PROBLEMAS LINGUISTICOS EN LA AUTOMATIZACION DE LOS SISTEMAS DE CLASIFICACION DOCUMENTAL

Moya, F.; Hípola, P. «Problemas lingüísticos en la automatización de los sistemas de clasificación documental». Procesamiento del Lenguaje Natural, octubre 1987, 5, págs. 73-85.

Los sistemas de clasificaci¢n de documentos m s utilizados en bibliotecas y centros de documentaci¢n cient¡ficos a veces resultan poco operativos para usuarios poco familiarizados con ellos y a menudo se convierten m s en un obst culo que en una ayuda para la recuperaci¢n de informaci¢n. Esta situaci¢n suele producirse especialmente con los lenguajes de clasificaci ¢n enciclop,dicos que se basan en las denominadas clasificaciones decimales. En este trabajo describe un sistema que est n preparando los autores concebido para actuar como interfaz entre notaciones codificadas en CDU y el usuario. Los procesos de codificaci¢n y decodificaci ¢n autom ticas han de superar los problemas usuales en cualquier programa int,rprete de lenguaje natural.

La clasificaci¢n de documentos que se realiza en centros de documentaci¢n y bibliotecas suele basarse en el uso de "lenguajes documentales", es decir, seg£n la cl sica definici¢n de Courrier (1976), ciertos tipos de lenguajes artificiales que permiten elaborar la representaci¢n formalizada de documentos y preguntas documentales que interesan a un grupo de usuarios, a fin de permitir recuperar los documentos que correspondan a las b£squedas.

El af n por elaborar lenguajes documentales con alto grado de condensaci¢n y formalizaci¢n ha producido que en algunos casos las indizaciones que se realizan sean s¢lo f cilmente descifrables por personal especializado. Esta es la situaci¢n que de hecho se suele producir en los centros en los que un usuario sin preparaci¢n especial debe enfrentarse con referencias indizadas por medio de la llamada Clasificaci¢n Decimal Universal (CDU).

Muy frecuentemente se han formulado cr¡ticas a la CDU, pero en la pr ctica su uso est fuertemente arraigado en Europa. Un comit, de la Federaci¢n Internacional de Documentaci¢n (FID) se encarga del mantenimiento de la clasificaci¢n desde que se edit¢ por primera vez a principios de siglo. Sin embargo, a pesar de que el sistema se utiliza en numerosos centros desde hace m s de ochenta a¤os, las indizaciones que se llevan a cabo con la CDU contin£an siendo para un gran sector del p£blico una retah¡la

ininteligible de d¡gitos y signos de puntuaci¢n.

Una soluci¢n apropiada para paliar este problema ser¡a poder contar con un sistema inform tico que act£e como interface entre las notaciones codificadas en CDU y el lenguaje natural del usuario. El sistema deber actuar por los menos como int,rprete en la decodificaci¢n de las notaciones. Pero, si pudiera contar con mecanismos propios de autoaprendizaje, podr¡a servir tambi,n para realizar codificaciones CDU a partir de informaciones suministradas en t,rminos de lenguaje natural.

En las siguientes p ginas se describen algunas caracter¡sticas del proyecto encaminado a dise¤ar un sistema de este tipo, en cuya preparaci¢n est n trabajando los autores.

La Clasificaci¢n Decimal Universal es un sistema de indizaci ¢n documental por temas (subject indexing), ya que los puntos de acceso que sirven para la recuperaci¢n de informaci¢n son las principales ideas que el indizador ha localizado en los documentos; de car cter enciclop,dico, pues est concebida para representar cualquier materia de conocimiento, no se limita a un dominio espec¡fico; de estructura precoordinada, debido a que los diferentes aspectos que componen un tema se encuentran ya coordinados antes de que se realice la indizaci¢n (si bien es verdad que existen procedimientos que permiten poscoordinar descriptores tanto en el proceso de indizaci¢n de documentos como en la b£squeda documental); y de vocabulario controlado, porque, previamente al an lisis documental, se ha establecido un listado de todos los descriptores que se pueden emplear.

