacer
  definir $mc' = Ctx\ (mc, c_e, r)$
  □ $mc_i \in CMC$
    $sim_i = Sim\ (mc', mc_i)$
    Si $sim_i > 0$ y mc_i no contiene a c_e
    $MCS = MCS$ □ $\{(mc_i, sim_i)\}$
  $r = r + 1$
mientras $mc' \neq mc$ y $MCS = \{\}$
Ordenar el conjunto $MCS \mid sim_i \geq sim_{i+1}$ (mc_i, sim_i), (mc_{i+1}, sim_{i+1}) $\in MCS$

Si $MCS = \{\}$ el método termina sin aportar resultados.

Paso 2: Obtener la similitud entre el concepto de entrada y cada uno de los conceptos de los mapas conceptuales más similares.

En este paso se aplica el *Similarity Flooding* entre el contexto *mc'* seleccionado en el paso anterior y los 5 mapas conceptuales de *MCS* con mayor valor de similitud respecto a *mc'*. Esta cantidad de *MC* fue determinada empíricamente, con el objetivo de lograr un balance entre eficiencia y eficacia, pues es poco probable que con *MC* poco similares se obtengan resultados favorables, y al mismo tiempo el costo computacional de la aplicación del *Similarity Flooding* entre dos mapas conceptuales es relativamente alto. En la aplicación del Similarity Flooding a cada nodo (*a*, *b*) del grafo de propagación de similitud se le asigna como valor de similitud inicial el obtenido al aplicar una métrica de similitud semántica entre los conceptos *a* y *b*. El seudocódigo de este paso es el siguiente:

$S_{ed} = \{\}$
$d = 1$
mientras d d» 5 y $d < |MCS|$
  mc_d es el mapa conceptual del par (mc_d, sim_d) en la posición d del conjunto *MCS*
  $S_{ed} = S_{ed}$ □ $SF\ (mc', mc_d, c_e)$
  $d = d + 1$
eliminar de S_{ed} todos los pares
  $(c_i, sim_i) \mid sim_i = 0$

Se obtiene como resultado de este paso el conjunto S_{ed} de pares conformados por conceptos de *MCS* y su similitud con respecto a c_e

Paso 3: Obtener la medida global de similitud.

Se refina la medida de similitud sim_i obtenida para cada concepto c_i en el paso anterior, a partir de su multiplicación por la similitud sim_{ctx} existente entre el contexto *mc'* del concepto c_e y un contexto mc_k' en el que se encuentra cada c_i en su mapa conceptual de origen. Este paso se realiza debido a que sim_i parece ser una medida de similitud relativa, que expresa la correspondencia entre c_e y c_i en el ámbito de sus contextos de origen, y con respecto a los demás conceptos en un mismo contexto, por lo cual es necesario ajustarla considerando la similitud entre dichos contextos. Se toma como radio inicial del contexto mc_k' el valor 3 pues a partir de este valor fueron apreciados mejores resultados en la comparación entre los contextos mediante el algoritmo de comparación entre mapas conceptuales utilizado. El seudocódigo de este paso se muestra a continuación:

Para cada $(c_i, s_i) \in S_{ed}$
  $r = 3$
  hacer
    $mc_k' = Ctx\ (mc_k, c_i, r)$, donde mc_k es el mapa conceptual de origen de c_i
    $sim_{ctx} = Sim\ (mc', mc_k')$
    $r = r + 1$
  mientras $sim_{ctx} = 0$ y $mc_k' \neq mc_k$
  $s_i = s_i * sim_{ctx}$

Ordenar $S_{ed} \mid sim_i$ e» sim_{i+1} (mc_i, sim_i), (mc_{i+1}, sim_{i+1}) $\in MCS$

Paso 4: Interpretar el resultado final.

