Ciencias de la Información Vol. 46, No. 1, enero-abril, pp. 63 - 71, 2015

Modelo de representación de textos basado en grafo para la minería de texto

Aramis Rodríguez-Blanco
Alfredo Simón-Cuevas
Ernesto Guevara-Martínez
Wenny Hojas-Mazo

La Minería de Texto constituye el proceso de descubrimiento de conocimiento, previamente desconocido y potencialmente útil, mediante la extracción automática de información desde diferentes recursos escritos. La estructuración del contenido textual en modelos de representación intermedia constituye un aspecto clave en este proceso. En el trabajo se propone un nuevo modelo de representación basado en grafos para la estructuración de contenidos textuales y un método para su construcción automática. El modelo está basado en la representación de frases conceptuales y las relaciones entre ellas, a partir de análisis de proximidad en el texto, para lo cual fueron definidas varias medidas de distancia. En el método propuesto se combinan técnicas de procesamiento de lenguaje natural, con patrones léxicos y recursos de conocimiento para extraer los conceptos, y fueron definidos dos métodos para identificar las relaciones: distancia más cercana y ventana contextual. Se concibió en tres fases fundamentales: pre-procesamiento, extracción de información, y refinado, y se evaluó experimentalmente con noticias de una colección de referencia. Los experimentos se orientaron a evaluar la cantidad de información contenida en los grafos resultantes, así como la precisión en la extracción automática de conceptos, en los cuales se obtuvieron resultados prometedores.

Palabras clave: modelos de representación de texto; minería de texto; extracción de información

RESUMEN

ABSTRACT

Text Mining constitutes the process of knowledge discovery, previously unknown and potentially useful, though the automatic information extraction from texts. The structuring of textual content in an intermediate representation models constitutes a key aspect in this process. The more employees' models are based on a list of significant terms, such as the vector space model, although recently the use of relational models in graph form has been increased. In the work a new graph-based representation model for the structuring of textual contents and a method for it automatic construction are proposed. The model is based on the representation of conceptual sentences and the relationships among them, through proximity analysis in the text, where several distance metrics were defined. It was conceived in three fundamental phases: pre-processing, information extraction, and a refine process, and it was experimentally evaluated with news of a reference collection. The experiments were guided to evaluate the quantity of the information contained in resulting graphs, as well as the precision in the automatic extraction of concepts, in which promising results were obtained.

Keywords: text representation model; tex tmining, information extraction

Introducción

En la actualidad, es significativa la cantidad de información que se genera y es almacenada en formato digital, sobre todo en el ámbito de las redes sociales, Internet y en las organizaciones en general. Se estima que alrededor del 80% de esa información está incluida en textos en lenguaje natural y en forma no estructurada (Gupta&Lehal, 2009; Sumathy &Chidambaram, 2013; Ramanathan &Meyyappan, 2013; Kumar et al., 2014). Aunque se ha trabajado en el desarrollo de sistemas de recuperación de información textual cada vez más sofisticados, el incremento del valor de esa información en los procesos de gestión de conocimiento y toma de decisiones en las organizaciones, está demandado de manera creciente el desarrollo de estas propuesta dirigidas al descubrimiento de conocimiento sobre ese tipo de información. En este escenario, la Minería de Texto surge como un nuevo campo multidiciplinario (Gupta, & Lehal, 2009; Singh, &Raghuvanshi, 2012; Ramanathan, & Meyyappan, 2013), Sumathy, &Chidambaram, 2013; desde donde se aportan soluciones para esta problemática.

La Minería de Texto (MT) se define como el proceso de descubrir automáticamente conocimiento, no trivial, previamente desconocido, y potencialmente útil en colecciones de textos no estructurados, a través de la extracción automática de información y la identificación de patrones ocultos e interesantes (Feldman&Sanger, 2007; Gupta&Lehal, 2009; Sumathy, &Chidambaram, 2013; Ramanathan& Meyyappan, 2013; Kumar et al., 2014). En este campo se abordan problemas tales como: la categorización y agrupamiento de textos, la generación automática de resumen, y la extracción de características (o patrones) de textos (Singh &Raghuvanshi, 2012; Ramanathan&Meyyappan, 2013), en cuya solución la extracción y análisis de conexiones entre conceptos del texto son aspectos valioso (Gupta&Lehal, 2009). En todo proceso de MT, la estructuración del contenido de los textos, un modelo de representación intermedia constituye un paso fundamental (Ramanathan& Meyyappan, 2013), de ello dependen los algoritmos y métodos de descubrimiento de conocimiento que se apliquen. Con este propósito, han sido empleados el Modelo de Espacio Vectorial, taxonomías de conceptos y grafos, de los cuales el primero es el más comúnmente usado (Shekhar et al. 2014; Chang & Kim, 2014), aunque los modelos basados en grafos están teniendo una gran demanda. Algunos de estos modelos que son empleados en tareas de MT son: grafos conceptuales (Montes et al., 2001; Thavamani & Rengarajan, 2014), redes semánticas (Shekhar et al., 2014), y mapas conceptuales (Palmeira et al., 2012).

En el presente trabajo se ha abordado la problemática referente a la representación de contenidos textuales en procesos de MT. Según se interpreta de lo reportado en (Ordoñes, &Gelbukh, 2010; Aggarwal & Zhai, 2012; Singh&Raghuvanshi, 2012), la búsqueda de nuevas alternativas para la representación de textos, que permitan no solo la inclusión de la semántica propia del lenguaje, y que esta representación sea exacta, sino que faciliten operaciones que conduzcan al descubrimiento de conocimiento, representa aún tema abierto a la investigación. En este sentido, se define una primera aproximación de un nuevo modelo de representación de textos basado en grafo, así como un método para su construcción de forma automática.

El nuevo modelo está inspirado en los fundamentos de los mapas conceptuales (Novak & Cañas, 2006) y el grafo de asociación (Medina et al., 2005) y se basa en la representación de frases conceptuales (formados por una o varias palabras) y las relaciones entre ellos según su proximidad en el texto, como rasgos representativos del contenido textual. La representación de conceptos, en lugar de términos simples, ofrece la posibilidad de capturar mayor información semántica, además de reducir la ambigüedad y llevar a cabo procesos discriminatorios (respecto a la relevancia) más efectivos (Zhong, &Wu, 2012). El análisis de la proximidad entre conceptos en un texto para la identificar y representar relaciones entre ellos, tomado de (Medina et al., 2005),posibilita descubrir patrones interesantes sobre la base del análisis de vínculos contextuales en el contenido. Este análisis es soportado por métricas definidas para el cálculo de la distancia entre conceptos en un texto.