Las rutinas que efect£an la decodificaci¢n autom tica de las notaciones se ocupan en un primer momento de gestionar una base de datos, compuesta por todas las notaciones que incluye la edici¢n abreviada castellana de la CDU, junto a sus correspondientes expresiones, y que, desde el punto de vista sem ntico, est organizada en forma de rbol, pues la estructura jer rquica de la clasificaci¢n consta de diez divisiones principales, cada una de las cuales posee sucesivas subdivisiones a las que les corresponde una notaci¢n num,rica de acuerdo con el sistema decimal. Las cifras de las notaciones se agrupan en cadenas de tres d¡gitos separadas por un punto.

```
\ar110,sa20
\font3\3\tab40\\font6\Ciencias sociales
\font3\33\tab40\\font6\Econom¡a
\font3\337\tab40\\font6\Pol¡tica aduanera
\font3\337.9\tab40\\font6\Sistemas aduaneros
\font3\337.91\tab40\\font6\Tratados de comercio (...)
\font3\337.912\tab40\\font6\Reciprocidad
\font2,ar60,sa0
```

Presentada una notaci¢n al sistema, ,ste localiza cu l es la expresi¢n que la CDU tiene asignada a esa cadena num,rica y entrega al usuario dicha expresi¢n en t,rminos 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 clasificaci¢n. Todas las cadenas num,ricas simples -incluidas las subdivisiones que genera el s¡mbolo "subdiv¡dase" y excluidas las subdivisiones anal¡ticas

introducidas por "punto cero" (@.0@), a las que nos referiremos m s adelante- se incluyen en la base de datos de forma que el acceso es inmediato. Analizada la notaci¢n que se presenta al sistema, un primer paso es confirmar si la cadena num,rica se encuentra, tal cual, como campo clave de alguno de los registros de la base.

Para poner a punto la base de datos est siendo necesario en muchas ocasiones modificar la redacci¢n de las expresiones que presenta la edici¢n abreviada. Por ejemplo, aunque en esta edici¢n se haya redactado:

\ar110,sa20
\font3\532\tab40\\font6\Mec nica de los fluidos. Hidr ulica
\font3\532.14\tab40\\font6\Densidad
\font2,ar60,sa0

nosotros, por nuestra parte, para simplificar el proceso decodificador, de momento preferimos modificar la redacci¢n de expresiones como ,sta, consignando en nuestra base:

\ar110,sa20
\font3\532.14\tab40\\font6\Densidad de los fluidos
\font2,ar60,sa0

En el futuro se arbitrar la optimizaci¢n del sistema, a fin de conseguir que, por medio de instrumentos software -que realicen combinaciones sint cticas elementales en lenguaje natural, o representaciones arb¢reas, etc.-, sea posible minimizar el consumo de memoria en la base de datos, eliminando elementos l,xicos redundantes.

La decodificaci¢n ha de seguir un proceso m s completo cuando en la notaci¢n presentada al sistema no s¢lo se incluye una subdivisi¢n de las tablas principales. Entonces pueden encontrarse dos posibilidades distintas:

\sa+5

1. Si es una indizaci¢n referida a un documento que trata sucesivamente de dos o m s temas, aparecer n se¤alados esos temas por sus notaciones correspondientes, unidas por el s¡mbolo de extensi¢n o de adici¢n.

2. Si se refiere a una "noci¢n tem tica compleja" en la que se relacionan dos o m s notaciones, aparecer una cadena num,rica m s elaborada, cuya sintaxis se articula por medio de los s¡mbolos de relaci¢n, de nexo indisoluble, los corchetes y por medio de los distintos auxiliares.

\sa-5

Cuando se encuentra el signo de la extensi¢n es porque el tema principal del documento est simbolizado por m s de un n£mero de las tablas principales y tales n£meros se encuentran contiguos en las tablas. Aparecen el primer y £ltimo n£mero separados por una barra (@/@). Cabe la posibilidad de que se halle reducido el tama¤o de la cadena, apareciendo consignados del £ltimo n£mero £nicamente los d¡gitos que var¡an con relaci¢n al primero, incluido el £ltimo punto que es com£n a los dos.

\ar110,sa20
\font3\546.13\tab40\\font6\Qu¡mica del cloro
\font3\546.14\tab40\\font6\Qu¡mica del bromo
\font3\546.13/.14\tab40\\font6\Qu¡mica del cloro y del bromo

