Filtres, H. │Rodríguez, P. R.
Vol.21 No. 1, enero-marzo, 2019 p. 32-44.
ISSN 1562-3297
Avances@ciget.vega.inf.cu

Avances
Centro de Información y Gestión Tecnológica

# Componente para la anotación semántica de información

## *Component for semantic information annotation*

**Hubert Viltres Sala[1], Paúl Rodríguez Leyva[2]**

[1]Máster en Ciencias Informáticas, profesor Asistente, Universidad de las Ciencias Informáticas, La Habana, Cuba, hviltres@uci.cu ; ID: orcid.org/0000-0002-5116-3665
[2]Máster en Ciencias Informáticas, profesor Asistente, Universidad de las Ciencias Informáticas, La Habana, Cuba, pleyva@uci.cu ; ID: orcid.org/0000-0002-2949-0766

**Para citar este artículo / to reference this article / para citar este artigo**

Filtres, H. & Rodríguez, P. (2019). Componente para la anotación semántica de información. *Avances*, 21(1), 32-44. Recuperado de http://www.ciget.pinar.cu/ojs/index.php/publicaciones/article/view/415/1407

**RESUMEN**

Para la recuperación de información se emplean técnicas de la web semántica que enriquecen la información mediante la extracción de conocimiento y realización de anotaciones. En este artículo se propone un componente para realizar anotaciones semánticas en documentos indexados, que utiliza el contenido almacenado en una ontología de dominio para realizar las anotaciones semánticas. En el proceso de anotación semántica se utilizan algoritmos de procesamiento del lenguaje natural que mejoran la calidad de la información almacenada.

Recibido: septiembre 2018
Aprobado: diciembre 2018

Las métricas de precisión y exhaustividad permitieron corroborar la calidad, pertinencia y relevancia de las anotaciones semánticas en la recuperación de información.

***Palabras clave**:* anotación, recuperación de información, semántica, ontología.

**ABSTRACT**

For the recovery of information semantic web techniques are used that enrich the information by extracting knowledge and making annotations. This article proposes a component for making semantic annotations in indexed documents, which uses the content stored in a domain ontology to perform semantic annotations. Natural language processing algorithms are used in the semantic annotation process to improve the quality of the stored information. Accuracy and completeness metrics allowed to corroborate the quality, pertinence and relevance of semantic annotations in information retrieval.

**Keyword:** annotation, information retrieval, semantics, ontology.

## INTRODUCCIÓN

En la actualidad existe más de un billón de páginas web que contienen información en diferentes formatos. La variedad de formatos de información disponible en la web plantea un reto para los Sistemas de Recuperación de Información (SRI) que necesitan ofrecer resultados relevantes a los usuarios cuando realizan una búsqueda. Los SRI utilizan diversos algoritmos en el procesamiento de la información almacenada en grandes colecciones de datos para seleccionar los documentos más relevantes según diferentes criterios.

Los SRI comparan los términos introducidos en la consulta con la información almacenada, mediante diferentes técnicas que determinan la frecuencia de aparición de los términos permitiendo seleccionar los documentos más relevantes según el índice de búsqueda creado. Los métodos tradicionales de procesamiento de la consulta dificultan entender la intención detrás de la pregunta realizada por el usuario y limita la capacidad del SRI de recuperar documentos relevantes (Viltres *et al*., 2018).

En su implementación los SRI utilizan los elementos básicos de Recuperación de Información (RI) y según Salton et al. (1983), Korfhage (1997), Baeza y Ribeiro (1999) y Blázquez (2013) para satisfacer las

necesidades de los usuarios deben rastrear, indexar, procesar y mostrar los resultados de búsquedas más relevantes; a través de diferentes técnicas de análisis de información. En investigaciones realizadas por Baeza y Ribeiro (1999) se plantea que los SRI deben analizar la RI desde el punto de vista computacional y humano, que permitan seleccionar los métodos y algoritmos necesarios en el procesamiento de la información y en la comprensión de la intención de búsqueda de los usuarios para recuperar documentos relevantes.