El método propuesto se concibió en tres fases:

- pre-procesamiento,
- extracción de información, y
- refinado,

según lo reportado en (Rodríguez&Simón, 2013). En la primera se aplican técnicas de procesamiento de lenguaje natural, soportado en el analizador sintáctico Freeling, para capturar la información sintáctica de las oraciones del texto y los elementos que la componen. En la segunda se combinan patrones léxicos definidos, con recursos de conocimiento para identificar conceptos, y se definen dos criterios referentes a la evaluación de la proximidad (distancia más cercana y ventana contextual), así como reglas las asociadas a cada uno de ellos para identificar los tipos de relaciones entre los conceptos. En la tercera y final son eliminados los conceptos que tengan menor relevancia, en cuanto a su vínculo contextual con el resto de los conceptos, a partir de la evaluación del peso de asociatividad acumulada; la métrica definida con este propósito. Se realizaron experimentos con noticias de la colección de la Agencia de Prensa EFE, en ellos se evaluó la cantidad de información que se incluyó en los grafos generados, siendo un objetivo reducir su complejidad y mantener una representación de la información relevante. También se evaluó la precisión en la identificación de conceptos por parte del método y se ejemplificó su aplicación sobre una de las noticias incluidas en el corpus de prueba.

El trabajo se organiza en lo adelante en cinco secciones. La Sección 2 aborda fundamentos de la MT y modelos de representación de textos. En la Sección 3 se formaliza el modelo propuesto, y un método para su construcción automática es descrito en la siguiente sección. En la Sección 5 se describen los experimentos realizados, y se resumen los análisis sobre los resultados. Finalmente se exponen conclusiones en la Sección 6.

Materiales y métodos

La Minería de Texto (MT) se define como el proceso de descubrir automáticamente conocimiento, no trivial, previamente desconocido, y potencialmente útil en colecciones textos no estructurados, a través de la extracción automática de información e identificación de patrones ocultos e interesantes (Feldman &Sanger, 2007; Gupta&Lehal, 2009; Sumathy, &Chidambaram, 2013; Ramanathan

&Meyyappan, 2013;Kumar et al., 2014). La MT o también conocida como: Análisis Inteligente de Textos, Minería de Datos Textuales, y más comúnmente como Descubrimiento de Conocimiento en Textos (KDT, por sus siglas en inglés) (Ramanathan &Meyyappan, 2013). Constituye un campo multidisciplinario en el que se aplican técnicas y métodos provenientes de la minería de datos, estadística, extracción de información, lingüística computacional, y procesamiento de lenguaje natural(Gupta, &Lehal, 2009; Singh, &Raghuvanshi, 2012; Sumathy, &Chidambaram, 2013; Ramanathan, &Meyyappan, 2013). En la MT se abordan problemas tales como: la categorización (clasificación supervisada) y agrupamiento de textos (clasificación no supervisada), la generación automática de resumen, y la extracción de características (o patrones) de textos (Singh &Raghuvanshi, 2012; Ramanathan&Meyyappan, 2013). En la solución de muchos de estos problemas, la extracción y análisis de conexiones entre conceptos del texto constituyen aspectos valioso (Gupta&Lehal, 2009).

La forma de estructuración del contenido de los textos en un modelo de representación intermedia, sobre el cual se puedan aplicar los procesos automáticos de análisis y descubrimiento, constituye uno de los pasos principales en el proceso de MT (Ramanathan &Meyyappan, 2013). Estos modelos de representación de forma general, pueden estar basados en palabras (bolsa de palabras o n-gramas), conceptos (jerarquía de conceptos o grafos semánticos), párrafos (ej. lista de párrafos o n-frases), entre otros (Kumar et al., 2014). El Modelo de Espacio Vectorial (modelo basado en palabras) es el más comúnmente empleado (Shekhar et al. 2014; Chang & Kim, 2014), sin embargo, en este modelo no se consideran las relaciones semánticas establecidas entre palabras, frases y párrafos, que apoyan al significado, la coherencia y unidad del discurso en el texto. En función de poder representar semánticamente la información textual para realizar procesos de minería más significativos (Aggarwal &Zhai, 2012), surge el uso de taxonomías (Basole et al., 2013) y con gran auge en la actualidad, el de los grafos (Chang & Kim, 2014). Según Chang y Kim, los modelos basados en grafos se caracterizan por cómo son representados los nodos y los arcos (Chang & King, 2013). Los nodos pueden representar palabras, sentencias, párrafos y el propio texto, y también conceptos, como componente semántico. Estos pueden ser homogéneos u heterogéneos, según si representan uno o más de dos componentes, y pesados o no pesados, si tienen algún valor de peso asociado (Chang & King, 2013). Los arcos pueden ser dirigidos y no dirigidos, si indican orden o no, etiquetados y no etiquetados, y pesados o no pesados, si tienen un valor de peso asociado (ej. frecuencia en la que están presente en el mismo momento los conceptos conectados) (Chang & King, 2013).Algunos de los modelos basados en grafos empleados en tareas de MT son los grafos conceptuales (Montes et al., 2001; Thavamani& Rengarajan, 2014), redes semánticas (Shekhar et al., 2014), y mapas conceptuales (Palmeira et al., 2012). En los grafos conceptuales se representan conceptos y relaciones conceptuales como nodos. Los conceptos tienen un tipo (clase de concepto) y un referente (la instancia de este tipo de objeto), y cada relación tiene uno o más (usualmente dos) arcos enlazados a un concepto (Ordoñes,& Gelbukh, 2010). En las redes semántica que se proponen en (Shekhar et al., 2014) los nodos representan sustantivos y las relaciones son identificadas a partir de relaciones de sinonimia, hiperonimia/piponimia y meronimia/holonimia que se hayan en WordNet. En (Palmeira et al., 2012), los mapas son construidos sobre la base de interconectar sustantivos y adjetivos por medio de relaciones etiquetadas con verbos.

En el trabajo se propone un nuevo modelo basado en grafo, como otra alternativa de representación con similitudes a los descritos anteriormente, en cuanto a la representación conceptos. Sin embargo, la nueva propuesta se distingue del resto, específicamente de los grafos conceptuales y mapas conceptuales, en cuanto a la identificación y representación de las relaciones, ya que en el nuevo modelo el rol de las relaciones es expresar fortalezas en cuanto a vínculos contextuales entre los conceptos según su proximidad. Aunque en los grafos conceptuales y mapas conceptuales las relaciones también indican vínculos contextuales, estos generalmente se identifican en el contexto de la oración, y en el modelo propuesto está concebido un contexto más amplio, haciendo posible vincular conceptos que estén en oraciones contiguas. También se aporta otro tipo de información para el análisis en el etiquetado de las relaciones.

