EL CONTENIDO DE LOS DOCUMENTOS TEXTUALES: SU ANÁLISIS Y REPRESENTACIÓN MEDIANTE EL LENGUAJE NATURAL

José Antonio Moreiro González
Gijón: Ediciones Trea, 2004.

La expresión y posterior descripción de los conceptos mediante las palabras, a las que acude, como referencia inevitable en los procesos y resultados del análisis de contenido documental, son objeto de la obra más reciente de José A. Moreiro titulada *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural.* Ya sea en utilización libre, o bien controlada mediante un sistema que la legitime para representar con pertinencia los conceptos, la palabra se convierte en fin y medio para el autor, el lector y el analista. Esas y otras consideraciones preliminares sobre la dicotomía fenomenológica de la palabra introducen al lector en el ámbito del análisis y representación documentales del contenido de los textos.

Al tratar sobre el análisis y representación del contenido de los documentos textuales, Moreiro parte de la exigencia de determinación que contienen las estructuras semánticas de esos documentos si queremos conocer su organización y discriminar las partes en que se concentra la información relevante. Apoyándose en la concepción de Báez sobre la intencionalidad del acto de comunicación, se centra en un objeto semántico, la información expresada mediante el lenguaje en articulaciones concretas, que conocemos como documentos.

El análisis de contenido y la intermediación documental son atendidos en el primer capítulo del libro, a la par que se reconocen las barreras existentes para la comunicación directa de los mensajes. Apoyándose en la teoría saussuriana, hace explícita la asociación de interdependencia compositiva de los documentos con sus dos estructuras, la externa y la abstracta, comprensibles para el lector documentalista que, sin actuar como autor del mensaje, se vuelve emisor o viabilizador de la intermediación necesaria entre el mensaje y su destino. Esa práctica reconoce los niveles descriptivos y de análisis que introducen al lector en las fases o momentos del proceso, más específicamente, el reconocimiento, la reducción y la representación. Luego son atendidos los referenciales semánticos que nos llevan a la comprensión de lo que se entiende por texto y por documento, y a la superación de la dicotomía significado/significante a través de la unión de los planos sintáctico, semántico y pragmático del discurso considerado como secuencia de microestructuras. La organización de los textos mediante macroestructuras que representan su significado global es también explorada en este primer capítulo, que finaliza con un análisis de la superestructura e identificación de las partes del texto, que vienen explicadas mediante dos esquemas típicos, el de las narraciones y el de la investigación experimental.

En el segundo capítulo, referido al reconocimiento o lectura de los documentos, se presentan los aspectos identificadores de la lectura con finalidad analítico-documental, contando inicialmente con los procesos inferenciales de ese proceso, que nos permiten comparar lo que ya sabemos con lo traído por el texto, concretándolo en las inferencias *elaborativas,* o proyecciones de nuestros esquemas cognitivos en el texto, en las inferencias *reductivas,* que nos permiten identificar lo esencial del mensaje y,

paralelamente, por medio de las inferencias *lógico-sintácticas,* para comprender cómo está construido el texto, así como para captar informaciones a partir de los conceptos que expresan las palabras, apoyados en las inferencias *léxicas.*

Luego se abordan las estrategias para realizar una adecuada lectura de los textos, siguiendo las dos fases del proceso de reconocimiento del documento, la lectura de situación y la lectura activa. Cuando el proceso de análisis se realiza por personas, el examen inteligente del texto se dirige a los lugares más ricos para obtener información, de acuerdo con las recomendaciones hechas por Anderson, y que puede estar punteado por las cuestiones retóricas claves como las de Lasswell o los criterios de Cicerón presentes en *De oratore,* que de alguna forma están representados en la gramática de casos de Filmore o se corresponden con las facetas del siempre actual método de Ranganathan para el análisis del contenido documental. Consecuentemente y para finalizar el segundo capítulo, el autor plantea unas recomendaciones para la reducción del texto, señalando tácticas, haciendo consideraciones y dando ejemplos de criterios susceptibles de ser aplicados en la práctica.