El resultado del método es un conjunto de elementos S_{ed}, donde cada elemento está compuesto por un concepto y el valor de similitud global que se sugiere con respecto al concepto de entrada c_e. Para su empleo en el tratamiento de la ambigüedad de c_e, y de acuerdo a una de las variantes propuestas para procesar los resultados, se propone proceder de la siguiente forma:

Sea (c_1, sim_1) el primer elemento de S_{ed}. Por tanto se asume que c_1 es el concepto más similar a c_e, según propone el método, y puede interpretarse como sinónimo de c_e. Si c_1 se encuentra en un *synset* en WordNet,

se le asigna dicho *synset* a c_e como propuesta de sentido. Si se encuentra en más de un *synset*, se procede a la determinación del synset más apropiado teniendo en cuenta los contextos de origen de c_e y c_1, mediante la aplicación de un método de desambiguación. Si c_1 no aparece en ningún synset en WordNet, es posible inducir un nuevo sentido, partiendo de las ocurrencias encontradas (Di Marco & Navigli, 2013; Nasiruddin, 2013). Por tanto, este nuevo sentido debe etiquetar a c_e y c_1 como sinónimos, y estar definido (a modo de glosa) a partir de un contexto de ocurrencia conformado por la intersección entre *mc* y el conjunto de los conceptos del mapa conceptual de origen de c_1. Este sentido tendría una estructura similar a la de un synset de WordNet.

Otra posible aplicación de los resultados de esta propuesta para tratar la ambigüedad semántica pudiera ser interpretar como sentido del concepto de entrada al conjunto de conceptos recuperados más similares y los valores de similitud respectivos, con un contexto de ocurrencia conformado por la intersección de los contextos en que se encuentran los conceptos en sus MCs de origen.

Resultados Experimentales

Diseño del experimento

En la evaluación del método para la reducción de ambigüedad en conceptos se utilizó como muestra un corpus de MCs en idioma español provenientes de diferentes fuentes de la literatura, al no existir un corpus de MCs de referencia que pueda utilizarse como recurso para este tipo de tareas.

La tarea fundamental de la evaluación consistió en determinar si los resultados de similitud del método guardaban alguna correspondencia o relación con la similitud semántica existente entre los conceptos, esta última calculada empleando una métrica basada en WordNet. Esta correspondencia se determinó empleando criterios estadísticos basados en el análisis de la correlación utilizando el coeficiente de correlación de Pearson. Para poder determinar la similitud semántica existente entre cada concepto de entrada y los conceptos recuperados, la pareja de conceptos debía tener *synsets* en WordNet; siendo necesario a la vez simular que el concepto de entrada se comportara como un concepto sin *synsets* en WordNet durante la ejecución del método, pues de otra forma no tendría sentido el experimento. Por tanto fueron seleccionados como conceptos de entrada un conjunto de conceptos de diferentes MCs del propio corpus, cuyas etiquetas estuvieran presentes en WordNet, y se adaptó el método de recuperación de conceptos similares para no hacer uso de los synsets de estos conceptos.

Para la selección de los conceptos que constituirían la entrada del método primeramente se identificaron dentro del propio corpus varios grupos de MCs que abordan dominios específicos, como por ejemplo: los relacionados con la microbiología (ej.: "Biología Celular", "Células"); con elementos de la química y física (ej.: "Aire", "Átomos", "Electrones", "Molécula", "Nitrógeno", "Universo"); y con la biología (ej.: "Animales", "Fauna", "Insectos","Plantas","Ser vivo"). El método se aplicó sobre estos MCs, al existir mayor información de estos dominios en el corpus. De estos MCs, se seleccionó para la realización de los experimentos un conjunto de conceptos que estuvieran presentes también en otros MCs del corpus. Los MCs de origen de los conceptos fueron excluidos del corpus durante la ejecución del método.

Tabla 1: Resultados de la aplicación del método para la recuperación de conceptos similares de un corpus de mapas conceptuales

Concepto	Mapa conceptual	CCR	Pos	Sim	DesvE	Corr	Prob
Aire	Aire	7	4	0.156	0.018	0.427	0.339
Nitrógeno	Aire	10	2	0.030	0.017	0.430	0.215
Electrones	Átomos	4	1	0.117	-	0.731	0.269
Positiva	Átomos	2	No	No	No	1.000	-
Células	Célula	4	1	0.106	-	0.936	0.064
Núcleo	Célula	8	1	0.062	-	0.873	0.005
Insectos	Insectos	7	1	0.357	-	0.562	0.190
Agua	Molécula	10	5	0.032	0.012	0.172	0.635
Gas	Molécula	25	1	0.058	0.000	0.184	0.378
Movimiento	Molécula	22	19	0.015	0.055	0.073	0.746
Organismos	Nitrógeno	12	4	0.017	0.008	0.005	0.989
Oxígeno	Plantas	5	1	0.093	-	0.662	0.224
Suelo	Plantas	6	3	0.015	0.013	0.070	0.895
Nitrógeno	Ser vivo	4	No	No	No	-0.425	0.575
Flora	Ser vivo	9	1	0.057	-	-0.302	0.429
Átomos	Universo	10	4	0.042	0.127	0.317	0.372
Materia	Universo	9	6	0.022	0.036	-0.018	0.963
Energía	Universo	4	2	0.017	0.026	-0.507	0.493
Electrones	Universo II	11	1	0.291	-	0.476	0.139
Sustancias	Universo II	5	1	0.055	-	0.520	0.369
Promedio		8.3	2.5	0.081	0.038	0.309	0.467
Desviación estándar		5.759	4.264	0.096	0.037	0.434	0.295
Muestra Total		174				0.256	0.001