Resultados y discusión

Modelo de Representación Propuesto

El modelo propuesto se ha definido sobre la base de los fundamentos de los mapas conceptuales (Novak &Cañas, 2006) y el grafo de asociación (Medina et al., 2005); este último usado en tareas de recuperación de información. Los mapas conceptuales son reconocidos para representar contenido textual en (Ordoñes&Gelbukh, 2010; Palmeira et al., 2012; Rodríguez & Simón, 2013). El modelo constituye una nueva alternativa de representación de textos basada en grafo, donde se representan conceptos (formados por una o varias palabras), en lugar de palabras simples, y relaciones entre ellos, como rasgos representativos del contenido textual. Los conceptos se definen como regularidades o patrones en eventos u objetos, o registros de estos, designados por una etiqueta (Novak & Cañas, 2006), también representan hechos, procesos, y situaciones, contenidos de las diferentes unidades semánticas que conforman el texto. Las frases conceptuales capturan mayor información semántica, son menos ambiguas y resultan ser más discriminativas (respecto a la relevancia) que los términos simples (Zhong, &Wu, 2012). Los conceptos se relacion según su proximidad en el texto, tomando en consideración lo reportadoen(Medina et al., 2005). Esto posibilita crear las bases para el descubrimiento de aquellos patrones interesantes en los textos, sobre la base del análisis de los vínculos contextuales entre los conceptos, lo cual se puede llevar a cabo aplicando algoritmos de minería de grafos (Aggarwal, & Wang, 2010), o través de mecanismos de consulta como el reportado en (Simón et al., 2008).

En el análisis de la proximidad entre los conceptos en el texto se asume que las oraciones, párrafos y secciones conforman unidades semánticas. Si dos unidades semánticas contienen las mismas frases conceptuales se presume que transmiten la misma idea, por lo que la semántica que se relaciona con estas unidades estructurales del texto se vincula directamente con los conceptos que la componen. Las relaciones entre los conceptos también dependen de las estructuras gramaticales formadas en el texto. Por tanto, para modelarlas se utiliza la unidad semántica más específica

que se forma a nivel de oración y párrafo, ya que determinan básicamente la distancia física entre los conceptos. Mientras más pequeña sea la distancia física entre dos conceptos, más fuerte se crea el vínculo contextual y semántico.

De acuerdo a lo anterior, cada documento es modelado por un grafo. Los vértices del grafo representan conceptos significativos, y reconocidos como homogéneos y no pesados; según (Chang, & Kim, 2013). Las artistas representan las relaciones entre los conceptos, son dirigidas y etiquetadas, y forman proposiciones (Novak & Cañas, 2006). Las etiquetas en las aristas expresan cuán fuerte es el vínculo contextual entre los conceptos según su proximidad. Matemáticamente, el modelo se define por un par ordenado *(C, Pr)*, siendo *C* un conjunto de conceptos definidos como $c^m = w_1,\ldots,w_d | w_d \in$ *palabras*, y *Pr* (siendo, $C \cap Pr = \emptyset$) un conjunto de proposiciones. Ψ_{MC} es la función de incidencia que asocia a cada proposición, un par ordenado de conceptos y una relación etiquetada con una frase de enlace. Si *p* es una proposición y $\Psi_{MC}(p)=(c_o;c_d;\ fe)$, entonces *p* conecta los conceptos c_o y c_d a través de la frase de enlace *fe*; también se puede decir que c_o domina a c_d. El concepto c_o constituye el origen de la relación y c_d el destino, *fe* es una etiqueta que define el tipo de relación que ese establece entre ambos conceptos. Se definieron tres tipos de relaciones para representar el nivel de vínculo contextual:

- *fuerte relación* (FR),
- *débil relación* (DR) y
- *poca relación* (PR),

por tanto, $fe = e / e \in \{FR, DR, PR\}$, cada uno de los cuales es definido según la distancia calculada entre los conceptos en el texto, considerando la distancia mínima en caso de que se obtengan varios valores de distancia entre dos conceptos.

La distancia entre dos conceptos puede ser calculada, tomando como referencia la distancia entre las unidades semánticas en el texto donde están presentes, tal es el caso de la oración y el párrafo (Medina et al., 2005), pero también a partir de la cantidad de palabras que los separa. Considerando la distancia por unidades semánticas, si se define (p_r,n_r) y (p_s,n_s) como el número del párrafo y el número de la oración en que se encuentran los término s*r* y *s*, respectivamente, la *distancia* (*D*) puede ser calculada mediante la Ecuación 1. Considerando la distancia por la cantidad de palabras, cada concepto posee una localización *l* dentro del texto, definida como la tripleta *(p, o, po)*, donde *p* es el número del párrafo, *o* es el número de la oración respecto al párrafo *p*, y *po* es la posición del concepto *c* en la oración, y la distancia (*D*) puede ser calculada mediante la Ecuación 2. Siendo $D_{min}(r, s)$ una función que determina la distancia mínima entre dos conceptos c_r y c_s, dado la distancia física calculada entre sus localizaciones usando la Ecuación 3.

$$D_{s,r} = \begin{cases} 0 & si\ (s = r) OR\ (o_{ps} = o_{pr}) \\ 1 & si\ p_s = p_r \\ |p_s - p_r| + 1 & otro\ caso \end{cases} \quad (1)$$

$$D_{s,r} = \begin{cases} 0 & si\ (s = r)\ OR\ (o_{ps} = o_{pr}) \\ D_{\min(s,r)} & otro\ caso \end{cases} \quad (2)$$

Sean, c_s y c_r dos conceptos, $l_s = (p_s, o_{ps}, po_s)$ y $l_r = (p_r, o_{pr}, po_r)$, sus respectivas localizaciones, tal que p_s *e» * p_r, *TO(o)* una función que calcula el tamaño (cantidad de palabras) de una oración *o* y *TP(p)* una función calcula el tamaño (cantidad palabras) de un de un párrafo *p*. La distancia física entre dos conceptos es calculada mediantes la ecuación:

$$D(l_s, l_r) = (T(o_{ps}) - po_s) + \sum_{i=o_{ps}+1}^{TO(p_p)} TO(o_{pi}) + \sum_{j=p_s}^{p_r} TP(p_j) + \sum_{k=1}^{o_{pr}} TO(o_{pk}) + po_r \quad (3)$$

Método para la Construcción Automática del Modelo de.....

