

**Problemas lingüísticos en la
automatización de los sistemas
de clasificación documental**

Fèlix de Moya

Pedro Hípola

E.U. de Biblioteconomía y Documentación

Universidad de Granada

PROBLEMAS LINGÜÍSTICOS EN LA AUTOMATIZACION
DE LOS SISTEMAS DE CLASIFICACION DOCUMENTAL

Félix de Moya

Pedro Hípola

E. U. de Biblioteconomía y Documentación

Universidad de Granada

La clasificación de documentos que se realiza en centros de documentación y bibliotecas suele basarse en el uso de "lenguajes documentales", es decir, según la clásica definición de Courrier (1976), ciertos tipos de lenguajes artificiales que permiten elaborar la representación formalizada de documentos y preguntas documentales que interesan a un grupo de usuarios, a fin de permitir recuperar los documentos que correspondan a las búsquedas.

El afán por elaborar lenguajes documentales con alto grado de condensación y formalización ha producido que en algunos casos las indizaciones que se realizan sean sólo fácilmente descifrables por personal especializado. Esta es la situación que de hecho se suele producir en los centros en los que un usuario sin preparación especial debe enfrentarse con referencias indizadas por medio de la llamada Clasificación Decimal Universal (CDU).

Muy frecuentemente se han formulado críticas a la CDU,

pero en la práctica su uso está fuertemente arraigado en Europa. Un comité de la Federación Internacional de Documentación (FID) se encarga del mantenimiento de la clasificación desde que se editó por primera vez a principios de siglo. Sin embargo, a pesar de que el sistema se utiliza en numerosos centros desde hace más de ochenta años, las indizaciones que se llevan a cabo con la CDU continúan siendo para un gran sector del público una retahíla ininteligible de dígitos y signos de puntuación.

Una solución apropiada para paliar este problema sería poder contar con un sistema informático que actúe como interface entre las notaciones codificadas en CDU y el lenguaje natural del usuario. El sistema deberá actuar por los menos como intérprete en la decodificación de las notaciones. Pero, si pudiera contar con mecanismos propios de autoaprendizaje, podría servir también para realizar codificaciones CDU a partir de informaciones suministradas en términos de lenguaje natural.

En las siguientes páginas se describen algunas características del proyecto encaminado a diseñar un sistema de este tipo, en cuya preparación están trabajando los autores.

La Clasificación Decimal Universal es un sistema de indización documental por temas (subject indexing), ya que los puntos de acceso que sirven para la recuperación de información son las principales ideas que el indizador ha localizado en los documentos; de carácter enciclopédico,

pues está concebida para representar cualquier materia de conocimiento, no se limita a un dominio específico; de estructura precoordinada, debido a que los diferentes aspectos que componen un tema se encuentran ya coordinados antes de que se realice la indización (si bien es verdad que existen procedimientos que permiten poscoordinar descriptores tanto en el proceso de indización de documentos como en la búsqueda documental); y de vocabulario controlado, porque, previamente al análisis documental, se ha establecido un listado de todos los descriptores que se pueden emplear.

Las rutinas que efectúan la decodificación automática de las notaciones se ocupan en un primer momento de gestionar una base de datos, compuesta por todas las notaciones que incluye la edición abreviada castellana de la CDU, junto a sus correspondientes expresiones, y que, desde el punto de vista semántico, está organizada en forma de árbol, pues la estructura jerárquica de la clasificación consta de diez divisiones principales, cada una de las cuales posee sucesivas subdivisiones a las que les corresponde una notación numérica de acuerdo con el sistema decimal. Las cifras de las notaciones se agrupan en cadenas de tres dígitos separadas por un punto.

3	Ciencias sociales
33	Economía
337	Política aduanera
337.9	Sistemas aduaneros

337.91 Tratados de comercio (...)

337.912 Reciprocidad

Presentada una notación al sistema, éste localiza cuál es la expresión que la CDU tiene asignada a esa cadena numérica y entrega al usuario dicha expresión en términos de lenguaje natural. El procedimiento es sumamente sencillo cuando se trata de notaciones que se refieren a un conjunto de conceptos previamente establecidos en la clasificación. Todas las cadenas numéricas simples -incluidas las subdivisiones que genera el símbolo "subdivídase" y excluidas las subdivisiones analíticas introducidas por "punto cero" (.0), a las que nos referiremos más adelante- se incluyan en la base de datos de forma que el acceso es inmediato. Analizada la notación que se presenta al sistema, un primer paso es confirmar si la cadena numérica se encuentra, tal cual, como campo clave de alguno de los registros de la base.