El proceso de indización y sus resultados, los índices, son objeto del tercer capítulo, cuya primera parte trata sobre el concepto de indización y su procedimiento. En la segunda parte, sobre los criterios y condiciones en que debe hacerse una buena indización, se describen los objetivos que el analista debe perseguir, tales como la especificidad, la relevancia, la exhaustividad y la precisión, destacándose como indicadores de evaluación la entropía, la procedencia de los términos, la profundidad, y el índice de consistencia. A continuación, atiende a la selección y a la asignación de términos, tareas que se inspiran en las necesidades de los usuarios y que se fundamentan, principalmente, en el contexto de las culturas a las cuales pertenecen y en sus experiencias personales. Acomete además los elementos del universo de posibilidades que hay para representar los conceptos seleccionados, desde los vocabularios controlados a los lenguajes libres, así como las posibles circunstancias que llevan a tomar la decisión de incluir un término como representante del contenido original.

Entre esas circunstancias, se encuentran las determinantes de los niveles de indización, conforme se aspire a hacer una propuesta más genérica o más selectiva: clasificación o categorización; indización superficial; indización profunda; indización exhaustiva e indización selectiva. Una extensa reflexión sobre los índices, su naturaleza y categorización, permite que lleguemos a la comprensión del universo categorial de alcance del concepto índice. Se trata de una exposición didácticamente irreprochable, mediante la cual concreta en la obra sus conocimientos sobre el tema, aplicando metodológicamente lo que demuestra en la teoría. Distingue inicialmente los índices libres, basados en palabras del texto, de los índices controlados, basados en conceptos. Entre los primeros, incluye los índices de documentos individuales (nombres propios, geográficos, topográficos y cronológicos), los de colecciones de documentos, entre los cuales destaca los índices esquemáticos, los índices de palabras y nombres, los permutados (tipo KWIC, KWOC y KWAC), los índices de unitérminos y los índices de citas. Por su parte, los índices basados en conceptos incluyen los índices analíticos de libros, revistas y bibliografías, los índices clasificados, los sistemas de índices coordinados de recuperación de la información mediante operadores lógicos y los boletines de índices sistemáticos. Mientras que la quinta parte del tercer capítulo está destinada a la relación entre índices e Internet, analizando cuestiones tocantes a la indización con motores de búsqueda, tanto en lo que se refiere a la recuperación por palabras-

clave, a los metadatos y a la indización de documentos digitales, como a la recuperación conceptual en Internet. La parte final del capítulo está dedicada a la indización automática. Describe modelos extractivos de carácter estadístico y probabilístico, cuyo origen fue coincidente con las primeras tentativas de conjugar la informática y la estadística con la documentación. La esencia del proceso descansa sobre la identificación automática de palabras-clave en el texto por la frecuencia con que aparecen, cuya fundamentación teórica está originada en la ley de Zipf. Nuevas formulaciones de esa Ley originaron otras técnicas de discriminación de los términos, la indización estadística de términos por frecuencia, conocida por las siglas IDF, a *Term frequency, inverse document frequency* (TFIDF), el método N-grams que modifica la ley de Zipf, para el tratamiento de las palabras compuestas y los *Stemmers* que utilizan la frecuencia con que aparecen secuencias de letras en el cuerpo de un texto para extraer la raíz de las palabras. Más allá de esas posibilidades, las relaciones semánticas entre los términos lingüísticos pueden ser establecidas por métodos de agrupamiento y clasificación.

Todavía en relación con la indización automática, se muestran los modelos analíticos de carácter lingüístico, que se derivan del procesamiento del lenguaje natural al que se aplican desde los años 60, bajo el impacto de las teorías lingüísticas y que están fundamentados en procesos analíticos de naturaleza morfológico-léxica, sintáctica, semántica o pragmática. Procedimientos y criterios en favor de un tratamiento inteligente y algunos programas de indización automática que combinan el modelo lingüístico con herramientas estadísticas son descritos para finalizar el capítulo.

Los lenguajes que representan el contenido de los documentos conforman el asunto del cuarto capítulo, en el que se ofrece un panorama abarcador de la variedad y de la evolución histórica de esos lenguajes, cuyo estudio debe atenerse, por un lado, a consideraciones de orden lingüístico y, por otro, a las condiciones funcionales y a las herramientas que se precisan para utilizarlas en contextos y necesidades determinados y, a su vez, también determinantes.