El método que se propone constituye una adaptación del reportado en (Rodríguez & Simón, 2013), en cuanto a los procesos de extracción de información, así como del proceso de refinado, cada uno de los cuales están orientados a la construcción del modelo de representación propuesto. En el nuevo método se incorpora la posibilidad de utilizar ontologías(Studer, Benjamins, &Fensel, 1998), como recurso externo de conocimiento para apoyar la identificación de conceptos. En las ontologías se formaliza la conceptualización de un dominio que puede ser general o específico, a través de clases, propiedades e instancias, y representa un conocimiento aceptado por una comunidad (Studer, Benjamins, &Fensel, 1998). Por tanto, expresan un conocimiento bien formado y que puede servir de referencia para la identificación de conceptos en el texto, a partir de las clases e instancias representadas. El método está definidoentres fases fundamentales:

- *pre-procesamiento*,
- *extracción de información y*
- *refinado*;

según se muestra en la Fig. 1.

Pre-Procesamiento del Texto

En esta fase se sigue básicamente el mismo flujo de trabajo que en (Rodríguez & Simón, 2013). Se inicia con la extracción del texto plano desde ficheros en varios formatos. El texto plano es segmentado en párrafos y oraciones, teniendo en cuenta criterios, como la segmentación por punto final, y no se consideran los puntos incluidos en las siglas y en los números con cifras decimales. En un siguiente paso, cada oración es fragmentada en el conjunto de *tokens* que la componen (ej. palabras, números, signos de puntuación, etc.), los cuales se obtienen mediante un algoritmo propio basado en la identificación de fronteras, según diferentes clasificaciones de tokens (Abney, 1991).

Luego se lleva a cabo el proceso de análisis morfo-sintáctico sobre cada oración, en el que se incluye el análisis sintáctico superficial y de dependencias (o profundo); usando el analizador sintáctico Freeling. En el análisis morfológico se etiqueta cada token con su raíz morfológica, su categoría gramatical, género, número, persona, y tiempo. El análisis sintáctico superficial se realiza sobre los tokens etiquetados de cada oración, y básicamente consiste en agrupar a varios tokens en chunks o constituyentes (Abney, 1991). Los chunks representan categorías gramaticales entre los que se encuentran los sintagmas nominales, sintagmas preposicionales y grupos verbales (Abney, 1991). Este proceso concluye con la representación de esta información es representada en un árbol sintáctico, es cual es utilizado fundamentalmente para identificar las frases conceptuales en el texto. El análisis de dependencias concluye con el árbol de dependencias, en el cual se representan relaciones de dependencia entre los diferentes elementos que componen la oración (etiquetados según su función gramatical), por ejemplo, el sujeto y los complementos de la oración constituyen subárboles. Este proceso está dirigido a la construcción de conceptos más complejos a partir de dependencias identificadas entre

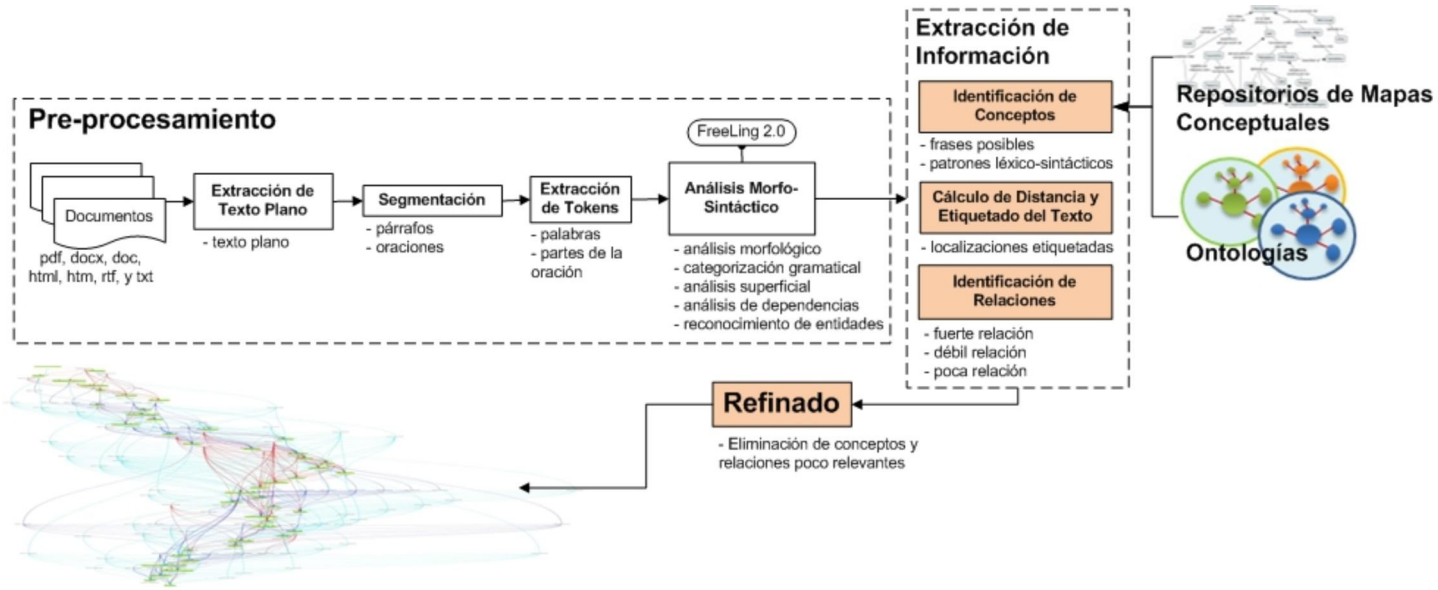

Figura 1. Método para la construcción automática del modelo de representación conceptual propuesto.

conceptos simples (formados por una sola palabra).

Extracción de Información

La identificación de conceptos se basa en identificar aquellas frases (formadas por una o varias palabras)que puedan tener un sentido conceptual. En este proceso se tiene en cuenta la información resultante del análisis sintáctico,fundamentalmente la representada en el árbol sintáctico, así como la conceptualización representada en los recursos de conocimiento definidos. El árbol sintáctico se analiza a partir de un conjunto de patrones léxicos, definidos a partir de un conjunto de categorías gramaticales que son relevantes para la identificación de conceptos, entre las que se encuentran: frases sustantivas y aquellas que indican entidades nombradas, como fechas y nombres propios. Los patrones expresan combinaciones de elementos del texto etiquetados con estas y otras categorías. Se han extendido los patrones, respecto a lo reportado en (Rodríguez & Simón, 2013), y una selección de ellos se muestra en la Tabla 1.

