Comentario

La indización en la red semántica: una solución interdisciplinaria

El análisis independiente de las estructuras gramaticales, tal como se le practica a partir del siglo XIX, aísla por el contrario el lenguaje, lo trata como una organización autónoma, rompe sus ligas con los juicios, la atribución y la afirmación. El paso ontológico que el verbo ser aseguraba entre el hablar y el pensar se ha roto; de golpe, el lenguaje adquiere un ser propio. Y es este ser el que detenta las leyes que lo rigen [1]

¿Acaso la Informática con las alternativas para recuperar información en la red hará obsoleta la construcción de tesauros documentales? ¿Tiene vergüenza nuestra disciplina de mencionar el término tesauro documental? ¿Abandonamos el intento de clasificar información y le otorgamos la tarea a los informáticos?

La Bibliotecología utiliza alternativas para indizar los contenidos documentales que luego serán recuperados por los usuarios de información. Ello es y ha sido a través de la construcción de herramientas lingüísticas para facilitar la mediación entre la información depositada o no en la red, a través de múltiples formas y técnicas de indización. El común denominador de éstas es la utilización de palabras simples o sintagmas, validados algunos en un proceso de normalización, como en el caso de los descriptores, palabras clave, términos, nombres de autores, nombres de monumentos, títulos de obras producto del pensamiento humano, u otros impuestos por la necesidad de los propios usuarios, como las folksonomías y su consecuencia, las nubes de palabras.

1 Michel Foucault (2005) *Las palabras y las cosas: una arqueología de las ciencias humanas.* México: Siglo XXI, p. 289

La representación conceptual de un objeto como una escultura, un texto impreso o digital, un discurso, una imagen fija o en movimiento facilita su almacenamiento y recuperación en acervos físicos disponibles en algún lugar o digitalizados en la red. La representación se logra a través de las entidades recuperables del recurso documental, las relaciones entre las entidades y con otros recursos documentales.

La ingente masa de información producida y disponible en medios manuales, electrónicos y una gran parte únicamente en la *World Wide Web* (*WWW*, también conocida como la *Web* o la Red), hace imposible una indización humana y la única forma de recuperar información es la indización automática. Periódicos, televisoras, radios, servicios meteorológicos, comercio, industria, economía generan grandes cantidades de información diaria que son tratados desde su inserción en los medios de transmisión masiva como objetos digitales manipulables para extraer posteriormente los datos o información que existe en ellos. El proceso automático de obtención de datos e información se logra a partir de una extracción terminológica, icónica o de metadatos de documentos audiovisuales.

Entre esta masa de información diversa existe otra de naturaleza más dinámica como los contenidos relacionados con noticias o publicidad que se clasifican con términos muy generales para enviar a usuarios predeterminados y que constituyen lo que se conoce como contenidos sindicados. El formato *RSS (RDF Site Summary)* basado en *XML (eXtensible Markup Language)* permite enviar contenidos a los suscriptores de un sitio en la Red: se trata de una forma de distribución selectiva de la información, donde además ocurren múltiples tipos de envíos, de acuerdo con la información clasificada para distribuir los contenidos sindicados.

Pastor Sánchez, Martínez Méndez y Rodríguez Muñoz[2] destacan un nuevo concepto en la relación hombre-máquina, debido a que se necesita tener en cuenta que el usuario visualizará enormes cantidades de información muy heterogénea. El

2 J. A. Pastor Sánchez, F. J. Martínez Méndes y J, V. Rodríguez Muñoz (2009) Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. Information Research, 4 (14) Available at http://InformationR.net/ir/14-4/paper422.html] (Consultado el 10 de agosto de 2011)

procesamiento del lenguaje natural o tecnologías del lenguaje humano con la creación de herramientas computacionales para realizar la indización automática es ahora una tarea universal y constante buscando emular el cerebro humano para que los sistemas de información permitan recuperar con mayor precisión la riqueza de datos que contienen.

Sin embargo existe también un trabajo humano de indización para preservar y conservar el conocimiento humano más valioso. En este sentido se observa el mantenimiento muy costoso de grandes repositorios físicos y digitalizados de información científica y artística muy calificada, en temas como la medicina, la agronomía, la biología, las matemáticas, la ingeniería, la química, las ciencias sociales, las humanidades y las artes. En la red coexisten sistemas libres en los que no existe calificación de la información contenida, con los repositorios mencionados anteriormente de información acreditada por la evaluación de pares.

Los sistemas de información que se refieren a información calificada se organizan por medio de clasificaciones con la finalidad de asignarle un orden de acuerdo a los contenidos documentales. Se indizan de acuerdo a las diferentes propiedades que poseen para llevar hacia el texto completo a los usuarios que los buscan o para informar al público de las características que los distinguen, entre los que se encuentran los contenidos temáticos apoyados en los encabezamientos de materia, tesauros documentales, vocabularios controlados o listas de palabras clave y los sistemas más poderosos con ontologías.

