Revista Cubana de Ciencias Informáticas
Vol. 7, No. 1, Enero-Marzo, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 55-67
http://rcci.uci.cu

Tipo de artículo: Artículo original
Temática: Inteligencia artificial
Recibido: 17/01/2013 | Aceptado: 14/03/2013

Método para la extracción de información estructurada desde textos

Method for structured-information extraction from texts

Aramis Rodríguez Blanco*, Alfredo J. Simón Cuevas

Facultad de Ingeniería Informática, Instituto Superior Politécnico "José Antonio Echeverría". Calle 114, No. 11901 e/ 119 y 127, Marianao, La Habana, Cuba. {aridriguezb, asimon@ceis.cujae.edu.cu}

Resumen

En el trabajo se presenta un método para la extracción de información estructurada desde textos escritos en idioma español, como base para el desarrollo de una propuesta de Minería de Texto. La información extraída es estructurada en forma de grafo, específicamente mediante un Mapa Conceptual, el cual constituye una forma de representación de conocimiento basada en conceptos significativos y sus relaciones en una estructura proposicional. El método propuesto permite procesar documentos de diferentes formatos, y combina el análisis sintáctico superficial y profundo o de dependencias, el reconocimiento de entidades, patrones lingüísticos y conocimientos de referencia almacenado en un corpus de Mapas Conceptuales, para identificar frases conceptuales y relaciones entre ellas, a ser extraídas y representadas en el Mapa Conceptual. SEINET constituye la herramienta que implementa el método propuesto, y a la cual se le han incorporado un conjunto de prestaciones que posibilitan un uso del método eficiente y flexible. Se exponen casos de estudio simples para ejemplificar el funcionamiento del método, y a su vez SEINET.

Palabras clave: minería de texto; extracción de información; mapas conceptuales; construcción automática de grafos desde textos

Abstract

In this work a method for extraction of information structured from Spanish texts is presented, as a base for a Mining of Text proposal development. Extracted information is structured in graph form, specifically in a Concept Map, which constitutes a knowledge representation form based on significant concepts and its relationships in a propositional structure. The proposed method allows to process documents of different formats, and it combines the superficial and deep syntactic analysis or of dependences, entities recognition, linguistic patterns and reference knowledge stored in a Concept Maps corpus, to identify conceptual sentences and relationships among them, to be extracted and represented in the Concept Map. SEINET constitutes the tool that implements the proposed method, and to which have been incorporated a group of benefits that facilitate the efficient and flexible use of the method. Simple cases of study are exposed to exemplify the operation method, and in turn SEINET.

Keywords: *text mining; information extraction; concept maps; automatic construction of graph from text.*

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 1, Enero-Marzo, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 55-67
http://rcci.uci.cu

Introducción

La gran cantidad de información, textual y no estructurada, que se encuentra almacenada y que continuamente se genera, sobre todo en la Web, demanda de iniciativas que propicien un mayor aprovecha miento de ese recurso - información - para el descubrimiento de conocimiento y la toma de decisiones. La *Minería de Texto (MT)* constituye el área de conocimiento dentro de la que se estudia esta problemática, y desde donde se generan soluciones para el *descubrimiento de conocimientos potencialmente útiles, y no explícito, en una colección de textos, a partir de la identificación y exploración de patrones interesantes* (Feldman et al., 1998). Un aspecto importante en el desarrollo de soluciones de MT lo constituye la representación intermedia que se utilice para la estructuración y almacenamiento de los contenidos extraídos en la etapa de pre-procesamiento, ya que sobre esa estructura es que aplican las técnicas de análisis para alcanzar el descubrimiento de conocimiento. El trabajo que se presenta aborda esta problemática.

Específicamente, se presenta una propuesta dirigida a la construcción, de forma automática, de una representación intermedia basada en gafos para representar y estructurar el contenido conceptual de textos, específicamente, se propone usar el Mapa Conceptual (MC). Los MCs constituyen *una herramienta para organizar y representar el conocimiento* (Novak y Cañas, 2008), en forma de *grafo dirigido* y *etiquetado*. Se componen de *conceptos* y *relaciones* que forman una estructura de *proposiciones*. Los conceptos representan eventos u objetos, o evidencias de ellos, especificados por una etiqueta y las relaciones están etiquetadas por una frase que establece el tipo de relación entre los conceptos. Las proposiciones se forman por dos o más conceptos interconectados mediante una *frase-enlace*, representando expresiones significativas (Novak y Cañas, 2008), y en ocasiones son consideradas como unidades semánticas o de significado. Un aspecto motivador y que aporta valor a esta propuesta es que se cuenta con un álgebra de consulta para repositorios de MCs, CMQL (Concept Maps Query Language) (Simón-Cuevas et al., 2008), que permite la búsqueda, exploración, recuperación de conocimiento desde diferentes vistas y perspectivas, así como la generación de nuevos conocimientos potencialmente útiles, a partir de la integración automática de conocimientos inicialmente aislados. Por lo que, CMQL, puede formar parte, o ser una base importante en una propuesta de MT a partir de que el contenido conceptual de los textos sea extraído y estructurado en forma de MC.

La construcción automática de MCs a partir de texto ha sido abordado por varios autores y hay contribuciones interesantes (Valerio y Leake, 2006; Valerio et al., 2008; Kowata et al., 2010; Estrada, 2011), pero en su gran mayoría dirigidas al idioma inglés, y solo la reportada en (Estrada, 2011) antecedente de la propuesta que se hace así, aborda el procesamiento de textos en idioma español. El método que se propone en este trabajo, permite el procesamiento de documentos de diferentes formatos y de forma masiva, combina el análisis sintáctico superficial y profundo o de dependencias, el reconocimiento de entidades, un conjuntos de patrones lingüísticos y conocimientos de referencia almacenado en un corpus de MCs, para identificar frases conceptuales y relaciones entre ellas, a ser extraídas y representadas en el MC.