Una vez identificado cada concepto, se etiqueta con la localización *l* que ocupa dentro del texto mediante la tripleta *(p, o, po)*. Cada concepto puede tener diferentes localizaciones, según su aparición en el texto, y la cantidad de localizaciones indica su frecuencia *f* de aparición, la cual también es registrada. A partir de las localizaciones, se lleva a cabo un proceso de cálculo de las distancias entre cada par de conceptos, usando la Ecuación 1 o 2, según considere el usuario. En esta etapa de la investigación se ha considerado conveniente estudiar el efecto de ambas medidas para el cálculo de la distancia entre conceptos, ya que de ellas depende la identificación de relaciones y sus tipos, así como la cantidad de información a incluir en el grafo resultante, su relevancia yla complejidad en su procesamiento computacional.

La identificación de las relaciones entre conceptos está concebida según dos criterios: distancia más cercana y ventana contextual, donde para cada uno de los cuales se emplea una métrica de cálculo de distancia específica, así como un conjunto de reglas del tipo Si-Entonces que interpretan

Tabla 1. Patrones léxicos para identificar *frases conceptuales* en textos en idioma español

Categorías	Patrones	Ejemplos	Patrones	Ejemplos
'sn"	[D | P | Z]+[<s-adj>]+NC	la gran avenida	W	5 de enero
sintagma nominal)	D | P | Z]+NC+[<s-adj>]	ropaelegante	VMN	aprender
	[D]+NP	mapa conceptual	P	ellos
	[Z]+NC	30 estudiantes		
	Z	10 millones		
's-adj" sintagma adjetivo)	([R]+[A | VBN])	muylejos / mal herido	<s-adj>+Fc+<s-adj>	pretty, nice and kind
	<s-adj>+(C | Fc)+<s-adj>	joven y fuerte		
'sadv" sintagma adverbial)	R	principalmente		

Leyenda: D: determinante; NC: sustantivo común; NP: sustantivo propio; Z: número o numeral; W: fecha; VBG: verbo gerundio; VMN: verbo nfinitivo; VBN: verbo participio; P: pronombre (ej. personal, demostrativo); C: conjunción; R: adverbio; +: concatenación de elementos; \|: lisyunción de elementos; <>: categoría sintáctica; []: elemento opcional; (): agrupación de elementos

el valor de distancia obtenido. Entre estos dos criterios radica como diferencia fundamental la cantidad de relaciones que son extraídas del texto, siendo esto en el primero muy superior, lo cual puede afectar el procesamiento computacional del grafo resultante. Para la identificación de relaciones según el criterio *distancia más cercana*, se emplea la Ecuación (1) y son definidas las siguientes reglas:

Sean c_a y c_b dos conceptos identificados;

R1: Si $D_{a,b} = 0$, entonces el vínculo contextual entre c_a y c_b es fuerte, y se etiqueta con 'FR';

R2: Si $D_{a,b} = 1$, entonces el vínculo contextual entre c_a y c_b es débil, y se etiqueta con 'DR';

R3:Si $(D_{a,b} \neq 0) \wedge (D_{a,b} \neq 1) \wedge (D_{a,b} \leq$ *umbral), entonces existe poco vínculo contextual entre* c_a *y* c_b*, y se etiqueta con 'PR'; el valor de umbral es definido por el usuario*.

Para la identificación de relaciones a partir del criterio*ventana contextual*, se emplea la Ecuación (2) y se define la *venta contextualVC(c)* de un concepto *c* como el conjunto de palabras que se encuentran en la vecindad (contexto) de *c*, siendo *c* el centroide, y cuyo tamaño es definido empíricamente por la cantidad media de palabras que incluyen los párrafos del texto. Considerando estas bases, son definidas las siguientes reglas:

Sean c_a y c_b dos conceptos, y w un coeficiente que define un límitede vecindad dentro de la ventana contextual donde las relacionesentre conceptos resulta ser débil (inicialmente se define empíricamente con valor 0,5);

R1: Si $D_{a,b} = 0$, entonces el vínculo contextual entre c_a y c_b es fuerte, y se etiqueta con 'FR'.

R2: Si $(D_{a,b} \neq 0) \wedge (D_{a,b} \leq w*VC(a))$, entonces el vínculo contextual entre c_a y c_b es débil, y se etiqueta con 'DR'.

R3: Si $(D_{a,b} \neq 0) \wedge (D_{a,b} > w*VC(a)) \wedge (D_{a,b} \leq$ *VC(a)), entonces existe pocovínculo contextual entre* c_a *y* c_b*, y se etiqueta con 'PR'. el valor de umbral es definido por el usuario*.

Para la identificación de relaciones a partir del criterio ventana contextual, se emplea la Ecuación (2) y se define la venta contextual *VC(c)* de un concepto *c* como el conjunto de palabras que se encuentran en la vecindad (contexto) de *c*, siendo *c* el centroide, y cuyo tamaño es definido empíricamente por la cantidad media de palabras que incluyen los párrafos del texto. Considerando estas bases, son definidas las siguientes reglas:

Sean c_a y c_b dos conceptos, y w un coeficiente que define un límite de vecindad dentro de la ventana contextual donde las relaciones entre conceptos resulta ser débil (inicialmente se define empíricamente con valor 0,5);

R1: Si $D_{a,b} = 0$, entonces el vínculo contextual entre c_a y c_b es fuerte, y se etiqueta con 'FR'.

R2: Si $(D_{a,b} \neq 0) \wedge (D_{a,b} \leq w*VC(a))$, entonces el vínculo contextual entre c_a y c_b es débil, y se etiqueta con 'DR'.

R3: Si $(D_{a,b} \neq 0) \wedge (D_{a,b} > w*VC(a)) \wedge (D_{a,b} \leq$ *VC(a)), entonces existe pocovínculo contextual entre* c_a *y* c_b*, y se etiqueta con 'PR'*.

Proceso de Refinado

Este es un proceso tiene como objetivo fundamental reducir el tamaño del grafo resultante de la fase de extracción de información, a partir de la definición de una estrategia de reducción de conceptos dirigida a representar solo los conceptos más significativos y sus relaciones. En este sentido, se estudiaron alternativas basadas en la frecuencia de aparición y en el análisis de los vínculos contextuales de los conceptos, a saber:

1. eliminar conceptos menos frecuentes, lo cual tiene el inconveniente de que generalmente la mayoría de los conceptos tienen bajas frecuencia y su eliminación podría provocar la pérdida de información valiosa;

2. eliminar los conceptos que tengan una cantidad de relaciones inferior a la media de las relaciones presentes en los conceptos del grafo, sin embargo, al no considerarse el tipo de relación también podría provocar la pérdida de información relevante;

3. eliminar aquellos conceptos que tengan menor relevancia en cuanto a su vínculo contextual con el resto de los conceptos en el grafo, teniendo en cuenta la cantidad de relaciones, así como los tipos.