La tendencia de los contenidos digitalizados en la Red es identificarse con el *URI (Identificador Uniforme de Recurso)*. Para recuperarlos son organizados aplicando modelos de metadatos como por ejemplo el *Dublin Core* que va definiendo las áreas de la descripción donde se expresan elementos formales como autor o título y de análisis de contenido como los temas y los identificadores, asociados a lenguajes de marcado con una sintaxis comprensible por las computadoras. Los objetos, recursos informativos o piezas documentales digitalizadas son dotados de una estructura que opera como puente entre los componentes documentales para identificarlos, evaluarlos y recuperarlos, ya sea al objeto en su totalidad o a cualquiera de sus partes. Otro medio de marcado de palabras claves para recuperar información es el hipertexto, consistente en una colección

de documentos organizados de acuerdo a una red de enlaces que permiten adquirir información en forma no secuencial.

En cuanto al acceso a las entidades, en los sistemas de información más evolucionados ello se logra con lenguajes como el *OWL (Ontology Web Language)*, que permiten comparar y combinar documentos con distintas estructuras sobre identificadores temáticos u otros datos, producto del análisis de elementos paradigmáticos de los diferentes contenidos documentales. Las estructuras de conocimiento son reconocidas mediante la asociación con tesauros, vocabularios controlados, glosarios, terminologías, diccionarios de sinónimos, listados de identificadores de monumentos, organizaciones o componentes de esta clase u orden.

En la red interconectada de transferencia electrónica mundial conocida como la Red Social, los usuarios utilizan *etiquetas* para identificar y categorizar objetos, recursos informativos o piezas documentales para compartir con otros usuarios interesados, constituyendo el fenómeno definido como indización o etiquetado social que a su vez genera las llamadas folksonomías. En definitiva, la indización social y las folksonomías apoyan las tareas de descripción, clasificación y recuperación de contenidos disponibles en la red y ambas se categorizan como expresiones de indización realizada por humanos, porque son estos los que seleccionan los términos que operan como entradas a la información.

Los recursos de la red aparecen con frecuencia acompañados de nubes de etiquetas que describen el contenido documental con palabras o expresiones léxicas ordenadas de forma alfabética extraídas del texto que representan. Pueden ser las cincuenta o cien palabras más utilizadas o las últimas más usadas para describir una colección. Las palabras más frecuentes aparecen con una fuente de tamaño mayor y el resto de las palabras estarán representadas con el tamaño proporcional a su frecuencia. El orden alfabético no permite observar la cercanía semántica de las integrantes del conjunto de etiquetas y pueden aparecer sinónimos o palabras en diferentes idiomas.

En la Red están ocurriendo situaciones que explican en parte el fenómeno de la participación de los humanos en la clasificación de información como la indización social, no sólo para representar el significado conceptual y mantener interconexión

con comunidades de miembros que comparten intereses, sino además conseguir posicionarse en sitios y motores de búsqueda que les otorguen visibilidad. Con respecto a este último aspecto comienzan a aparecer opiniones que alertan acerca de una conducción hacia la información que las máquinas escogen para los humanos. El hombre tiene un tiempo diario acotado para informarse o para producir información y en un mundo tan lleno de la misma es probable que se deje influir por las necesidades definidas por los autómatas indicándoles qué leer y cómo escribir. Por ahora esa dirección vectorial sobre la información es hacia información enciclopédica, comercial o de entretenimiento.

En este año *Eli Pariser*, activista de Internet, publicó un libro que ha levantado polémica sobre las guías hacia la información transferida en la Red. Su título es *The Filter Bubble: What the Internet is Hiding From You* donde manifiesta su percepción de peligrosidad sobre la personalización del perfil de los usuarios de motores de búsqueda como Google; la predicción impactante de *Netflix* (el videoclub *online* que ahora llega a América Latina), *Amazon* y *Pandora* (un servicio en línea de canciones) que definen si algún usuario va a disfrutar una película, un libro o un disco en particular, y hace las recomendaciones apropiadas; también comenta como Facebook actualiza con frecuencia a los amigos de un usuario con los que interactúa más, filtrando a aquellos con quienes tiene menos en común. Los buscadores como *Google* y *Yahoo* obtienen ganancias porque venden a los anunciantes las entradas a sus páginas, mediante la opción de asociar las palabras que el anunciante considere más vendedora de sus productos.

Por supuesto que la transferencia de información de los motores de búsqueda no tiene relación con los sistemas de información, que según la expresión de Schlögl[3] tienen un grado de estructuración, formalización e integración de sus elementos y, como consecuencia, ofrecen además de piezas documentales evaluadas por pares, la libertad de escoger de acuerdo a los intereses personales que dominan una búsqueda en particular.

Desde hace unos cuantos años se ha escrito mucho sobre la Red semántica y la opción de las ontologías para indizar y recuperar

3 C. Schlögl (2005). Information and knowledge management: dimensions and approaches. Information Research 10 (4). http://informationr.net/ir/10-4/paper235.html 268 (Consultado el 20 de agosto de 2011)

información Las ontologías son consideradas como un depósito activo de conceptos donde se establecen conexiones entre los símbolos de una lengua y sus referentes en determinado ámbito de conocimiento que permiten realizar la búsqueda evitando la ambigüedad léxica y/o estructural.