En la primera parte del capítulo se analizan los lenguajes naturales, distintos inicialmente en su modalidad general y científica, así como en su utilización documental que se plasma de forma libre o controlada. En lo tocante a la tipología de los lenguajes documentales, objeto de la segunda parte del capítulo, aparece dispuesta así: el lenguaje libre, representado por listas de unitérminos, listas de palabras-clave y glosarios; los lenguajes controlados, representados por las listas de encabezamientos de materias y tesauros; y por fin, los lenguajes codificados o sistemas de clasificación. Luego, discute la indización hecha utilizando lenguaje libre o mediante lenguajes controlados, presentando sus respectivas características, ventajas e inconvenientes, para después profundizar específicamente en la información representada mediante tesauros. En esta sección del trabajo se tratan las relaciones terminológicas que se producen en los tesauros: las equivalencias, las definitorias, las jerárquicas y clasificatorias, o las de asociación; asimismo las fases típicas en la construcción de un tesauro, desde la terminológica, pasando por la fase documental, por las formas de presentación jerárquica y alfabética, hasta alcanzar la elaboración de índices y la fase de difusión. Si bien no desciende hasta la conformación de los elementos que componen el tesauro, los descriptores, quizás porque el objeto del libro sea estudiar el lenguaje y no sus elementos constitutivos, consultables en cualquiera de las normativas al uso.

Se extiende aún hacia la superestructura del documento tesauro cuando describe el plan global de su presentación, atendiendo también a las tendencias que se siguen

en la actualidad para construirlos, a los tesauros consultables en línea, e incluso hasta llegar a los mapas conceptuales de redes semánticas, método más articulado de representar el conocimiento en el campo de la inteligencia artificial, y más próximo a la función comunicativa del lenguaje. Se destaca su inmersión en la propuesta de los *Topic maps,* como nueva posibilidad de proporcionar acceso a la información digital existente en diferentes redes semánticas, aunque no menosprecie los límites de las posibles aplicaciones de ese nuevo paradigma, incluyendo en el capítulo un ilustrativo cuadro en el que se reflejan las nuevas posibilidades de relación entre conceptos que ofrece un mapa conceptual, un *topic map* o un tesauro. Además de la presentación de un modelo de aplicación basado en las exigencias para la generación y gestión automáticas de tesauros, alcanza a proponer una compleja generación automática de tesauros de verbos, sus fines, posibles aplicaciones y modalidades de organización. Finaliza el capítulo disertando sobre otros esquemas de representación, en especial las ontologías representativas del conocimiento en inteligencia artificial, o aquellas derivadas de las técnicas de ingeniería de software.

El quinto y último capítulo está dedicado al resumen científico, abordándolo desde su naturaleza y finalidad. Luego atiende a las reglas básicas de su representación, por más que el autor reconozca que quien elabora un resumen no es solamente un intermediador, si no también un creador cuya tarea transciende las cadencias preestablecidas. Así, denomina valores a las consideraciones que marcan la pauta en la construcción y redacción del resumen: entropía; pertinencia, coherencia; corrección lingüística o gramatical; y estilo. Tras detenerse en los diferentes modelos de resúmenes, discute, luego, sobre su procesamiento automático, desde los primeros métodos extractivos, pasando por los modelos lingüísticos y cognitivos, hasta llegar a la síntesis de documentos múltiples. Se cierra el capítulo con un análisis de los criterios necesarios para evaluar la elaboración de los resúmenes, básicamente su grado de reutilización y el traslado que hacen de la superestructura del original, así como su calidad técnica, su tamaño, y la densidad y cohesión.

Como resultado de la lectura de esta monografía, puede concluirse que las enseñanzas que trasmite se presentan dentro de un contexto teórico exhaustivamente analizado y discutido, siempre desde una aproximación actual al análisis documental especialmente en lo tocante a la amplitud temática, a la calidad, y al uso crítico del marco referencial.

Ampliamente ilustrado, el texto alcanza el equilibrio necesario para hacerse al mismo tempo profundo e interesante. Además de presentar el estado de la cuestión desde una perspectiva lingüística, el autor atiende a la preocupación de los indizadores por establecer relaciones entre el lenguaje natural y los lenguajes documentales. Puede afirmarse que la obra es bienvenida por su erudición, inovación y estilo.

El libro se integra en una colección de títulos correspondientes al área de la Información-Documentación, todos de buen nivel gráfico, hecho que revela la consistencia de la producción editorial de Ediciones TREA.

Leilah Bufrem. Departamento de Ciência e Gestão da Informação
Universidade Federal do Paraná (Brasil)
bufrem@milenio.com.br