El método que se propone ha sido implementado a través de la herramienta SEINET, que representa las siglas de: *S*istemas para la *E*xtracción de *IN*formación *E*structurada desde *T*extos. En SEINET se incorporan un conjunto de prestaciones que posibilitan un uso del método eficiente y flexible, tal es el caso de técnicas de paralelismo para el procesamiento masivo de documentos y el pre-procesado, un editor de MC para refinar los resultados sin necesidad de un editor extra, reportes de estadísticas para el estudio y experimentación del método, la generación de CXL (Cañas et al., 2006) para almacenar el MCs y posibilitar la reutilización de ese conocimiento, entre otras. A través de la exposición de casos de estudio simples se ejemplifica la aplicación del método, y la herramienta SEINET, sobre textos en español, donde se puede apreciar la calidad los resultados, los cuales han sido comparados con una propuesta antecedente.

Minería de texto

La *Minería de Texto (MT)* surge como un enfoque particular del proceso de descubrimiento de conocimiento, específicamente, orientado al descubrimiento en fuentes textuales y no estructuradas. Se puede definir como un *proceso de descubrimiento de conocimientos potencialmente útiles, y no explícito, en una colección de textos, a partir de la identificación y exploración de patrones interesantes* (Feldman et al., 1998). En la MT se utilizan técnicas provenientes de la inteligencie artificial, de la gestión del conocimiento, de la minería de datos y del aprendizaje automático, así como del procesamiento de lenguaje natural y de la recuperación de información, desde donde se proveen métodos y herramientas para identificar, organizar y comprender la sintaxis y la semántica de los contenidos en lenguaje natural no estructurados presentes en los textos. Algunos de los elementos a descubrir en las colecciones de documentos son cosas tales como: tendencias, desviaciones y asociaciones (Montes y Gómez, 2001).

La MT involucra un conjunto de fases (Feldman y Sanger, 2007):

1. pre-procesamiento de la colección de documentos (ej. categorización de textos, extracción de información, extracción de términos);
2. la estructuración y almacenamiento de los contenidos extraídos en una representación intermedia (modelos vectoriales, relacionales, lista de palabras, entre otros);
3. la aplicación de técnicas de análisis sobre la representación intermedia (tales como análisis de distribución, clustering, análisis de tendencias, y reglas de asociación), con el objetivo de llegar al descubrimiento;
4. la visualización de los resultados.

El trabajo que se presenta constituye una propuesta dirigida a la construcción, de forma automática, de una representación intermedia basada en gafos para representar y estructurar el contenido conceptual de textos, específicamente, un MC. El uso de los MC como forma de representación intermedia, alcanza un mayor valor, y al mismo tiempo ha constituido una fuente motivacional para el desarrollo del trabajo, la disponibilidad un álgebra de consulta para repositorios de MC, CMQL (Concept Maps Query Language) (Simón-Cuevas et al., 2008) que permite la búsqueda, exploración, recuperación de conocimiento desde diferentes vistas y perspectivas, así como la generación de nuevos conocimientos potencialmente útiles, a partir de la integración automática de conocimientos inicialmente aislados. En este sentido, CMQL puede ser considerado una alternativa base en una propuesta de MT a partir de que el contenido conceptual de los textos sea extraído y estructurado en forma de MC.

Mapas Conceptuales

El MC fue creado por J. D. Novak como una forma de instrumentar la teoría de aprendizaje significativo, la que se sustenta en que el nuevo conocimiento es adquirido a partir de lo que ya se conoce y esto se realiza a través de un proceso constructivista (Ausubel y Novak, 1989). En este escenario, Novak define un MC 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* (Novak y Gowin, 1984). Otra definición más general es la que los reconoce como *una herramienta para organizar y representar el conocimiento* (Novak y Cañas, 2008), en forma de *grafo dirigido* y *etiquetado*.

Los MC se componen de conceptos y relaciones formando proposiciones. Los conceptos son definidos como regularidades percibidas en eventos u objetos, evidencias de ellos, especificados por una etiqueta, generalmente formada por una palabra (simple), pero es posible usar más de una (compuesta). Las relaciones están etiquetadas por una frase que establece el tipo de relación entre los conceptos. Las proposiciones se forman por dos o más conceptos

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 1, Enero-Marzo, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 55-67
http://rcci.uci.cu

interconectados mediante una frase-enlace, representando expresiones significativas (Novak y Cañas, 2008), y en ocasiones son consideradas como unidades semánticas o de significado. En la Figura 1. se muestra un ejemplo de un MC en el que se representa e interrelacionan algunos de los conceptos antes mencionados, así como otros que lo caracterizan.

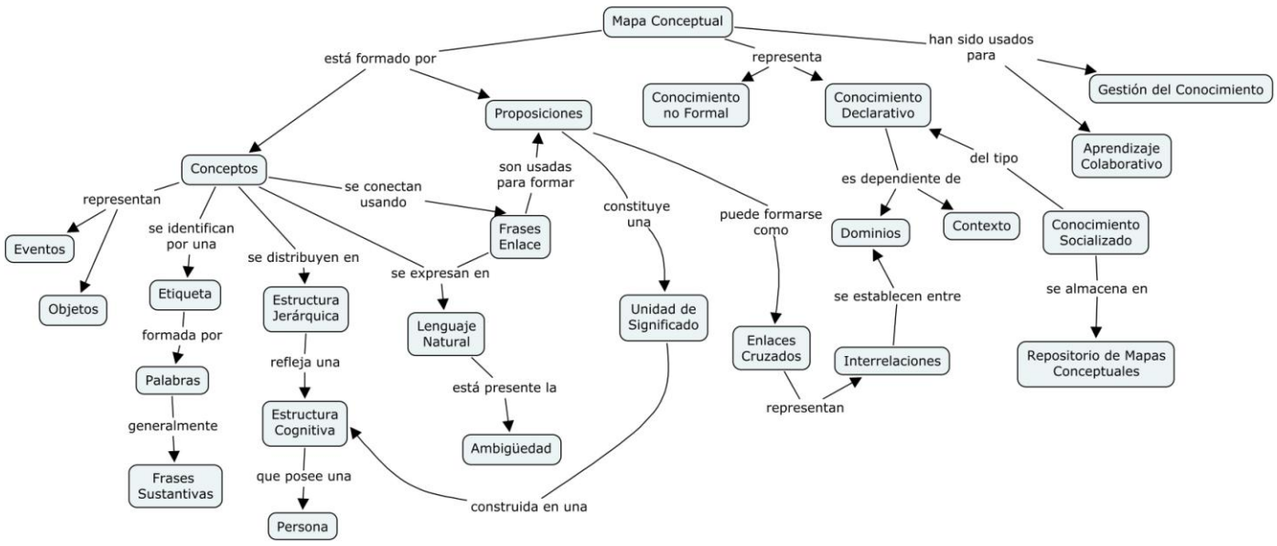


Figura 1. Representación (ejemplo de MC) de conceptos significativos que caracterizan a un MC como forma de representación de conocimiento.