Para poner a punto la base de datos está siendo necesario en muchas ocasiones modificar la redacción de las expresiones que presenta la edición abreviada. Por ejemplo, aunque en esta edición se haya redactado:

532 Mecánica de los fluidos. Hidráulica

532.14 Densidad

nosotros, por nuestra parte, para simplificar el proceso decodificador, de momento preferimos modificar la redacción de expresiones como ésta, consignando en nuestra base:

532.14 Densidad de los fluidos

En el futuro se arbitrará la optimización del sistema, a fin de conseguir que, por medio de instrumentos software -que realicen combinaciones sintácticas elementales en lenguaje natural, o representaciones arbóreas, etc.-, sea posible minimizar el consumo de memoria en la base de datos, eliminando elementos léxicos redundantes.

La decodificación ha de seguir un proceso más completo cuando en la notación presentada al sistema no sólo se incluye una subdivisión de las tablas principales. Entonces pueden encontrarse dos posibilidades distintas:

1. Si es una indización referida a un documento que trata sucesivamente de dos o más temas, aparecerán señalados esos temas por sus notaciones correspondientes, unidas por el símbolo de extensión o de adición.

2. Si se refiere a una "noción temática compleja" en la que se relacionan dos o más notaciones, aparecerá una cadena numérica más elaborada, cuya sintaxis se articula por medio de los símbolos de relación, de nexo indisoluble, los corchetes y por medio de los distintos auxiliares.

Cuando se encuentra el signo de la extensión es porque el tema principal del documento está simbolizado por más de un número de las tablas principales y tales números se encuentran contiguos en las tablas. Aparecen el primer y último número separados por una barra (/). Cabe la posibilidad de que se halle reducido el tamaño de la cadena.

apareciendo consignados del último número únicamente los dígitos que varían con relación al primero, incluido el último punto que es común a los dos.

546.13	Química del cloro
546.14	Química del bromo
546.13/.14	Química del cloro y del bromo

El símbolo de adición (+), dotado de una carga significativa equivalente al de extensión, ha sido empleado por el indizador siempre que las notaciones que se enumeran no se encuentren contiguas en la tabla sistemática. En este caso no se hallará nunca abreviación del último número.

625.31 + 625.35	Ferrocarriles de vía estrecha y ferrocarriles de vía ancha
-----------------	--

Como se puede observar, con la adición y la extensión se verifica un agrupamiento de nociones en cierta forma asimilable a la unión que produce el operador OR inclusivo del álgebra de Boole.

La relación, indicada con el colon (:), señala la interacción entre las nociones representadas por cada uno de los números que se han utilizado. En muchas ocasiones el símbolo sirve para indicar que estas relaciones son ellas mismas tema principal del documento. Se entiende que la relación es bidireccional. Este símbolo resulta un tanto ambiguo, pues no explica la naturaleza de la relación.

31	Estadística
324	Elecciones. Plebiscitos. Referendua
31:324	Estadística electoral

Si aparece el doble colon (::), símbolo de nexa

indisoluble, es porque el indizador deseó dejar constancia de una relación unidireccional:

17	Moral
7	Arte
17::7	La influencia de la moral en el arte

Esta última notación nunca podrá ser interpretada como "la influencia del arte en la moral"

Los corchetes ([]) ofrecen la posibilidad de establecer ecuaciones lógicas en cierta forma asimilables a la del álgebra de Boole.

La aparición de los denominados "auxiliares" permite identificar clasificaciones "por facetas". Los auxiliares comunes, que pueden encontrarse aplicados a cualquier subdivisión de las tablas principales, son los siguientes:

AUXILIAR COMUN	COMIENZA POR
de forma	(0 (paréntesis cero)
de tiempo	" (comillas)
de lugar	(1/9 (paréntesis...)
de raza y nación	(= (paréntesis igual)
de lengua	= (igual)
de p. de vista	.00 (punto cero cero)

Por su parte, los auxiliares especiales se encontrarán empleados sólo en aquellas partes de la CDU en las que está expresamente autorizado su uso. El apóstrofo (') expresa la síntesis de varios conceptos, cuando de dos o más

componentes se crea una noción que tiene su propia individualidad. Se usa, por ejemplo, dentro de la química para referirse a aleaciones. El guión (-) introduce elementos, componentes, propiedades y otros detalles de la materia del número principal de la CDU. Las series de punto cero (.01/.09) identifican clases y subclases de diferentes características: actividades, proceso, etc.

Es posible acumular notaciones que incluyan varios símbolos y auxiliares:

368.42.008 (460) "1973" (058)

368.42	Seguro de enfermedad
.008	En organización
(460)	En España
"1973"	En el año 1973
(058)	Anuario

El orden de los items que componen una notación compleja de CDU no está previamente determinado, si bien se suele recomendar que se redacten las notaciones comenzando por los conceptos más genéricos, descendiendo progresivamente hacia lo más específico.