A la información disponible en la web se le aplica un proceso de extracción y almacenamiento para mejorar la identificación del contexto de los documentos indexados y dotar a los SRI de una potente base de conocimiento que les permita brindar a los usuarios información relevante. En Rodríguez (2014) se define la Extracción de Información (EI) como "cualquier proceso que selectivamente organiza y combina los datos que se encuentran de manera implícita o explícita en uno o más textos".

La EI utiliza el análisis del lenguaje natural para localizar partes específicas de información, mediante la unión y organización de datos de forma implícita o explícita en uno o más textos (Cunningham, 2006; Rodríguez, 2014; Blandón, 2017). Para mejorar el proceso de RI y satisfacer la necesidad de información de los usuarios se utilizan sistemas que extraen fragmentos de texto con un significado relevante (Blandón, 2017).

La información se puede encontrar en diferentes formatos expresada en lenguaje natural y para mejorar su procesamiento se propone realizar anotaciones semánticas a los documentos indexados (Rodríguez, 2014; Blandón, 2017). La anotación de información es el proceso que permite asociar conceptos, relaciones, comentarios o descripciones a un documento o fragmento de texto para mejorar el proceso de inferencia de conocimiento (Oliveira & Rocha, 2013; Rodríguez, 2014; García, 2015 & Vállez, 2015).

En el proceso de anotación semántica se identifican formalmente las relaciones entre conceptos y documentos para enriquecer el contexto de la información. Permite convertir estructuras sintácticas en estructuras de conocimiento, mediante la creación y asignación de etiquetas semánticas a los documentos para dotarlos de significado y mejorar el acceso a la información. En el proceso de anotación se realizan varias tareas para identificar la información y entre las principales se encuentran:

- Reconocimiento de Nombres de Entidades: permite clasificar cada palabra de un documento en un conjunto de categorías predefinidas (nombres propios, lugares, fechas y organizaciones) para disminuir la ambigüedad del lenguaje natural y la existencia de fenómenos lingüísticos como la metonimia, polisemia o elipsis (Rodríguez, 2014). Para la identificación de las entidades

nombradas se utilizan métodos de aprendizaje: supervisado, semisupervisado y no supervisado.

- Extracción de términos: permite la identificación y extracción automática de candidatos a términos (conceptos en dominios específicos) a partir del análisis de grandes colecciones de datos.

La identificación de las entidades nombradas y la extracción de los términos permiten mejorar el proceso de anotación semántica de la información. Para realizar la anotación semántica se utilizan diferentes técnicas y herramientas que permiten extraer el conocimiento de la información indexada.

## MATERIALES Y MÉTODOS

Los SRI para optimizar la RI y seleccionar información relevante utilizan diferentes métodos que combinan ontologías, procesamiento de lenguaje natural, sistemas basados en conocimiento, técnicas de extracción y anotación de conocimiento. La extracción y anotación de información permite generar nuevo conocimiento y mejorar el proceso de recuperación de información relevante y personalizada para los usuarios.

La anotación semántica permite insertar en un documento etiquetas que representan elementos ontológicos (conceptos, relaciones, atributos e instancias). El proceso de anotación puede realizarse manual, automático o semiautomático mediante el análisis, extracción y marcado de la información para enriquecer semánticamente la información (Legaz, 2015; Rodríguez, 2014 & Rosell, 2016).

Anotación manual: permite transformar textos en lenguaje natural a estructuras de conocimientos, mediante el análisis manual de los documentos para identificar las entidades y seleccionar los conceptos que aparecen en el texto y establece las relaciones. Para realizar la anotación manual se necesita disponer de expertos que realicen anotaciones precisas (Sánchez, 2014; Legaz, 2015; García, 2015 & Vállez, 2015).

Anotación automática: el sistema analiza los textos expresados en lenguaje natural mediante un conjunto de reglas y realiza las anotaciones con respecto a las ontologías definidas mediante técnicas de procesamiento natural. Las anotaciones se realizan de forma automática y la calidad depende de las reglas definidas para seleccionar los conceptos más adecuados (Legaz, 2015; García, 2015 & Vállez, 2015).

Anotación semiautomática: permite que los sistemas y los usuarios puedan realizar anotaciones sobre los documentos para evitar las deficiencias del enfoque anterior. El sistema identifica los conceptos y el usuario valida si es el más adecuado para realizar la anotación semántica (Rodríguez, 2014; Legaz, 2015; García, 2015 & Vállez, 2015).