La última variante fue la seleccionada para la reducción de los conceptos menos significativos. La relevancia el vínculo contextual de un concepto c_i representado en el grafo, es medida a través de un peso de asociatividad acumulada *(PAA)*, el cual es calculado a través de *PAA(c) = 0,75*CFR(c) + 0,50*CDR(c) + 0,25*CPR(c)*; siendo *CFR(c)* la cantidad de relaciones *FR*, *CDR(c)* la cantidad de relaciones *DR*, y *CFR(c)* la cantidad de relaciones *PR*. En esta expresión, las relaciones de entrada y salida tienen en el mismo peso y los coeficientes q\ue ponderan cada tipo de relación son definidos según el nivel de importancia que tiene cada tipo de relación para el análisis contextual. Luego de calcular el *PAA* para cada concepto extraído en la fase anterior, son representados en el grafo aquellos con mayor valor, de acuerdo a un por ciento que defina el usuario según su interés. También se implementaron filtros paraque el usuario seleccione que tipo de relaciones incluir en el grafo resultante, para facilitar su análisis.

Resultados Experimentales

Los experimentos fueron realizados sobre un corpus conformado por textos (T) de 10 noticias, seleccionadas de la colección de la Agencia de Prensa EFE del año 1994 (referencia para la evaluación de sistemas de recuperación de información en el marco de CLEF[1]). En la Tabla 2 se resume una caracterización de los textos, donde

Tabla 2. Caracterización del corpus de prueba.

Características	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Cantidad de Párrafos	4	2	8	15	6	3	14	14	17	4
Promedio de Palabras por Párrafo	33	45	51	49	34	31	47	43	36	46

[1] *Siglas correspondientes al inglés : Cross Language Evaluation Forum*.

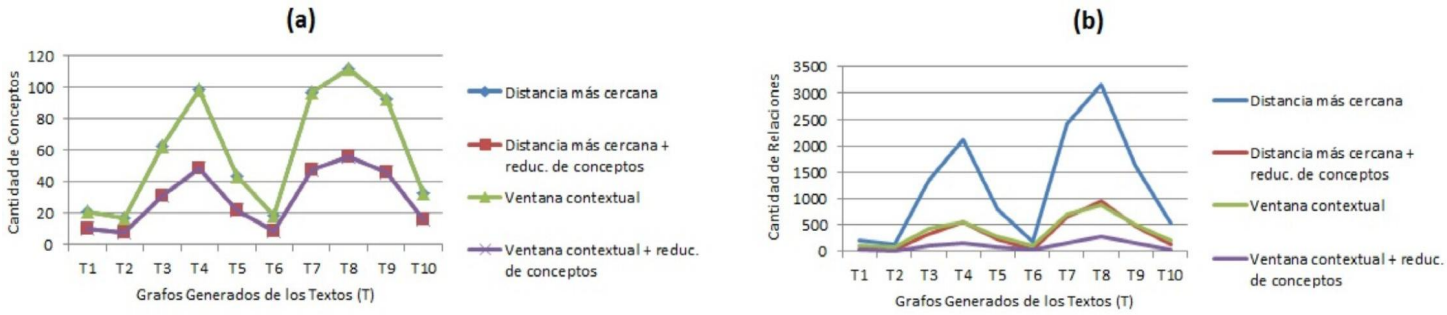

Figura 2. Método para la construcción automática del modelo de representación conceptual propuesto.

se puede apreciar la diversidad en cuanto los tamaños de los mismos.

Se realizaron varios experimentos dirigidos basicamente a la evaluación de la cantidad de información contenida en los grafos generados, según las diferentes variantes de identificación de relaciones contextuales descritas anteriormente. La cantidad de información es medida por la cantidad de conceptos y relaciones representadas en el grafo e influye en la complejidad del grafo, lo cual es un aspecto importante a tener en cuenta para el procesamiento computacional posterior a realizar para descubrimiento de conocimiento. Constituye un reto el lograr representar la información más relevante del texto, sin incrementar notablemente la complejidad del grafo. Fueron evaluados los resultando del método usando el criterio de distancia más cercana y ventana contextual, con y sin el mecanismo de reducción de los conceptos definido en el proceso de refinado, y los resultados se muestran en las Fig. 2a y 2b, respectivamente.

En los resultado mostrados en la Fig. 2a, evidencian los beneficios de la estrategia de reducción de conceptos definida en el proceso de refinado, ya que se logra reducir la cantidad de conceptos en un 50 % aprox., sin perder la información relevante (los conceptos representados son los de mayor vínculo contextual).En la evaluación del criterio de distancia más cercana, se utilizó umbral=*3*, esto significa que las relaciones débiles solo se identifican entre conceptos que se encuentran en párrafos contiguos.

Según se muestra en la Fig. 2b, un valor superior de umbral podría provocar un incremento notable de la complejidad del grafo resultante. Este incremento sin embargo, no repercute en el valor de la información a representar, ya que las relaciones entre los conceptos que se encuentran a una distancia mayor de 3 se pueden considerar irrelevantes o que aportan poca información. En el caso del criterio de ventana contextual, los experimentos se realizaron con *w = 0,5*, lo que significa que se identificaría fuerte relación entre dos conceptos cuando estos estén a una distancia física inferior o igual a la mitad de la cantidad media de palabras por párrafo del documento. Se aprecia la reducción de la complejidad del grafo al usar el criterio de ventana contextual, en lugar de la distancia más cercana, según la Fig. 2b, ya que se reducen en un70 % aprox. de cantidad de relaciones, y al mismo se representan los vínculos contextuales más significativos. En la Tabla 3 se resumen otros indicadores que fueron obtenidos de los experimentos, al observarse la disminución promedio de la complejidad de los grafos usando ventana contextual.