"-" indica que el tamaño de la muestra no es suficiente para realizar el cálculo.
CCR: Cantidad de conceptos recuperados.
Pos: Posición del concepto sinónimo (*S*) del concepto de entrada, con mayor valor de similitud, si existe.
Sim: Similitud global obtenida para el concepto *S*.
DesvE: Desviación estándar de los valores de similitud mayores o iguales que el de *S*.
Corr: Coeficiente de correlación de Pearson, entre los resultados del método y los valores de similitud semántica.
Prob: Valor de probabilidad (*p-value*) obtenido durante la prueba de correlación.

Resultados generales

De los resultados obtenidos se descartaron aquellos cuyo valor de similitud global se encontraba por debajo del umbral 0.01, considerando estos valores como poco significativos. Siguiendo este criterio, se excluyeron del análisis, 3 conceptos de entrada que no aportaron ningún resultado mayor o igual al umbral seleccionado. Finalmente fueron seleccionados 20 conceptos, derivadas de 10 MCs diferentes. De estos 20 conceptos, solo 2 de ellos (el 10 %) no aportaron en sus resultados ningún concepto semánticamente similar proveniente del corpus, pero sí conceptos que permiten definir al concepto en cuestión, como por ejemplo el concepto "Positiva" (referente a la carga eléctrica) para el cual se obtuvo "Negativa" (otro tipo de carga eléctrica) y "Núcleos" (los centros de un átomo cargados positivamente). En la Tabla 1 se muestran los resultados de forma detallada.

Resulta significativo que, de los 18 conceptos de entrada para los cuales el método aportó algún concepto sinónimo, en 9 de ellos (el 50%) se obtuvo un sinónimo como mejor propuesta del método; y para 3 de los 9 conceptos restantes, se obtuvo ubicado entre la segunda o tercera posición de los resultados. Sin embargo, es necesario señalar que en los casos en que un sinónimo no ocupa la primera posición en los resultados, la dispersión de los valores de similitud global proporcionados por el sinónimo de mayor valor de similitud y los conceptos en posiciones superiores no es despreciable, lo cual puede apreciarse mediante el análisis de la desviación estándar de las muestras con respecto a losvalores de similitud obtenidos.

Análisis de correlación

Como parte de la validación de la propuesta, se investigó la existencia de correlación entre los resultados del método, y la similitud semántica existente entre los conceptos. En el cálculo de la similitud semántica fue usada la propuesta de Wu y Palmer (Wu & Palmer, 1994), que se encuentra entre las métricas de este tipo con mejores resultados (Scriver, 2006). En el análisis de correlación se empleó el coeficiente de correlación de Pearson, asumiendo un nivel de significancia (α) de 0.01.

Las muestras iniciales estuvieron conformadas por los resultados de la ejecución del método para cada concepto de entrada. Para la mayoría de estas muestras, se observó que los valores de probabilidad superaron a α para la hipótesis H_0 (que plantea que no existe correlación), por lo que no fue posible comprobar la existencia de correlación alguna. Esto pudo deberse al reducido tamaño de las muestras, compuestas por 8.3 conceptos como promedio. Sin embargo, para la muestra conformada por el conjunto de los 174 conceptos obtenidos por el método en las diferentes ejecuciones, se obtuvo un valor de probabilidad de la hipótesis H_0 de 0.001, muy inferior al nivel de significancia seleccionado para estahipótesis, y un valor de correlación de 0.256. Por tanto puede afirmarse, con una probabilidad de error de 0.1%, que existe una correlación entre los resultados del método y la correspondiente similitud semántica, aunque esta correlación puede ser débil.