Refieren además los referidos autores que las anotaciones semánticas a los documentos posibilitan que los sistemas informáticos procesen y recuperen información con mayor calidad y relevancia para los usuarios. Entre las principales herramientas para realizar la anotación semántica se encuentran KIM, Armadillo, MnM y SemTag que utilizan ontologías para mejorar la identificación del significado de las palabras.

Las principales propuestas analizadas utilizan el enfoque basado en modelo semántico para realizar el proceso de indexación de información y almacenan las anotaciones separadas del documento original; tiene como principal deficiencia que solo analizan documentos de texto. En la presente investigación se considera que utilizar la técnica automática posibilita realizar anotaciones a los documentos recuperados de la web de forma eficiente. Para realizar las anotaciones se emplean ontologías de dominio como base de conocimiento.

**2.1 Ontologías**

Definen la base de conocimiento que permite desambiguar los términos identificados en el documento indexado. Según Gruber (1993) "una ontología es una especificación explícita de una conceptualización" y permite mejorar la RI. Las ontologías generalmente se utilizan para especificar y comunicar el conocimiento del dominio de una manera genérica y son útiles para estructurar y definir el significado de los términos (Fernández, 2015). La utilización de ontologías permite mejorar el proceso de anotación semántica de documentos al dotar los SRI de una base de conocimiento amplia y diversa.

Entre los principales métodos para mejorar la RI se describe la anotación semántica de información, que permite transformar el texto original obtenido al rastrear la web en un documento enriquecido a partir de incluir diversos términos que mejoran la comprensión de la información almacenada (Rodríguez, 2014; Legaz, 2015; García, 2015; Otero, 2017). Las anotaciones semánticas reducen el espacio entre el lenguaje natural y la representación computacional de la información enlazando los términos de un documento con su representación semántica en la ontología que representa el conocimiento de forma estructurada (Otero, 2017). En un documento un término puede tener varios significados o varios términos pueden referirse a un mismo concepto añadiendo complejidad el procesamiento de la información.

Según Otero (2017) para disminuir la ambigüedad de la información y aumentar la precisión de los resultados de búsquedas, los sistemas de anotación semántica deben explotar el contexto del término analizando su significado en el documento. La utilización de ontologías en el proceso de anotación semántica permite identificar mejor el significado de cada término del documento.

## 2.2. Métodos científicos empleados

Para el desarrollo de la investigación se emplearon métodos científicos generales y mixtos que

ofrecieron resultados de interés relacionados con el procesamiento semántico en sistemas de recuperación de información. Los métodos científicos empleados fueron:

Analítico-sintético: A partir del análisis de los referentes teóricos y la bibliografía relacionada con la investigación se descompone el problema científico en elementos por separado para profundizar su estudio y sintetizarlos en la propuesta de solución.

Hipotético-deductivo: mediante la observación y el análisis del fenómeno en cuestión y a través de reglas lógicas de deducción, se formuló una hipótesis que será comprobada en el proceso de validación del modelo propuesto.

Histórico-lógico: Para determinar los antecedentes, tendencias y particularidades de la recuperación de información con anotación semántica, en función de comprender mejor el objeto de estudio de la investigación.

Análisis documental: se realizaron consultas de libros, artículos científicos, proyectos de investigación para el estudio de los referentes teóricos.

Encuesta: para obtener toda la información necesaria sobre la interacción de los usuarios en los Sistemas de Recuperación de Información.

Experimental: Se aplica con datos provenientes de la base de datos de las búsquedas realizadas. Se emplean las métricas de evaluación (métricas de exactitud predictiva y exactitud en clasificación) debidamente fundamentadas para analizar los resultados. Se establecen indicadores que permiten realizar mediciones de los resultados.

Criterio de expertos: se utiliza el escalamiento de Likert. A partir de su aplicación a expertos se evaluaron los elementos teóricos que fundamentan la investigación.

Técnica Iadov: se aplicó para evaluar y obtener retroalimentación del nivel de satisfacción de los usuarios actuales y potenciales con respecto al modelo propuesto.