Las principales aplicaciones de los MCs son en la pedagogía, como apoyo al proceso de enseñanza y el aprendizaje, aprovechando las bondades que brinda siendo muy intuitivo para las personas. Sin embargo, otras de las aplicaciones importantes lo constituye su uso como herramienta para la captura de conocimiento tácito de experto, así como apoyo en varias de las actividades básicas de la gestión del conocimiento, a saber: crear, generar, compartir, transferir, capturar, almacenar conocimiento, entre otras. En este trabajo se propone emplear los MC como herramienta de representación intermedia del contenido de documentos en español, en el sentido de que éstos sean construidos automáticamente a partir de los documentos, como resultado del método para la extracción de información estructurada en grafo que se presenta más adelante.

Construcción automática a partir de textos

Varios autores han tratado el tema de la construcción automática de MCs a partir de textos (Valerio y Leake, 2006; Valerio et al., 2008; Kowata et al., 2010; Estrada, 2011), y se han hecho propuestas, caracterizadas en su mayoría por ser soluciones dirigidas a textos en idioma inglés. Valerio y Leake proponen un algoritmo de extracción de información a partir de documentos, cuya información es utilizada en la construcción de un MC parcial, a ser refinado posteriormente por expertos. La información se extrae de los documentos en línea, y la construcción del MC está planteada como una relación 1-1 (MC-documento), pero puede transformada en una relación n-m, introduciendo la segmentación por tópicos. Inicialmente el documento es segmentado y cada sentencia o segmento es analizada sintácticamente usando el algoritmo de Charniak (Charniak y Johnson, 2005). En la extracción de conceptos, se determina que una palabra forma una frase conceptual, si es sustantivo o adjetivo, abordándose primero las frases más simples, que son las más cercanas a las hojas en un árbol de dependencias, luego las más complejas. En (Valerio et al., 2008) se reporta una propuesta de aplicación de la construcción automática de MC desde textos para la clasificación de documentos.

En (Kowata et al., 2010), se presenta un método de construcción de MC a partir de texto en el que se incluyen las siguientes tareas (en ese mismo orden de ejecución, y cada una depende de la salida de la anterior):

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 1, Enero-Marzo, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 55-67
http://rcci.uci.cu

1. Extracción de Texto Plano,
2. Segmentación del Texto
3. Extracción de tokens
4. POS Tagging
5. Reconocimiento de elementos Centrales Candidatos (conceptos y enlaces candidatos)
6. Intérprete de dependencias
7. Constructor del MC

En esta propuesta se consideran las frases sustantivas, verbales y preposicionales como primeros candidatos a ser elementos principales de los MCs. Se indica que la fragmentación de la oración es una tarea importante para el reconocimiento de los principales elementos candidatos a incorporar al MC y que cada fragmento es creado de un conjunto de patrones lingüísticos formalmente descritos por expresiones regulares. El uso de métodos lingüísticos para la construcción de MC a partir de texto también se considera en (Valerio y Leake, 2006), donde también se usan frases sustantivas y verbales para extraer conceptos y relaciones.

Desarrollo

Método de extracción de información estructurada

El método que se propone toma como base las propuestas reportadas en (Simón et al., 2004; Estrada, 2011), y propone una variante mejorada a partir de tomar en consideración algunos aspectos tenidos en cuenta en las propuestas reportadas en (Valerio y Leake, 2006; Valerio et al., 2008; Kowata et al., 2010), aunque con la característica que al igual que en (Simón et al., 2004; Estrada, 2011), la nueva versión también está dirigida al procesamiento de textos en idioma español.

En el proceso de extracción y estructuración de información se parte de una información textual (contenida en un fichero o introducida manualmente) en lenguaje natural no estructurado, y se ejecutan un conjunto de tareas que pueden ser agrupadas en tres etapas, a saber: *pre-procesamiento*, *extracción de información* (frases conceptuales y relaciones) y *refinado y construcción del MC*, donde como resultado se estaría obteniendo de forma automática un MC que representa el conjunto de frases conceptuales (como nodos) que fueron identificadas (todas las posibles) y las relaciones entre ellas. En el caso de las relaciones que se identifican, a diferencia del resto de las propuestas reportadas, es posible identificar relaciones no explícitas en el texto, las cuales se pueden detectar a partir del uso de un corpus de MC como recursos de conocimiento con el que se puede dotar al procesamiento que se ejecuta en este método. El corpus de MCs también facilita la identificación de frases conceptuales. En la Figura 2. se muestra el esquema general del método.