Análisis de resultados

De los 23 conceptos evaluados, 20 aportaron resultados significativos, para un 86.96% de cobertura. De ellos, 12 aportaron un resultado correcto entre las tres primeras propuestas, para un 60% de precisión. Se determinó estadísticamente que existe cierta correlación lineal entre los valores de similitud propuestos por el método, y los valores obtenidos empleando una métrica de similitud semántica.

Entre los factores que pueden haber incidido en los resultados se encuentran los siguientes: El corpus de MCs utilizado no es un corpus formalizado y de referencia para la realización de experimentos, por lo cual es posible que los mapas conceptuales que lo componen no esten representando el conocimiento adecuadamente; las métricas de similitud semántica alcanzan correlaciones de hasta 0.85 con respecto al criterio humano (Scriver, 2006), por lo que, al emplearlas como marco de referencia, es posible que un margen de error determinado se haya extendido al experimento realizado; otro problema puede ser el conocimiento no representado en WordNet; y por último, puede no existir suficiente información contextual en los mapas conceptuales recuperados del corpus, lo que está relacionado con la calidad y completitud del conocimiento representado en el corpus utilizado.

Conclusiones y trabajos futuros

En este trabajo se ha presentado un método que permite obtener, para un concepto determinado en un mapa conceptual, aquellos conceptos que sean similares en un corpus de mapas conceptuales. Esto se realiza por medio de la combinación de un algoritmo de comparación entre mapas conceptuales, para determinar similitudes entre contextos, y el algoritmo de correspondencias entre grafos *Similarity Flooding*, para determinar la correspondencia entre conceptos. A traves de la combinación de estas técnicas se logra una cierta correlación entre los resultados del método y los valores que puede aportar una métrica de similitud semántica, en lo cual pueden haber influido las características del corpus utilizado. Por esto es recomendable en trabajos futuros utilizar corpus con características diferentes para analizar el comportamiento del método propuesto ante estas variaciones; incorporar otros recursos como fuente de información adicional, como pudieran ser otras ontologías además de WordNet; y por último mejorar la validación de esta propuesta, por ejemplo a partir de su empleo integrando sistemas más complejos.

Bibliografía

Agirre, E., & Edmonds, P. (2006). Introduction. In E. Agirre & P. Edmonds (Eds.), Word Sense Disambiguation: Algorithms and Applications. Netherlands: Springer.

Agirre, E., & Martinez, D. (2001). Knowledge Sources for Word Sense Disambiguation. Text, Speech and Dialogue. Lecture Notes in Computer Science, 2166/2001, 1-10. doi: 10.1007/3-540-44805-5_1

Bentivogli, L., Forner, P., Magnini, B., & Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. Paper presented at the Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources, Geneva, Switzerland.

Cañas, A., Leake, D., & Maguitman, A. (2001). Combining concept mapping with cbr: Experience-based support for knowledge modeling. Paper presented at the Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference.

Cañas, A., Valerio, A., Lalinde, J., Carvalho, M., & Arguedas, M. (2003). Using WordNet for Word Sense Disambiguation to Support Concept Map Construction. LNCS 2857, Springer-Berlin 350-359.

Cañas, A. J., & Carvalho, M. (2004). Concept Maps and AI: an Unlikely Marriage? Paper presented at the Proceedings of Simpósio Brasileiro de Informática Educativa (SBIE 2004), Brasil.

Climent, S., Rodríguez, H., & Gonzalo, J. (2008). Definition of the links and subsets for nouns of the EuroWordNet project.

Costa, J. V., Rocha, F. E. L., & Favero, E. L. (2004). Linking phrases in Concept Maps: A Study on the Nature of Inclusivity. Paper presented at the Concept Maps: Theory, Methodology, Technology. Proceedings of the First International Conference on Concept Mapping, Pamplona, España de la Iglesia Campos, M. (2012). Métodos para el tratamiento de la ambigüedad semántica en mapas conceptuales. Máster en Informática Aplicada, Instituto Superior Politécnico José Antonio Echeverría, La Habana, Cuba.

Di Marco, A., & Navigli, R. (2013). Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. Computational Linguistics, MIT Press, 39(3), 709-754.