## RESULTADOS Y DISCUSIÓN

El componente para la anotación semántica de información posibilita procesar mejor la información y añadir conocimiento al documento almacenado para que sea comprensible por los SRI. Para el correcto funcionamiento del componente se necesita como entrada los documentos almacenados en el SRI y un repositorio ontológico para desambiguar los términos y extraer el conocimiento que será anotado. La propuesta que se describe en la presente investigación (*figura 1*) se fundamenta en investigaciones realizadas por Rodríguez (2014), Legaz (2015), García (2015), Guillén, Lloret y Gutiérrez (2016) y Otero, (2017) para realizar la anotación semántica de información.

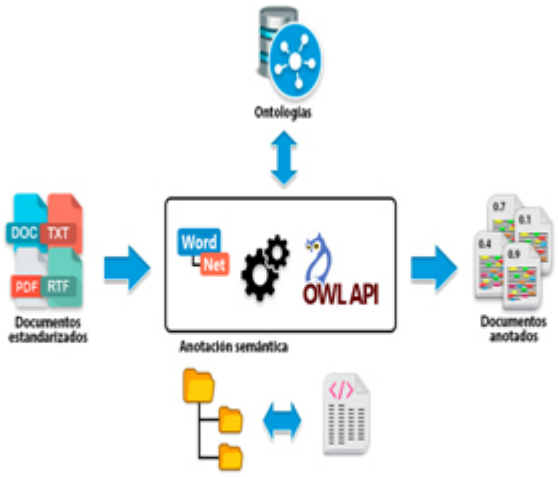


**Figura 1**. Representación y anotación semántica de documentos.

Para realizar las anotaciones semánticas en el componente se implementa un método que obtiene los términos del documento y los asocia con los conceptos representados en la ontología. La propuesta de anotación semántica se basa en la modificación de la metodología de Rodríguez (2014) que consiste en identificar y extraer los términos, asociar los términos a un concepto y almacenar la anotación del documento mediante ontologías de dominio y WordNet. Como paso final se crear un índice semántico que permite mejorar la RI.

En la primera fase del método se obtienen los documentos expresados en lenguaje natural y se procesa la información para identificar los términos que integran el conjunto de palabras clave en el texto. Después de procesado el texto se extraen los términos y se calcula su frecuencia de aparición dentro del texto a través de algoritmos basados en la Ley de Zipf.

El componente obtiene un documento en lenguaje natural y procede a extraer los términos mediante el algoritmo TF-IDF (Salton *et al*., 1983) para crear la lista de términos candidatos a convertirse en conceptos en la ontología de dominio. A la lista de términos candidatos (figura 2 y 3) se le aplica el reconocimiento de entidades nombradas, para disminuir la ambigüedad del lenguaje natural y seleccionar los conceptos que mayor significado aportan.


**Figura 2.** Identificación de los términos en un texto en lenguaje natural


**Figura 3.** Identificación de los conceptos asociados a los términos

Se asigna al término un único concepto que se utiliza para realizar la anotación semántica. Las palabras del texto relacionadas con los conceptos identificados en la ontología son anotadas y se construye un índice semántico para facilitar la RI (figura 4).

Para calcular el índice semántico se aplica el algoritmo TF-IDF (Salton *et al*., 1983) para determinar la frecuencia de aparición de los conceptos de la ontología en el documento (López *et al*., 2016; Álvarez, 2014; Otero, 2017). En el cálculo de la relevancia entre los conceptos y el documento se analizan las relaciones de jerarquía de los conceptos y se aplica una métrica de similitud semántica basada en el camino entre conceptos (Navas, 2016).

La métrica enriquece los conceptos al identificar relaciones directas y asociativas establecidas en la ontología. Al aplicar TF-IDF (Salton *et al*., 1983) a un documento se genera su índice semántico formado por un vector que contiene todos los conceptos con su respectivo valor. Para cada concepto se determina su índice y se almacena en el documento en el formato concepto-valor del índice.


**Figura 4**. Documento con anotaciones semánticas

El componente almacena el documento con el índice semántico calculado y las anotaciones semánticas realizadas (*figura 4*). Para almacenar las anotaciones se aplica el enfoque basado en un modelo semántico que establece que las anotaciones deben ser almacenadas separadas del documento original (Rodríguez, 2014).