En cualquier caso, por lo visto hasta ahora queda claro que el estricto formalismo de la sintaxis de todos estos símbolos y auxiliares, en combinación con el control y la precoordinación del sistema, garantiza un grado elevado de univocidad semántica, gracias al cual existen pocas posibilidades de que se produzcan en el proceso decodificador interpretaciones polisémicas. Aprovechando

estas características resulta posible organizar el programa interprete sin que aparezcan excesivos ruidos. La representación arbórea de los conceptos puede generarse con gran facilidad, dada la estructura jerárquica de la CDU. Pero, para ofrecer decodificaciones más "amigables" con el usuario, preferimos asignar a cada posible combinación sintáctica algún sintagma -generalmente de tipo funcional gramatical- o nexos ("y", "en relación con", "durante", "de", etc.) que traduzca en elementos léxicos de lenguaje natural el conjunto finito de conectores sintáctico-semánticos que la CDU conoce.

Por otra parte, el proyecto que se ha concebido pretende partir de la misma base de datos con la que se efectúa la decodificación para desarrollar una aplicación que facilite el proceso interactivo entre usuario y base de datos, a fin de que el sistema pueda generar notaciones CDU utilizando la información léxica que suministre el usuario.

Se ha pensado que la aplicación gire en torno a la gestión de un fichero invertido de formas léxicas y expresiones compuestas, consignadas en términos de lenguaje natural. Cada entrada del fichero contaría al menos con un puntero que señale a su término lexical canónico correspondiente. Los punteros de los lexicales canónicos remitirían a las expresiones contenidas en la base de datos, que a su vez estarían relacionadas con los números propios de la CDU.

En algunos casos, bastará con entregar al sistema una

descripción del tema en lenguaje natural. Desechadas las palabras "vacías" de significado y localizados los términos significativos fundamentales, en muchas ocasiones el programa estará preparado para ofrecer de forma inmediata una o varias posibles codificaciones, de manera que el usuario pueda elegir la que, por su caracterización semántica, parezca pertinente. Así, por ejemplo, al intentar codificar un tema como "enfermedades que sufren las abejas", localizada la expresión "enfermedades de los animales" (591.2), dado que para referirse a los distintos animales no existe una subdivisión preceptuada en las tablas, el sistema deberá estar preparado para relacionar esta noción con la que le corresponde a la forma léxica "abeja" (indizable por 595.799): 591.2:595.799.

Si el sistema no soluciona de esta forma el problema, se comenzará un proceso más completo. Dado que el principio básico de la CDU consiste en que se clasifiquen los documentos en primer lugar de acuerdo con las disciplinas a las que pertenecen, hasta conseguir determinar, de la forma lo más específica posible, el tema principal, en este proceso se deberá, por medio de la interacción con el usuario, a través de preguntas y respuestas, avanzar por las ramas del árbol, hasta llegar al grado de especificidad en la descripción que se desee -y la CDU lo permita-. Esto se haría a través de la identificación de los elementos léxicos que el usuario elija.

Antes de considerar definitiva una expresión -y su correspondiente notación numérica- considerada a juicio del usuario pertinente, el sistema actuará de guía en un rastreo horizontal y vertical por el árbol que constituyen las tablas principales de la CDU para asegurarse de que el concepto identificado se encuentra en el contexto más apropiado de la clasificación y en el grado de especificidad oportuno. Además, la existencia, en muchos lugares de la CDU, de símbolos de reenvío facilita el desarrollo de desplazamientos conceptuales "en red" dentro del árbol de conocimiento CDU, similares a los que son usuales en programas que se encargan de la gestión de tesauros.

Para que todo esto sea posible, se han previsto procedimientos que solventen ciertos problemas frecuentes. En primer lugar, el listado de formas y expresiones léxicas no siempre contendrá todas las palabras que el usuario formule. Por eso, el fichero invertido deberá permanecer abierto siempre, dispuesto a crecer por atribuciones que se desprendan en el proceso de interacción con el usuario. Cada vez que una forma léxica del usuario no se encuentra en el fichero invertido, el sistema, tras las preguntas al usuario, ha de quedar en condiciones de localizar, en primer lugar, formas de significado equivalente -si es posible- y, en segundo lugar, debe incorporar la nueva forma o expresión al fichero y asignarle punteros. De esta manera, si en el fichero inverso no aparece la expresión "viaje de novios", a través de estos procedimientos de "autoaprendizaje" será

posible añadir la expresión al fichero, con unos punteros que remitan a "matrimonio" (392.5) y "viajes" (910). Así el sistema queda cada vez más enriquecido.

Por otra parte, como paso previo a todos los anteriores, se ha tenido en cuenta que un mismo elemento léxico puede representar un tema principal o cualquier otra de las especificidades temáticas "por facetas". Por ello es necesario prever un cuestionario "por facetas" y así se evitará, por ejemplo, que ante un documento de título "Diccionario de filología" se interpretara que el término "diccionario" forma parte del tema principal, o que "La casa de Bernarda Alba" sea indizada como vivienda unifamiliar.

Los problemas que acompañan constantemente la implementación de un programa de estas características son los que se refieren a la interpretación semántica de los términos del lenguaje natural. Un documento titulado "Las enfermedades de las abejas" podría clasificarse dentro de la medicina o de la biología según si estos animales son los agentes o los pacientes de la enfermedad.