García, F., & Artiles, S. (2004). Simetría de la Técnica de Mapas Conceptuales y la Dimensión Informacional de la Gestión del Conocimiento en las Organizaciones: GECYT como Caso de Estudio Proceedings of the First International Conference on Concept Mapping 2004 (CMC'04). Pamplona, España.

Kleinberg, J. (1998). Authoritative Sources in a Hyperlinked Environment. Paper presented at the Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms.

Kowata, J. H., Cury, D., & Silva Boeres, M. C. (2010). A Review of Semi-Automatic Approaches to Build Concept Maps. Paper presented at the Concept Maps: Making Learning Meaningful. Proceedings of Fourth Int. Conference on Concept Mapping, Viña del Mar, Chile.

Leake, D. B., Maguitman, A., & Cañas, A. (2002). Assessing Conceptual Similarity to Support Concept Mapping. Paper presented at the FLAIRS Conference.

Magnini, B., & Cavaglia, G. (2000). Integrating Subject Field Codes into WordNet. Paper presented at the Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece.

Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2002). The Role of Domain Information in Word Sense Disambiguation. Natural Language Engineering, 359–373.

Marshall, B., Chen, H., & Madhusudan, T. (2006). Matching Knowledge Elements in Concept Maps using a Similarity Flooding Algorithm. Decision Support Systems, 42(3), 1290-1306.

Melnik, S., Garcia-Molina, H., & Rahm, E. (2002). Similarity flooding: a versatile graph matching algorithm and its application to schema matching. Paper presented at the Proceedings of the 18th International Conference on Data Engineering (ICDE '02), San Jose, Ca.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: an On-line Lexical Database. International Journal of Lexicography, 3 (4), 235-244.

Nasiruddin, M. (2013). A State of the Art of Word Sense Induction: A Way Towards Word Sense Disambiguation for Under-Resourced Languages. Paper presented at the TALN/ RECITAL, France.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. ACM Comput. Surv.(41, 2), 69.

Novak, J. D., & Cañas, A. J. (2006). The Theory Underlying Concept Maps and How to Construct Them.

Novak, J. D., & Gowin, D. B. (1984). Learning How to Learn. New York: Cambridge University Press.

Scriver, A. D. (2006). Semantic Distance in WordNet: A Simplified and Improved Measure of Semantic Relatedness. Master of Mathematics in Computer Science Thesis presented to the University of Waterloo in fulfilment of the thesis requirement for the degree of Master of Mathematics in Computer Science, University of Waterloo, Ontario, Canada.

Simón Cuevas, A. J. (2008). Herramientas para el perfeccionamiento de los Sistemas de Gestión de Conocimiento basados en Mapas Conceptuales. Tesis presentada en opción al grado científico de Doctor en Ciencias Técnicas, Instituto Superior Politécnico "José Antonio Echeverría", Ciudad de La Habana.

Simón Cuevas, A. J., Ceccaroni, L., Rosete, A., Suárez, A., & de la Iglesia Campos, M. (2008). A Concept Sense Disambiguation Algorithm for Concept Maps. Paper presented at the Concept Mapping - Connecting Educators, Proceedings of the 3rd International Conference on Concept Mapping (CMC 2008), Tallinn, Estonia; Helsinki, Finland.

WordNet: An Electronic Lexical Database. (1998). The MIT Press.

Wu, Z., & Palmer, M. (1994). Verb Semantics and Lexical Selection. Paper presented at the Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (ACL'94), Las Cruces, New Mexico.

Recibido: 16 de junio de 2014.
Aprobado en su forma definitiva:
28 de agosto de 2014

Manuel de la Iglesia Campos
Facultad de Ingeniería Informática, Instituto Superior Politécnico José Antonio Echeverría, Cujae, La Habana. CUBA.
Correo electrónico:
miglesia@ceis.cujae.edu.cu

Alfredo Simón Cuevas
Facultad de Ingeniería Informática, Instituto Superior Politécnico José Antonio Echeverría. Cujae, La Habana. CUBA.
Correoelectrónico:
asimon@ceis.cujae.edu.cu

José A. Olivas
Departamento de Tecnologías y Sistemas de Información, Escuela Superior de Informática, Universidad de Castilla La Mancha.
Correo-electrónico:
JoseAngel.Olivas@uclm.es