Para validar el componente propuesto se diseñó un experimento con 100 documentos indexados en un sistema de recuperación de información. De la colección de documentos seleccionados se identificaron las temáticas (Cultura, Deporte, Cuba, Ciencia y tecnología) y los términos relevantes para realizar la anotación semántica. Los documentos son agrupados por categorías (*tabla 1*) y se le aplica el método de anotación semántica para identificar los términos y los conceptos asociados.

Tabla 1. Colección de documentos por categorías

| Categoría | Título del documento | Cantidad de términos |
|---|---|---|
| Cultura | Alicia Leal Veloz. Dibujante y pintora cubana | 4162 |
| Deporte | Deporte en Cuba | 3819 |
| Cuba | Cuba. Archipiélago del Mar de las Antillas | 500 |

Después de realizada la anotación semántica se diseña un experimento para comprobar los resultados brindados a los usuarios al introducir una consulta en un SRI sin anotación y con el componente propuesto implementado. Para medir el funcionamiento de la propuesta se utilizaron las métricas de Precisión (P) y

Exhaustividad (E) que permiten comprobar la calidad de los resultados obtenidos. Para comprobar la calidad de los resultados obtenidos se diseñan 10 consultas en lenguaje natural, asociadas a diversas temáticas relacionadas con las necesidades de información de los usuarios en un SRI.

La propuesta de validación se sustenta en las investigaciones realizadas por Rodríguez (2014), Guillén, Lloret y Gutiérrez, 2016 y Otero, 2017. Los valores de precisión con anotación semántica obtenidos fueron aceptables similares a investigaciones de referencia en la anotación semántica de documentos. Los resultados obtenidos demuestran que realizar anotaciones a los documentos permite recuperar información relevante. Adicionalmente se realizó una consulta a expertos donde la concordancia demostró un nivel alto de satisfacción con la aplicación del componente propuesto.

Para validar la aplicabilidad del componente para la anotación semántica se aplicó una consulta a expertos mediante el método de escalamiento de Likert. Se definen los siguientes aspectos a valorar por los expertos:

1. Extracción de la información de los documentos

2. Anotación semántica de los documentos

La selección de los expertos se realizó en función del nivel de conocimiento de los candidatos a expertos relacionados con la lingüística y la RI. Se valoró el nivel de conocimiento del idioma español para analizar la semántica de la información. Después de seleccionados los expertos, se sometió a su consideración una encuesta y se procesan los resultados para determinar la valoración de los expertos sobre la aplicabilidad del componente. Para procesar los datos se determina el índice porcentual (IP).

Donde:

MA: muy de acuerdo, DA: de acuerdo, SI-NO: ni de acuerdo ni en desacuerdo, ED: en desacuerdo, CD: completamente en desacuerdo, IP: Índice porcentual, lo que se corresponde con la alternativa A de Likert.

El análisis sobre los resultados obtenidos de aplicar el escalamiento de Likert, al proceso de anotar semánticamente un documento evidenció alta valoración por parte de los expertos en relación a la fundamentación teórica, la utilización de una ontología de dominio y el almacenamiento de las anotaciones semánticas. Se obtuvieron criterios favorables para aplicar el componente en la RI y sugerencias que ayudaron a mejorar el proceso de anotación semántica.

## CONCLUSIONES

El componente propuesto para realizar la anotación semántica de información permite mejorar la relevancia de los resultados de búsqueda brindados a los usuarios en un Sistema de

Recuperación de información, lo que contribuye a la toma de decisiones individuales y de las organizaciones.

Para mejorar la precisión y relevancia de los resultados de búsqueda brindados a los usuarios se realizan anotaciones semánticas en los documentos, lo que exige una preparación de los especialistas vinculados a la gestión de información en las universidades y centros de investigación.

Las métricas de Precisión y Exhaustividad del componente incorporado al modelo, triangulados con los resultados de la consulta a expertos, demostraron que los resultados obtenidos fueron satisfactorios

## AGRADECIMIENTOS

De ser necesario expresar agradecimientos de los autores, se podrán incluir en este acápite, en una muy breve mención, a personas o instituciones que contribuyeron con financiamiento u otro tipo de colaboración.