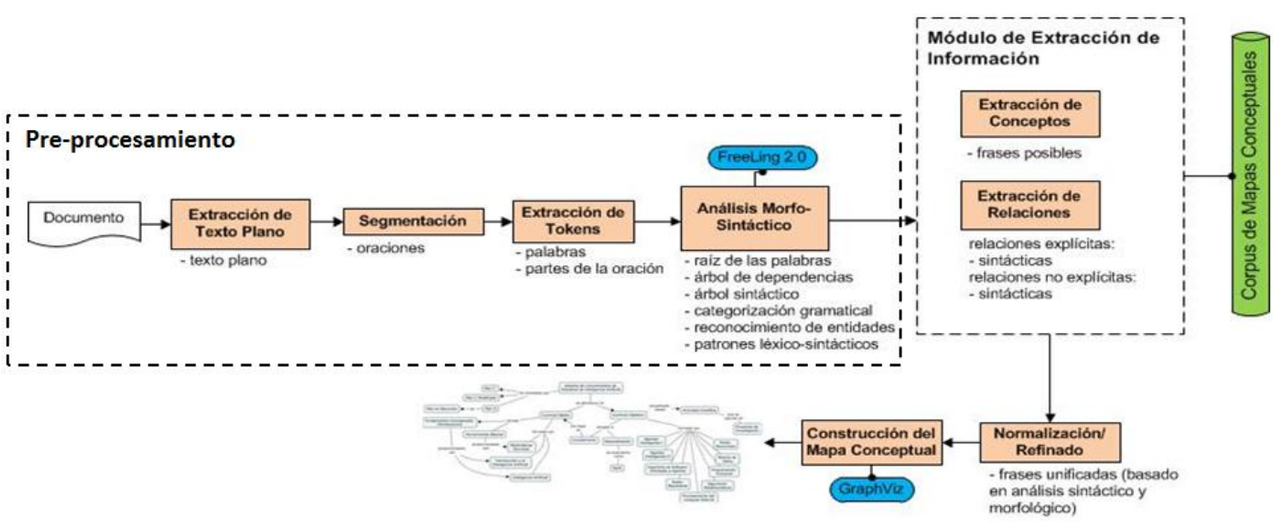


Figura 2. Esquema Funcional del Método de Extracción de Información Estructurada en forma de MC

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 1, Enero-Marzo, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 55-67
http://rcci.uci.cu

Pre-Procesamiento del texto

Extracción de Texto Plano: El método parte de información textual almacenada en uno o varios ficheros o proporcionado de forma manual. En esta tarea se extrae el texto plano de la fuente, y en el caso de ficheros se emplean librerías implementadas en lenguaje Java para la extracción del texto plano de ficheros con formato: pdf, docx, doc, html, htm, rtf, y txt, brindando una mayor cobertura del método.

Segmentación de Texto: La segmentación de texto consiste en desfragmentar el mismo en párrafos y oraciones, las cuales se segmentan utilizando un algoritmo para la determinación de sus fronteras, teniendo en cuenta varios tipos de segmentación, principalmente, la segmentación a partir de los *puntos finales*, para su identificación se tuvo en cuenta cuales son las funciones que puede jugar un punto en la oración. De esta forma, se obtiene una lista de oraciones a ser procesada, como resultado de la segmentación.

Extracción de tokens: Cada oración está compuesta por *tokens*, que no son más que cada una de las partes de la oración, o sea, palabras, números, signos de puntuación, etc. Este proceso divide cada oración en un conjunto de *tokens*, los cuales serán la base de análisis posteriores. Los *tokens* se identifican en una oración mediante un algoritmo que realiza una lectura de la misma, identificando las fronteras entre las diferentes clasificaciones de tokens.

Análisis morfo-sintáctico: El análisis morfo-sintáctico del texto se realiza a cada oración por separado. Inicialmente se etiqueta cada *token* con lo que se determina que es su raíz morfológica y su categoría gramatical. A partir de los *tokens etiquetados* se realiza el *Análisis Sintáctico Superficial (*ASS*)* del texto, el cual consiste en agrupar a varios *tokens* en lo que en la bibliografía se le conoce como *chunks*.

Los *chunks* constituyen estructuras gramaticales, como grupos verbales, sintagmas preposicionales, sintagmas nominales, etc., las cuales se organizan en forma de árbol y se agrupan en un conjunto consecutivo de subárboles. Luego del ASS se realiza el *Análisis Sintáctico Profundo* o conocido también como *Análisis de Dependencias* (AD), con el cual se determinan las dependencias entre las diferentes estructuras gramaticales identificadas en el ASS, dando como resultado otra estructura en forma de árbol. El ASS es realizado empleando la herramienta libre FreeLing 2.0. Freeling es una herramienta de código abierto que provee servicios de análisis del lenguaje natural como análisis morfológico, análisis sintáctico y de dependencias, etiquetador de categorías sintácticas, divisor de oraciones, reconocedor de entidades, fechas, números, magnitudes físicas, monedas y anotación de sentidos basado en WordNet (Miller et al., 1993). Otro elemento importante en la elección de FreeLing, es que es una de las pocas herramientas que proporcionan funcionalidades para el procesamiento y análisis de documentos en idioma español.

En el caso del AD es realizado mediante un algoritmo propio elaborado como parte de la solución propuesta, y con el fin de obtener un resultado superior al obtenido con el algoritmo de dependencias que proporciona FreeLing. La propuesta de AD parte del resultado del ASS realizado por FreeLing y devuelve un árbol de dependencias como salida. Para ello se ejecutan tres tareas principales:

1. el refinado del ASS devuelto por FreeLing;
2. la transformación de la estructura superficial a la estructura de dependencias; y
3. la determinación de las dependencias.

En la primera tarea se resuelve la mayoría de los problemas sintácticos no resueltos por la versión 2.0 de FreeLing y prepara el ASS para la segunda y tercera tarea, de forma tal que esta se pueda ejecutar con calidad. La segunda tarea transforma la estructura del árbol del ASS, de manera que cada estructura quede representada por un árbol de dependencias. La tercera tarea crea las dependencias entre las raíces de los subárboles de la lista unificándolos en un solo árbol de dependencias, y en el caso de no haber sido unificados todos los subárboles de la lista, entonces el árbol de dependencia resultante consistiría en una floresta, en la cual las raíces de cada uno de los arboles están al mismo

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 1, Enero-Marzo, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 55-67
http://rcci.uci.cu

nivel. En el análisis morfo-sintáctico también se lleva a cabo la tarea de reconocer todas las entidades incluidas en el documento y según sean identificadas por FreeLing, ya que se ha considerado de principio que todas las entidades, pueden ser consideradas como frases conceptuales potenciales, lo que se trata a continuación.

Extracción de información

El módulo de extracción de información representa el núcleo del método propuesto, dado que aquí es donde se extraen los contenidos de información claves para la estructuración y construcción del MC, tal es el caso de las frases que expresan conceptos potenciales, como las relaciones que se establecen entre ellos y sus correspondientes etiquetas. En este proceso se tiene en cuenta toda la información sintáctica obtenida en la fase anterior y se dispone de un mecanismo de trabajo con un corpus de MCs, el cual es conformado según los intereses de dominio del usuario.