Adicionalmente, se comprobó la calidad del proceso de extracción de conceptos, al evaluar cuán correctos eran los conceptos representados en los grafos generados. Para esto fue definida una métrica de precisión *(P)*, según la Ecuación 4, en la que se tuvieron en cuenta:

- los conceptos correctos *(CC)*,
- aceptados *(CA)*, e
- incorrectos *(CI)*.

$$P = \frac{AC}{AC + AA + AI} \quad (4)$$

Son considerados conceptos aceptados, aquellos formados por más de una palabra, y que una de ellas no debe formar parte de dicho concepto. Los conceptos incorrectos son aquellas palabras o frases que no tienen un significado conceptual. Luego de obtenida la precisión para cada grafo generado, estos fueron clasificados según la distribución que se muestra en la Tabla 4. Los resultados obtenidos se muestran en la Fig. 4, donde se observa que todos los grafos han sido evaluados entre Bien y Muy Bien, lo cual indica que la identificación de conceptos tiene una alta precisión. Se obtuvo como promedio una precisión del 80 %, para constatar la efectividad del método propuesto en este sentido.

Tabla 4. Clasificación de grafos generados según su precisión

Evaluación	Rango de Valores de Precisión
Muy bien	90-100
Bien	70-90
Regular	60-70
Mal	0-60

A modo de ejemplo, se utilizó la noticia: 'Guinea-Obiang Presidente Sugiere Rechazará Ayuda Exterior Condicionada', y el grafo generado automáticamente se muestra en la Fig. 6. En la visualización del grafo cada tipo de relación se distingue con un color diferente, siendo de color

Tabla 3. Resumen de los resultados de los experimentos.

Métricas	Distancia más Cercana	Distancia más Cercana + Reduc. Conceptos	Ventana Contextual	Ventana Contextual + Reduc. Conceptos
AvgC	59,8	29,5	59,8	29,5
AvgR	1251	344,6	382,6	105,8

AvgC: promedio de conceptos extraídos por documento; AvgR: promedio de relaciones extraídas por documento.

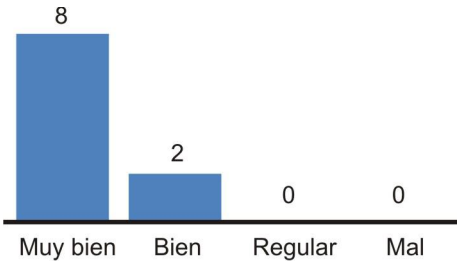

Figura 3. Resultado de la clasificación de los grafos generados

rojo las FR, azul claro las PR y azul oscuro las DR. Según se puede interpretar del grafo, la noticia trata sobre convulsiones políticas por la que atraviesa el presidente Obiang, debido a desórdenes sociales del pueblo, provocado por la ayuda extranjera. Esto se obtubo a partir de los fuertes vínculos contextuales identificados entre los conceptos: 'Obiang','convulsiones políticas', 'desórdenes sociales', 'pueblo guineano', y 'ayuda extranjera'. El grafo construido puede ser refinamiento usando la herramienta CmapTools[2].

Conclusiones y trabajo futuro

Como resultado del presente trabajo se describió y formalizó un nuevo modelo basado en grafo para la representación de textos a ser aplicado en procesos de MT, a través del cual se pueden representar los conceptos y las relaciones entre ellos según sea su proximidad dentro del texto. De igual manera se presentó un método para su construcción automática a partir de textos en diferentes formatos. El tratamiento de la proximidad, como criterio para representar relaciones entre conceptos, así como el etiquetado de estas, crea bases para descubrir patrones interesantes y conocimiento nuevo, a partir del procesamiento sobre los vínculos contextuales entre los conceptos; constituyendo aspectos distintivos, respecto a similares del estado del arte. Los resultados experimentales obtenidos con la aplicación del método sobre las noticias de prueba, sugieren que a través de los patrones definidos se alcanza una adecuada precisión en la extracción de conceptos y que la ventada contextual constituye el criterio más apropiado para la identificación de las relaciones. Por otra parte, la estrategia de refinado definida posibilita reducir la complejidad del grafo resultante, sin afectar la representación de la información más relevante extraída del texto. No obstante, en investigaciones serán estudiadasfuturas otras alternativas para determinar la relevancia de los conceptos en el grafo y serán ajustados los parámetros a tener en cuenta para definir el tamaño de la ventana contextual, en aras de lograr mayor precisión en la representación que se obtiene del texto.

Bibliografía

Abney, S. (Ed.) (1991). Parsing by Chunks. Kluwer Academic Publishers.

Aggarwal, C. C., & Wang, H., (Eds.) (2010). Managing and Mining Graph Data.Springer Science+Business Media.

Aggarwal, C. C., &Zhai, C., (Eds.) (2012). Mining Text Data.Springer Science+Business Media.

Basole, R. C., Seuss, C. D., & Rouse, W. B. (2013). IT innovation adoption by enterprises: Knowledge discovery through text analytics, Decision Support Systems, vol. 54, 1044–1054.

Chang, J. Y., & Kim, I. M. (2013). Analysis and Evaluation of Current Graph-Based Text Mining Researches, Advanced