\font2,ar60,sa0
El s¡mbolo de adici¢n (@+@), dotado de una carga significativa equivalente al de extensi¢n, ha sido empleado por el indizador siempre que las notaciones que se enumeran no se encuentren contiguas en la tabla sistem tica. En este caso no se hallar nunca abreviaci¢n del £ltimo n£mero.
\ar110,sa20
\font3\625.31 + 625.35\tab40\\font6\Ferrocarriles de v¡a estrecha y ferrocarriles de v¡a ancha
\font2,ar60,sa0
Como se puede observar, con la adici¢n y la extensi¢n se verifica una agrupamiento de nociones en cierta forma asimilable a la uni¢n que produce el operador OR inclusivo del lgebra de Boole.
La relaci¢n, indicada con el colon (@:@), se¤ala la interacci¢n entre las nociones representadas por cada uno de los n£meros que se han utilizado. En muchas ocasiones el s¡mbolo sirve para indicar que estas relaciones son ellas mismas tema principal del documento. Se entiende que la relaci¢n es bidireccional. Este s¡mbolo resulta un tanto ambiguo, pues no explica la naturaleza de la relaci¢n.
\ar110,sa20
\font3\31\tab40\\font6\Estad¡stica
\font3\324\tab40\\font6\Elecciones. Plebiscitos. Referendum
\font3\31:324\tab40\\font6\Estad¡stica electoral
\font2,ar60,sa0
Si aparece el doble colon (@::@), s¡mbolo de nexo indisoluble, es porque el indizador dese¢ dejar constancia de una relaci¢n unidireccional:
\ar110,sa20
\font3\17\tab40\\font6\Moral
\font3\7\tab40\\font6\Arte
\font3\17::7\tab40\\font6\La influencia de la moral en el arte
\font2,ar60,sa0
Esta £ltima notaci¢n nunca podr ser interpretada como "la influencia del arte en la moral"
Los corchetes (@[]@) ofrecen la posibilidad de establecer ecuaciones l¢gicas en cierta forma asimilables a la del lgebra de Boole.
La aparici¢n de los denominados "auxiliares" permite identificar clasificaciones "por facetas". Los auxiliares comunes, que pueden encontrarse aplicados a cualquier subdivisi¢n de las tablas principales, son los siguientes:

AUXILIAR COMUN	COMIENZA POR	
de forma	@(0@	(par,ntesis cero)
de tiempo	@"@	(comillas)
de lugar	@(1@/@9@	(par,ntesis...)
de raza y naci¢n	@(=@	(par,ntesis igual)
de lengua	@=@	(igual)
de p. de vista	@.00@	(punto cero cero)

Por su parte, los auxiliares especiales se encontrar n

empleados s¢lo en aquellas partes de la CDU en las que est expresamente autorizado su uso. El ap¢strofo (@'@) expresa la s¡ntesis de varios conceptos, cuando de dos o m s componentes se crea una noci¢n que tiene su propia individualidad. Se usa, por ejemplo, dentro de la qu¡mica para referirse a aleaciones. El gui ¢n (@-@) introduce elementos, componentes, propiedades y otros detalles de la materia del n£mero principal de la CDU. Las series de punto cero (@.01@/@.09@) identifican clases y subclases de diferentes caracter¡sticas: actividades, proceso, etc.

Es posible acumular notaciones que incluyan varios s¡mbolos y auxiliares:

```
\ar110,sa20
\font3\368.42.008 (460) "1973" (058)
\font3\368.42\tab40\\font6\Seguro de enfermedad
\font3\.008\tab40\\font6\En organizaci¢n
\font3\(460)\tab40\\font6\En Espa¤a
\font3\"1973"\tab40\\font6\En el a¤o 1973
\font3\(058)\tab40\\font6\Anuario
\font2,ar60,sa0
```

El orden de los items que componen una notaci¢n compleja de CDU no est previamente determinado, si bien se suele recomendar que se redacten las notaciones comenzando por los conceptos m s gen,ricos, descendiendo progresivamente hacia lo m s espec¡fico.

En cualquier caso, por lo visto hasta ahora queda claro que el estricto formalismo de la sintaxis de todos estos s¡mbolos y auxiliares, en combinaci¢n con el control y la precoordinaci¢n del sistema, garantiza un grado elevado de univocidad sem ntica, gracias al cual existen pocas posibilidades de que se produzcan en el proceso decodificador interpretaciones polis,micas. Aprovechando estas caracter¡sticas resulta posible organizar el programa int,rprete sin que aparezcan excesivos ruidos. La representaci¢n arb¢rea de los conceptos puede generarse con gran facilidad, dada la estructura jer rquica de la CDU. Pero, para ofrecer decodificaciones m s "amigables" con el usuario, preferimos asignar a cada posible combinaci¢n sint ctica alg£n sintagma -generalmente de tipo funcional gramatical- o nexo ("y", "en relaci¢n con", "durante", "de", etc.) que traduzca en elementos l,xicos de lenguaje natural el conjunto finito de conectores sint ctico-sem nticos 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 efect£a la decodificaci¢n para desarrollar una aplicaci¢n que facilite el proceso interactivo entre usuario y base de datos, a fin de que el sistema pueda generar notaciones CDU utilizando la informaci¢n l,xica que suministre el usuario.