Extracción de conceptos: El proceso de extracción de conceptos consiste en identificar aquellas frases (conjunto de palabras) o palabras simples que pueden tener un sentido conceptual. Para la identificación de estos conceptos potenciales se definen un conjunto de patrones lingüísticos, formulados a partir de un conjunto de categorías gramaticales las que se muestran en la Tabla 1. En la selección de las categorías gramaticales relevantes para la identificación de conceptos se tuvieron en cuenta los resultados reportados en (Villalon et al., 2010), como por ejemplo que el 80% de los conceptos correspondían a frases sustantivas lo cual sugiere que existe una estrecha relación entre sustantivos y conceptos. Esto hace considerar que las frases sustantivas representadas en el árbol sintáctico superficial puedan ser identificadas como conceptos potenciales de forma directa a ser incluidos en el MC resultante. La tarea iniciar de extracción está dirigida a usar el corpus de MC, específicamente, identificando conceptos que estén en alguno de los MC y también en el texto. Luego se procede al uso de los patrones lingüísticos para identificar conceptos en el árbol sintáctico resultante del ASS. En la Tabla 2 se muestran algunos ejemplos de patrones lingüísticos.

Tabla 1. Categorías Gramaticales Consideradas en la formulación de Patrones Lingüísticos

Categorías Gramaticales	**Ejemplo**	
NC: sustantivo común.	*casa*	
NP: sustantivo propio.	*Unión Europea*	ONU
A: adjetivo.	*grande*	
R: adverbio.	*ahora*	
Z: número.	*Treinta y dos*	*32*
W: fechas.	*10 de mayo de 2001*	
VMN: verbo en infinitivo.	*Vivir*	

Tabla 2. Formulación y Ejemplificación de Algunos Patrones Lingüísticos.

Patrones Lingüísticos	**Ejemplos**	
((D))+(*)	*esta casa*	El 32
((D))+(NC | NP)+(A | VMP)	*casa grande*	*casa destruida*
((D))+(NC)+(Z)	*La habitación 32*	
((D))+(A)+(NC)+(Z)	*nueva habitación 32*	
((D))+(<(A)+(<(Fc)+(A)>)!+(CC)+(A)>)+((NC)	*La alta y vieja casa*	
((D))+(NC)+(<(A)+(<(Fc)+(A)>)!+(CC)+(A)>)	*La casa alta, ancha, larga y vieja*	
Leyenda: D: Determinantes; SP: Preposición; VMP: forma verbal en pasado participio en función adjetiva; Fc: , (coma); CC: Conjunción; ( ): Término obligatorio; ( )!: Término repetido 0 o más veces; (( )): Término opcional; (*): Categoría gramatical (enunciada en Tabla 1); < >: Lista de términos; | : Disyunción		

La aparición en el texto, de una secuencia de las posibles combinaciones que se formulan a través de los patrones, definidas por signos de separación y conjunciones, constituye una lista de conceptos. Algunos conceptos se complementan con otros a través de preposiciones para evitar la pérdida de sentido conceptual. Los conceptos identificados son almacenados en una lista, para poder ser usados en la identificación de relaciones.

Extracción de relaciones: La extracción de relaciones va dirigida a dos caminos fundamentalmente, la *extracción de relaciones explícitas* y la *extracción de relaciones implícitas*, siendo esta última aquel tipo de relación entre conceptos que no aparece de forma evidente en el texto que se procesa. Las relaciones explícitas generalmente ocurrirán entre frases conceptuales que ese encuentran en una misma oración, y en muy poca medida entre conceptos que estén en diferentes partes del texto, lo que no ocurre con las relaciones implícitas, que si fundamentalmente conectan conceptos que no están en la misma oración, aunque esto dependerá de la manera en que el conocimiento de referencia ha sido representado en los MCs del corpus.

La *extracción de relaciones explicitas* se realiza a partir de un proceso de identificación de estas relaciones sobre el árbol de dependencias resultante del AD. Aquí se estarían aprovechando las dependencias que existen entre las estructuras gramaticales para crear reglas que conecten a los conceptos entre ellos. Esto permite conectar a uno o más conceptos con otro(s) a través de una misma frase de enlace. Los conceptos que se conectan a través de frase de enlaces verbales, se conectan mediante un mecanismo que identifica las relaciones verbales que existen entre los sujetos de cada oración simple con los complementos que pertenecen a su contexto de dependencia. Las relaciones en las que la frase de enlace está compuesta por una conjunción subordinada y una frase verbal indican subordinación o dependencia. Estas relaciones se determinan mediante la dependencia de una conjunción subordinada de una estructura conceptual, la cual se conecta con el sujeto de la oración subordinada a la conjunción. Las relaciones en las que la frase de enlace son preposiciones, utilizan la ventaja de las dependencias para relacionar a varios conceptos consecutivos con varios conceptos estructurados de la misma manera. Los conceptos relacionados en el árbol de dependencia se construyen a partir de la identificar en la lista de conceptos, un posible concepto, que se construye a partir de la información relacionada en un subárbol que tiene como raíz un token etiquetado con una de las categorías gramaticales enunciadas en la Tabla 1. Las frases de enlace, en caso de contener una frase verbal se construyen a partir de recolectar la información referente a un grupo verbal, identificado en el árbol de dependencias como un subárbol que tiene como raíz un *token* etiquetado como forma verbal, y como hijos directos, los modificadores de la frase de enlace, así como los conceptos que se relacionan a través de la misma. En caso de ser una frase de enlace preposicional, se determina los conceptos orígenes, que pueden estar agrupados en un subárbol, con los conceptos destinos, que deben estar agrupados en un subárbol subordinado al nodo de la frase preposicional. El uso del AD permite abarcar más relaciones, que con el uso del ASS, como se reporta en (Estrada, 2010). Además, la posibilidad de extraer relaciones incorrectas disminuye considerablemente. Aunque hay que reconocer que la calidad de las relaciones está influenciada, en gran medida, por la calidad del AD.

La *extracción de relaciones implícitas* se basa en el uso del corpus de MCs como conocimiento de referencia y bien formado, cuyos MCs se recomiendan que sean de dominio específico, vinculado con el MC que se procesa. Este tipo de relaciones de idéntica cuando existe una relación entre conceptos representados en alguno de los MCs del corpus y que también están en el texto.

Al final, los conceptos que no son relacionados se catalogan como “conceptos huérfanos” y se extraen a una lista aparte la cual será incluida en la lista final de conceptos cuando se genere el MC de modo que pueda brindarse la mayor información posible sobre el texto.