## REFERENCIAS BIBLIOGRÁFICAS

Álvarez, M.A. (2014). *Detección de similitud en textos cortos considerando traslape, orden y relación semántica de palabras.* (Tesis de Maestría). Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla, México.

Baeza, R. & Ribeiro, B. (1999). *Modern information retrieval*. New York: ACM press. vol. 463.

Blandón, J.C. (2017). *Extracción de instancias de una clase desde textos en lenguaje natural independientes del dominio de aplicación.* (Tesis Doctoral). Universidad Nacional de Colombia - Sede Medellín, Colombia.

Blázquez, M. (2013) Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos. *MEI*, II, 4 (7), 115.

Cunningham, H. (2006). Information Extraction, Automatic. In: Keith Brown, (Editor-in-Chief) *Encyclopedia of Language & Linguistics*, 5(2), 665-677.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), 199–220.

Guillén, A., Lloret, E. & Gutiérrez, Y. (2016). TLH Suite: herramienta para la anotación semántica de información. *RISTI-Revista Ibérica de Sistemas e Tecnologias de la información* (18), 99-103.

Korfhage, R. (1997) *Information Storage and Retrieval*. New York: John Wiley, 1997.

Legaz, M.D. (2015). *Integración de información biomédica basada en tecnologías semánticas avanzadas*. (Tesis doctoral). Universidad de Murcia.

López, T., Troyano, J.A., Ortega, J. & Enriquez, F. (2016). Una

aproximación al uso de word embeddings en una tarea de similitud de textos en español. Sociedad Española para el Procesamiento del Lenguaje Natural, *Revista Procesamiento del lenguaje Natural*, 57, 67-74. ISSN 1135-5948, Universidad de Sevilla, España.

Navas, M. (2016) *Modelo de paráfrasis semántica de similitud de documentos*. (Tesis de maestría). ETSI_Informática. Universidad Politécnica de Madrid.

Oliveira, P. & Rocha, J. (2013). *Semantic annotation tools survey.* En Computational Intelligence and Data Mining (CIDM), IEEE Symposium on. IEEE, pp. 301-307. DOI: 10.1109/CIDM.2013.6597251

Otero, E.N. (2017) Descubrimiento de grafos en datos enlazados para la anotación semántica de documentos. (Tesis doctoral). Universidad de Santiago de Compostela, Galicia, España.

Rodríguez, M.A., Valencia-García, R., García-Sánchez, F. & Samper-Zapater, J. (2014). Creating a Semantically-Enhanced Cloud Services Environment through Ontology Evolution. *Future Generations in Computer Systems*, 32, 295–306.

Rodríguez, M.A., Valencia-García, R., García-Sánchez, F. & Samper-Zapater, J. (2014). Ontology-based annotation and retrieval of services in the Cloud. *Knowledge-Based Systems*, 56, 15-25.

Rodríguez, M.Á. (2014). *Extracción semántica de información basada en evolución de ontologías.* (Tesis doctoral). Universidad de Murcia, España.

Rosell, Y. (2016) UH-WEB: Propuesta de diseño de un CMS semántico para la Universidad de La Habana. (Tesis doctoral). Granada, Granada.

Salton, G., Fox, E.A. & Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM*, 26(11), 1022-1036.

Sánchez, J.L. (2014). *Linked Data para la Generación de Conocimiento Financiero a partir de la Extracción de Información Semiestructurada*. (Tesis Doctoral). Universidad Carlos II de Madrid, España.

Vállez, M. (2015). Exploración de procedimientos semiautomáticos para el proceso de indexación en el entorno web. (Tesis doctoral). Universidad de Barcelona, España.

Filtres, H., Rodríguez, P., Febles, J.P. & Estrada, V. (2018). Procesamiento Semántico de información en sistemas de recuperación de información. *Revista Cubana de Ciencias Informáticas*, 12(1), 102-106. Recuperado de http://scielo.sld.cu/pdf/rcci/v12n1/rcci08118.pdf

Fernández, J.A (2015) Ontología, funciones y discurso en el videojuego. Universidad de Costa

Rica. Revista de Humanidades de Costa Rica, 7(1), 1-21. doi.org/10.15517/h.v7i1.27641