Las ontologías pueden trabajar si la información está previamente estructurada, dotada de *valor semántico*, entendido este como el otorgado por operaciones informáticas a partir de los metadatos. Sin metadatos para crear los contenidos que definan el valor de las palabras o expresiones léxicas en los documentos, las computadoras no pueden interpretar el significado y relacionarlo para recuperar información, o traducirlo o interaccionar para formar el "mapeo conceptual" que necesita la ontología. La Red semántica depende del funcionamiento de las ontologías para operar de acuerdo a los objetivos y metas establecidas con la finalidad de recuperar en la red con *valor semántico*. Por lo tanto hasta que no existan documentos marcados y ontologías formadas a partir de ellos, no habrá Red semántica.

La Bibliotecología durante años se ha dedicado a producir herramientas lingüísticas que permiten traducir los conceptos de una pieza documental después del análisis de la misma a palabras o sintagmas para recuperar los contenidos. Las críticas recibidas al trabajo realizado han sido muchas y constantes porque no siempre las expresiones léxicas representan el lenguaje de los usuarios del sistema de información. Las palabras juegan malas pasadas a los humanos, sobre todo cuando representan contenidos documentales porque el lenguaje es un ilimitado y vasto territorio. Sin embargo es seguro que a las máquinas también se las jugarán.

El afán por reducir a la máxima sencillez algo tan complicado como la representación de contenidos documentales ha llevado a olvidar que en la base del fenómeno está el razonamiento del ser humano que por ahora no puede ser sustituido. Además, el lenguaje natural tiene sus peculiaridades léxicas y terminológicas como la sinonimia, polisemia, las variaciones lingüísticas, el uso de metáforas, las expresiones locales, entre otras. Las ontologías podrán imitar el trabajo humano, pero los campos semánticos definidos en un tesauro documental tendrán que ser la base para comparar el trabajo que haga la máquina con la

mente humana en campos de conocimiento de un valor alto en el nivel social.

El tesauro documental es el producto más evolucionado entre los vocabularios controlados por el énfasis en la estructura conceptual de una disciplina, más que en la léxica, con independencia de un contenido sostenido por palabras simples o sintagmas que hagan visibles los contenidos documentales. La tecnología disponible ha hecho evolucionar al tesauro, integrarlo con otras herramientas como los *Topic Maps* o los mapas conceptuales y aprovechar los espacios dinámicos de la Red para intercambiar ideas y realizar colaboración científica entre múltiples interesados en su construcción. En este sentido se cuenta con wikis y blogs que generan nuevos mecanismos de participación y difusión para el trabajo cooperativo.

Entre los soportes están en desarrollo para estructurar el tesauro documental el *Simple Knowledge Organization System (SKOS)*, basado en *RDF/RDFS, Resource Description Framework/ Resource Description Framework Schema)* para representar esquemas conceptuales que actualmente elabora el *World Wide Web Consortium (W3C)*, pero con el cual ya existen programas. En el mercado comercial se ofrecen programas como el *Visual Thesaurus*, diccionario y tesauro interactivo, cuya finalidad también es crear mapas de palabras relacionadas en forma jerárquica y de asociación y el *Think Map*, programa con una arquitectura modular que permite integrar documentos, bases de datos relacionales, combinar información cualitativa y cuantitativa y recuperar cualquier información diseñada de acuerdo a las tareas específicas a realizar en cualquier tipo de institución

La construcción de ontologías al igual que la de tesauros documentales necesita de una gran especialización, esfuerzo y una gran dedicación porque es un proceso de costo elevado, difícil en la investigación del uso del lenguaje y propenso a errores. La tendencia en los tesauros documentales es hacer énfasis en la estructura conceptual, relacionar los diferentes términos para un mismo concepto y definir las relaciones léxicas para vincular los términos con su significado en un idioma o varios, creando una herramienta lingüística renovada con una mayor flexibilidad para responder a múltiples exigencias de indización y recuperación. De hecho se observa una evolución hacia el tesauro conceptual (Pastor Sánchez , Martínez Méndez y Rodríguez Muñoz, *op. cit.*)

La respuesta a la pregunta inicial es que la obsolescencia del tesauro documental no se ha dado. El avance de una disciplina no se basa en la negación de lo hecho anteriormente, sino en retomar el planteamiento conceptual y analizarlo. Si bien la Bibliotecología tiene un fuerte lazo con la Informática, esta no es la columna vertebral del quehacer disciplinario. No se justifica la cautela que se observa hacia el uso de los tesauros documentales, porque ha habido una evolución en la construcción de la herramienta, respaldada por un trabajo interdisciplinario y continúa siendo útil en la clasificación de información. Entre los juegos del lenguaje en el campo bibliotecológico se denominan con la estructura arbórea definida para el tesauro documental a vocabularios controlados o glosarios. Sin embargo, la fijación y normalización de la terminología de una disciplina es la base de su cuerpo teórico y el uso de una denominación y significado asociados al mismo objeto de estudio son síntomas de una disciplina madura.

Catalina Naumis Peña