Refinado y construcción del Mapa Conceptual

Esta constituye la última fase del método. Luego de haber extraído los conceptos y relaciones explícita y no explicita entre ellos y formada las proposiciones, se procede a eliminar posibles errores que puedan existir en dichas proposiciones, para lo cual se ejecutan una serie de reglas, las cuales indicaran que proposiciones deben ser eliminadas. Por ejemplo, son eliminadas las proposiciones en las que el concepto origen es igual al concepto destino, o en las que alguno de los dos conceptos contenga a la frase de enlace. Además se eliminan las proposiciones repetidas, que son las que cada concepto en una proposición sea igual o contenga a su correspondiente en otra proposición y que las frases de enlace sean iguales o una incluya a la otra.

Luego de concluido el refinado de las proposiciones, se procede con la normalización de los conceptos, que no es más que un proceso de unificación de conceptos, a partir de un algoritmo de comparación sintáctica, unificándose aquellos conceptos muy similares sintácticamente. También se unifica las frases de enlace, pero en este caso no solo es suficiente con que sean iguales, sino que se toma en cuenta su función en el MC, ya que si se unifica incorrectamente se le puede estar atribuyendo a un concepto origen, los conceptos destinos a los que se une a través de la frase de enlace otro concepto origen. Por tanto para unificar se debe verificar que en todas las proposiciones en la que esta esté, los conceptos origen se enlacen con los mismos conceptos destinos. Las frases desenlazadas se eliminan posteriormente. Se consideran como iguales aquellas frases de enlace en las que las palabras sean las mismas en el mismo orden. Por último, luego que se ha refinado la información extraída y estructurada en proposiciones se procede a la última tarea y es la integración de todas las proposiciones para construir finalmente el Mc y visualizarlo.

Implementación y ejemplificación

El método propuesto ha sido implementado en una herramienta experimental que se ha bautizado con el nombre de *SEINET*, correspondiente a las siglas de Sistema para la Extracción de Información Estructurada en Textos, fundamentalmente orientada al procesamiento de textos en español. Algunas de sus cualidades más relevantes son las siguientes:

- soporte para el procesamiento masivo de documentos, es decir, dada una colección de documentos la herramienta es capaz de procesar cada uno de los documentos de forma automática, y en un tiempo razonablemente breve.
- posibilidad de exportar el MC resultante a formato CXL (Concept Mapping Extensible Language) (Cañas et al., 2006), lo que posibilita que pueda ser editado en CmapTools (Cañas et al., 2004), así como reutilizado ese contenido por terceras aplicaciones.
- aplicación de técnicas de paralelismo en el procesamiento masivo de documentos, así como en actividades de la fase de pre-procesamiento.
- incorporación de un editor de MC que permite modificar/refinar el MC resultante sin necesidad de disponer de CmapTools.
- generación de reportes estadísticos sobre el procesamiento y construcción de los MCs útiles para la realización de pruebas continuas al método.
- parametrización de categorías gramaticales y los patrones lingüísticos para facilitar el estudio del método en diferentes escenarios.
- implementación del método basada en componentes, lo que permite la reutilización de sus partes funcionales.

En la Figura 3. se muestran diferentes vistas de la interfaz de SEINET.

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 1, Enero-Marzo, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 55-67
http://rcci.uci.cu

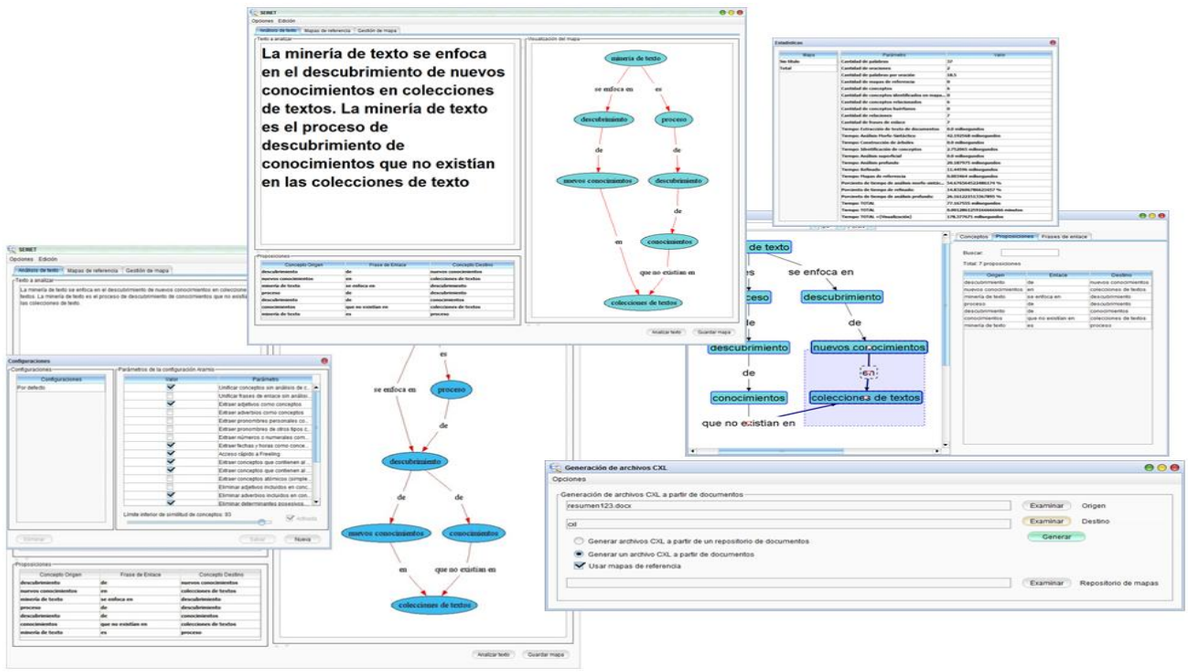


Figura 3. Vistas de la interfaz de SEINET.