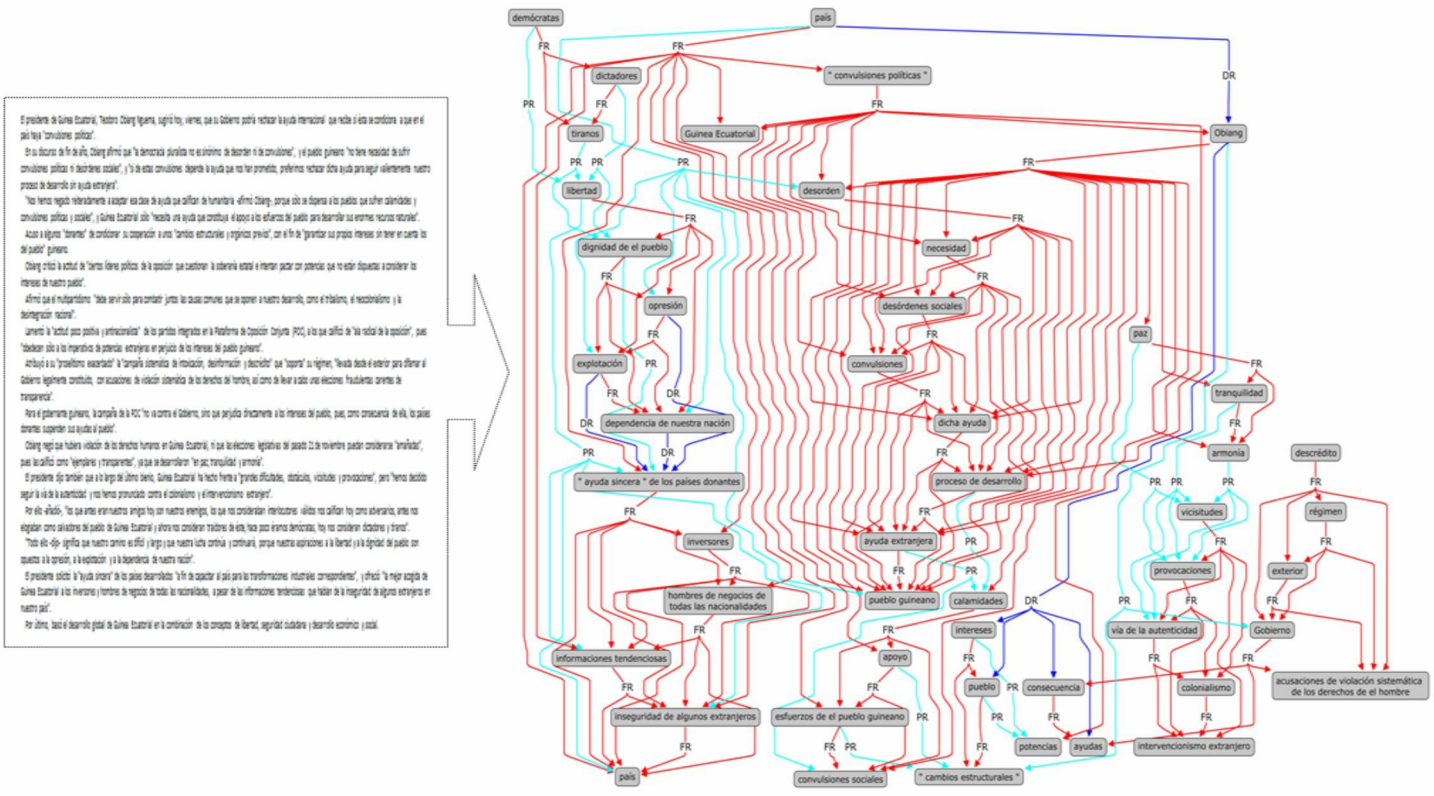

Figura 3. Grafo generado a partir de una noticia (texto a la izquierda)

[1] *http://cmap.ihmc.us/*

Science and Technology Letters, vol. 42, 100-103.

Chang, J. Y., & Kim I. M. (2014). Research Trends on Graph-Based Text Mining, Int. J. of Smart Home, 8(4), 37-50.

Feldman, R., & Sanger, J. (2007).The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. EUA, New York:Cambridge University Press.

Gupta, V., &Lehal, G. S., (2009).A Survey of Text Mining Techniques and Applications, Journal of Emerging Technologies in Web Intelligence, 1(1), 60-76.

Kumar, S., Agrawal, M., Rajput, S., & Kumar, S. (2014).An Information Retrieval(IR) Techniques for text Mining on web for Unstructured data, Int. J. of Adv. Research in Comp. Science and Software Engineering, 4(2), 67-70.

Medina, J., Guevara, E., Hernández, J., Hechavarría, A., & Hernández, R. (2005). Similarity Measures in Documents using Association Graphs. En Proceedings of CIARP'05, LNCS,vol. 3773, pp. 741–751, Springer Berlin Heidelberg.

Montes, M., Gelbukh, A., López-López, A., & Baeza-Yates, R. (2001). Un Método de Agrupamiento de Grafos Conceptuales para Minería de Texto. Procesamiento de Lenguaje Natural, Vol. 27, 115-122.

Novak, J. D., & Cañas, A. J. (2006). The Origins of the Concept Mapping Tool and the Continuing Evolution of the Tool (higher resolution for printing). Information VisualizationJournal, 5(3), 175-184.

Ordoñes, S. &Gelbukh, A.(2010). Representación computacional del lenguaje natural escrito, Ingeniería, 15(1), 6-21.

Palmeira, C., Chaves, R., Cavalcante H., &Favero, E. (2012). A Requirements Elicitation and Analysis Aided by Text Mining, International Journal of Computer Science and Network Security, 12(6), 122-128.

Ramanathan, V., &Meyyappan, T. (2013) Survey of Text Mining, En Proceedings of International Conference on Technology and Business Management, pp. 508-514.

Rodríguez, A., & Simón, A. (2013). Método para la Extracción de Información Estructurada desde Textos, Revista Cubana de Ciencias Informáticas, 7(1), 1-15.

Shekhar, Ch., Sharan, A., &Lata, M. (2014). Semantic Graph Based Approach for Text Mining. En Proceedings of 2014 Int. Conf. on Issues and Challenges in Intelligent Computing Techniques (ICICT), pp. 596-601. IEEE Press.

Simón, A., Ceccaroni, L., Rosete, A., Suarez, A., &Victoria, R. (2008). A Support to Formalize a Conceptualization from a Concept Maps Repository. En Proceedings of CMC'08. pp. 68-75.

Singh, P. D., &Raghuvanshi, J. (2012) Rising of Text Mining Technique: As Unforeseen-part of Data Mining, International Journal of Advanced Research in Computer Science and Electronics Engineering, 1(3), 139-144.

Studer, R., Benjamins, V., &Fensel, D. (1998). Knowledge engineering: principles and methods. Data and Knowledge Engineering, 25(1-2), 161–197.

Sumathy, K. L., & Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues – An Overview, International Journal of Computer Applications, 80(4), 29-32.

Thavamani, C., &Rengarajan, A. (2014).Mining Conceptual Relations from Textual Web Content Using Leximancer, IOSR Journal of Computer Engineering, 16(5), 24-27.

Zhong, N., Li, Y., & Wu, S-T. (2012). Effective Pattern Discovery for Text Mining, IEEE Transactions on Knowledge and Data Engineering, 24(1), 30-44.

Recibido: 9 de diciembre de 2014.
Aprobado en su forma definitiva:
16 de enero de 2015

Aramis Rodríguez Blanco
Instituto Superior Politécnico José Antonio Echeverría (CUJAE), La Habana, Cuba.
Correo-e.: aridriguezb@ceis.cujae.edu.cu

Alfredo Simón Cuevas
Instituto Superior Politécnico José Antonio Echeverría (CUJAE), La Habana, Cuba.
Correo-e.: asimon@ceis.cujae.edu.cu

Ernesto Guevara Martínez
Instituto Superior Politécnico José Antonio Echeverría (CUJAE), La Habana, Cuba.
Correo-e.: eguevara@ceis.cujae.edu.cu

Wenny Hojas Mazo
Instituto Superior Politécnico José Antonio Echeverría (CUJAE), La Habana, Cuba.
Correo-e.: wrojas@ceis.cujae.edu.cu