Se ha pensado que la aplicaci¢n gire en torno a la gesti¢n de un fichero invertido de formas l,xicas y expresiones compuestas, consignadas en t,rminos de lenguaje natural. Cada entrada del fichero contar¡a al menos con un puntero que se¤ale a su t,rmino lexical can¢nico correspondiente. Los punteros de los lexicales can¢nicos remitir¡an a las expresiones contenidas en la base de datos, que a su vez est n relacionadas con los n£meros propios de la CDU.

En algunos casos, bastar con entregar al sistema una descripci¢n del tema en lenguaje natural. Desechadas las palabras "vac¡as" de significado y localizados los t,rminos significativos fundamentales, en muchas ocasiones el programa estar preparado para ofrecer de forma inmediata una o varias posibles codificaciones, de manera que el usuario pueda elegir la que, por su caracterizaci¢n sem ntica, parezca pertinente. As¡, por ejemplo, al intentar codificar un tema como "enfermedades que sufren las abejas", localizada la expresi¢n "enfermedades de los animales" (591.2), dado que para referirse a los distintos animales no existe una subdivisi¢n preceptuada en las tablas, el sistema deber estar preparado para relacionar esta noci¢n con la que le corresponde a la forma l,xica "abeja" (indizable por 595.799): 591.2:595.799.

Si el sistema no soluciona de esta forma el problema, se comenzar un proceso m s completo. Dado que el principio b sico 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 m s espec¡fica posible, el tema principal, en este proceso se deber , por medio de la interacci¢n con el usuario, a trav,s de preguntas y respuestas, avanzar por las ramas del rbol, hasta llegar al grado de especificidad en la descripci¢n que se desee -y la CDU lo permita-. Esto se har¡a a trav,s de la identificaci¢n de los elementos l,xicos que el usuario elija.

Antes de considerar definitiva una expresi¢n -y su correspondiente notaci¢n num,rica- considerada a juicio del usuario pertinente, el sistema actuar de gu¡a en un rastreo horizontal y vertical por el rbol que constituyen las tablas principales de la CDU para asegurarse de que el concepto identificado se encuentra en el contexto m s apropiado de la clasificaci¢n y en el grado de especificidad oportuno. Adem s, la existencia, en muchos lugares de la CDU, de s¡mbolos de reenv¡o

facilita el desarrollo de desplazamientos conceptuales "en red" dentro del rbol de conocimiento CDU, similares a los que son usuales en programas que se encargan de la gesti¢n de thesauros.

Para que todo esto sea posible, se han previsto procedimientos que solventen ciertos problemas frecuentes. En primer lugar, el listado de formas y expresiones l,xicas no siempre contendr todas las palabras que el usuario formule. Por eso, el fichero invertido deber permanecer abierto siempre, dispuesto a crecer por atribuciones que se desprendan en el proceso de interacci¢n con el usuario. Cada vez que una forma l,xica 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 expresi¢n al fichero y asignarle punteros. De esta manera, si en el fichero inverso no aparece la expresi¢n "viaje de novios", a trav,s de estos procedimientos de "autoaprendizaje" ser posible a¤adir la expresi¢n al fichero, con unos punteros que remitan a "matrimonio" (392.5) y "viajes" (910).

As¡ el sistema queda cada vez m s enriquecido.

Por otra parte, como paso previo a todos los anteriores, se ha tenido en cuenta que un mismo elemento l,xico puede representar un tema principal o cualquier otra de las especificidades tem ticas "por facetas". Por ello es necesario prever un cuestionario "por facetas" y as¡ se evitar , por ejemplo, que ante un documento de t¡tulo "Diccionario de filolog¡a" se interpretara que el t,rmino "diccionario" forma parte del tema principal, o que "La casa de Bernarda Alba" sea indizada como vivienda unifamiliar.

Los problemas que acompa¤an constantemente la implementaci¢n de un programa de estas caracter¡sticas son los que se refieren a la interpretaci¢n sem ntica de los t,rminos del lenguaje natural. Un documento titulado "Las enfermedades de las abejas" podr¡a clasificarse dentro de la medicina o de la biolog¡a seg£n si estos animales son los agentes o los pacientes de la enfermedad.