En la Tabla 3. se muestra, a modo de ejemplo los resultados que obtiene SEINET y CMAG (Estrada, 2011) para un texto simple. Los MCs que se muestran en dicha tabla han sido construidos automáticamente por ambas herramientas. Aunque en amos MC se logra representar casi la totalidad del contenido del texto simple de ejemplo, se aprecia que a través de SEINET se logra una mayor calidad en cuanto a la precisión en la extracción de frases conceptuales, obteniéndose conceptos con una estructura más simple y no están repetidos, como sucede en la salida de CMAG. También se aprecian mejoras en la extracción de relaciones, evidenciándose en estas últimas, una mayor coherencia y mayor cantidad de relaciones entre conceptos que la identificada en CMAG. Se aprecia además como el resultado del proceso de refinado o normalización contribuyen a que se logre un MC en el que se integra todo el contenido, lo que no ocurre en el MC obtenido por CMAG. La ampliación de los patrones lingüísticos en la identificación de conceptos y relaciones, así como y la aplicación del AD han incidido en un aumento de la calidad en la salida de SEINET en comparación con la de CMAG. Por otro lado, el uso de una técnica de acceso rápido a la herramienta FreeLing y ha permitido que el tiempo de ejecución disminuya considerablemente (CMAG: 6.2 seg. vs. SEINET: 0.1 seg.). Los conceptos tienen una estructura más simple y no están repetidos, como sucede en la salida de CMAG, esto confirma la calidad del proceso de identificación de conceptos.

En la Tabla 4. se muestra otro ejemplo, pero en este caso, solo ejecutando SEINET pero al que se le ha incorporado un MC formado por una proposición como conocimiento de referencia para el procesamiento del texto de la Tabla 3.

Otra de las mejoras incluidas en la salida de SEINET es la identificación de la relevancia de los conceptos y las relaciones mediante su representación con diferentes colores. En el caso de los conceptos, los que son extraídos solamente del texto, se representan gráficamente con un color aleatorio, mientras que los que además están presentes en los mapas de referencia, se representan usando un color predefinido y que contrasta con el color del resto de los conceptos.

Tabla 3. Ejemplificación de procesamiento de texto simple con SEINET y CMAG (Estrada, 2011).

Texto de ejemplo:
Los Mapas Conceptuales están formados por un conjunto de proposiciones formadas por dos conceptos unidos a través de una frase de enlace. La construcción automática de Mapas Conceptuales permite el procesamiento de la información almacenada en los textos de manera eficiente.

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 1, Enero-Marzo, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 55-67
http://rcci.uci.cu

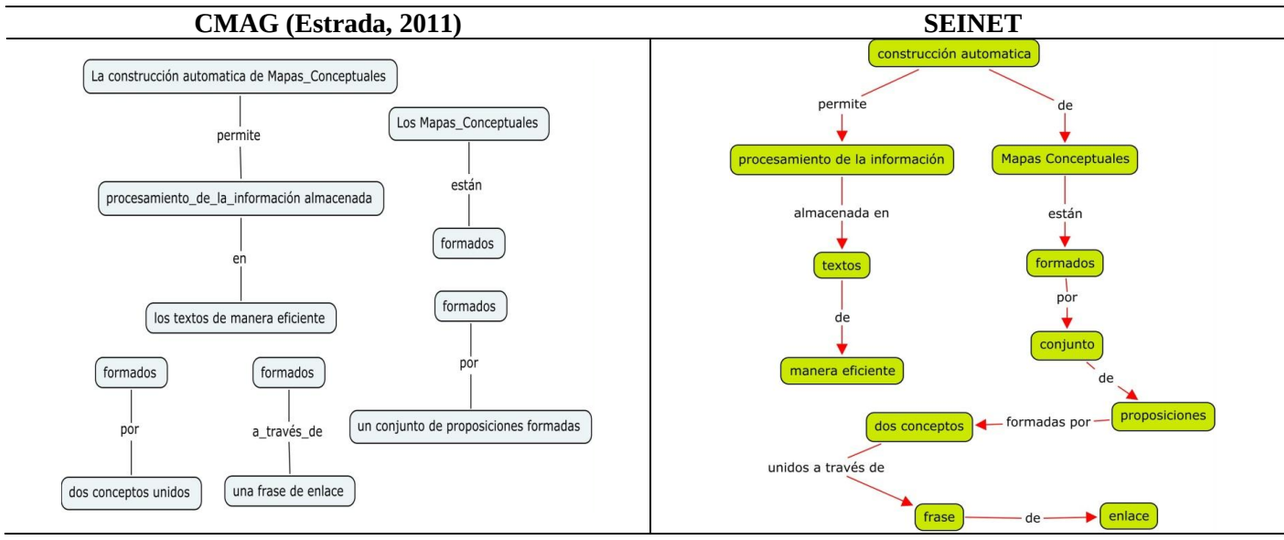
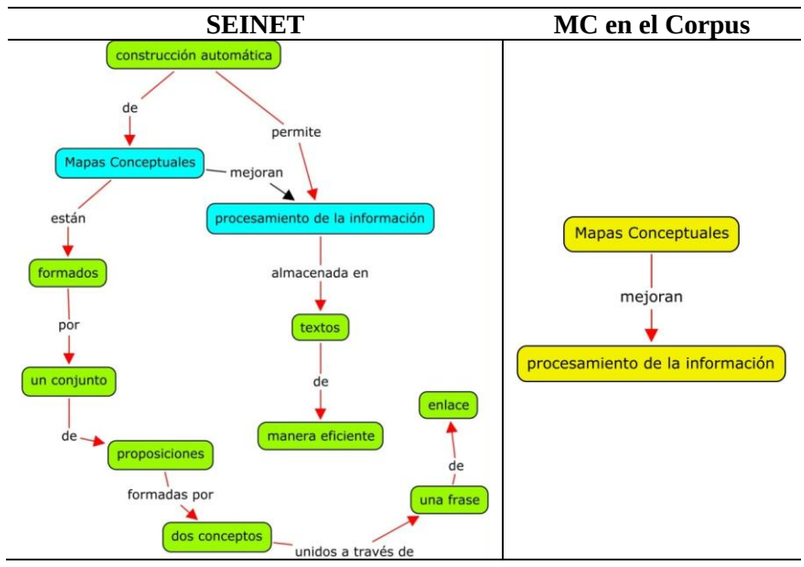


Tabla 4. Ejemplificación usando un MC de referencia en el corpus.



De igual manera ocurre con las relaciones, las cuales se representan de color rojo, si son relaciones explícitas, o sea, que son identificadas en el texto, mientras que las relaciones implícitas, que se extraen de los MCs de referencias, se representan de color negro. Ambas situaciones son apreciadas en la Tabla 4, donde aquí se precia como se ha incorporado al MC resultante (a la izquierda) una relación implícita entre los conceptos '*mapas conceptuales*' y '*procesamiento de la información*' complementando la información extraída del texto y evidenciando la utilidad del uso de MCs de referencia en un corpus configurable.

Conclusiones

En el trabajo se ha presentado un método para la extracción de información estructurada desde textos escritos en idioma español, con beneficios para la Minería de Texto y que representa bases para futuros desarrollo en esa dirección. La información extraída automáticamente es estructurada a través de un MC. Las tareas de pre-procesamiento brindan soporte para diferentes formatos de textos, y se combina el análisis sintáctico superficial y profundo o de dependencias, así como el reconocimiento de entidades, para aportar la mayor cantidad de información sintáctica posible a la fase de identificación y extracción de frases conceptuales y relaciones explícitas e implícitas. Los patrones lingüísticos definidos y el conocimiento en el corpus de MCs como referencia posibilitan aprovechar en

mayor medida la información sintáctica suministrada en la fase anterior, y al mismo tiempo mejorar la calidad en la identificación de conceptos y relaciones, con respecto a otras propuestas. Se presenta SEINET como herramienta experimental que implementa el método propuesto, a la que se han incorporado un conjunto de prestaciones que contribuyen a un uso más eficiente y flexible del método, así como la reutilización de las funcionalidades que implementa de la información estructurada en MCs resultante. Los casos de estudio, aunque simples, permiten ejemplificar el método propuesto, así como el funcionamiento de SEINET, y apreciar resultados parcialmente satisfactorios, considerando que el grado de complejidad, del problema tratado en este trabajo es nada despreciable.

Referencias

- AUSUBEL, D. y NOVAK J. D., "Psicología Educativa," México Trillas ed, 1989.
- CAÑAS A. J., y CARVALHO M., Concept Maps and AI: an Unlikely Marriage?", Revista Brasileira de Informática na Educação, 2004.
- CAÑAS A. J., HILL G., CARFF R., SURI N., LOTT J., GÓMEZ G., ESKRIDGE T. C., ARROYO M., y CARVAJAL R., CMapTools: A Knowledge Modeling and Sharing Environment, en Proc. of the First International Conference on Concept Mapping, Universidad Pública de Navarra: Pamplona, Spain, 2004, pp. 125-133.
- CAÑAS A. J., HILL G., BUNCH L., CARFF R., ESKRIDGE T., Y PÉREZ C., KEA: A Knowledge Exchange Architecture Based On Web Service, Concept Maps and CmapTools, en Proc. Of Second International Conference on Concept Mapping (CMC'06), Vol. 1, San José, Costa Rica, 2006, pp. 304-310.
- CHARNIAK E., y JOHNSON M., Coarse-to-fine n-best parsing and Maximum Entropy discriminative reranking. ACL'05, 2005.
- FELDMAN R., SANGER J., The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York, Cambridge University Press, 2007, 410 p.
- FELDMAN R.; FRESKO M.; KINAR Y.; LINDELL Y.; LIPHSTAT O.; RAJMAN M.; SCHLER Y.; ZAMIR O., Text Mining at the Term Level, en Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), 1998.
- KOWATA J. H., CURY D. y BOERES M. C. S., Concept Maps Core Elements Candidates Recognition From Texts, en Proc. of Fourth International Conference on Concept Mapping (CMC'10), Viña del Mar, Chile, 2010.
- MILLER G., BECKWIDTH R., FELLBAUM C., GROSS D., y MILLER K., Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography, 3(4), 1993, pp. 235-244.
- MONTES Y GÓMEZ M., Minería de texto: Un nuevo reto computacional, Instituto Nacional de Astrofísica, Optica y Electrónica (INAOE), México, 2001, Disponible en: http://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf

- ESTRADA E., CMAG: Herramienta para la Construcción Automática de un Mapa Conceptual a Partir de un Texto No Estructurado en Lenguaje Natural, Tesis de Diploma, Facultad de Ingeniería Informática, Instituto Superior Politécnico "José Antonio Echeverría", 2011.
- VILLALON J., CALVO R. A. y MONTENEGRO R., Analysis of a gold standard for concept map mining – How humans summarize text using concept maps, en Proc. of Fourth International Conference on Concept Mapping (CMC'10), Viña del Mar, Chile, 2010, pp. 14-22.
- NOVAK J. D. y GOWIN D. B., Learning How to Learn, New York, Estados Unidos, 1984.
- NOVAK J. D., y CAÑAS A. J., The Theory Underlying Concept Maps and How to Construct Them, Technical Report IHMC CmapTools 2006-01 (Rev 2008-01), Florida Institute for Human and Machine Cognition, Pensacola Fl, 32502, USA, 2008.
- SIMÓN A. J., ROSETE A., PANUCIA K. y ORTIZ A., Aproximación a un método para la representación en Mapas Conceptuales del conocimiento almacenado en textos, con beneficios para la Minería de Texto, I Simposio Cubano de Inteligencia Artificial (SiCIA'04), 10ma Convención y Feria Internacional Informática 2004, C. Habana, Cuba, 2004.
- SIMÓN A., CECCARONI L., ROSETE A., SUAREZ A., y VICTORIA R., A Support to Formalize a Conceptualization from a Concept Maps Repository, en Proc. of the Third Int. Conference on Concept Mapping. Tallinn University, Tallinn, Estonia, 2008, pp. 68-75.
- VALERIO, A. y LEAKE D. B., Jump-Starting Concept Map Construction with Knowledge Extracted from Documents, en Proc. Of Second International Conference on Concept Mapping (CMC'06), Vol. 1, San José, Costa Rica, 2006, pp. 296-303.
- VALERIO, A., LEAKE D. B., y CAÑAS A. J., Associating Documents To Concept Maps In Context, en Proc. Of Third International Conference on Concept Mapping (CMC'08), Vol. 1, Tallinn University, Tallinn, Estonia, 2008, pp. 114-121.