CLASIFICACIÓN AUTOMÁTICA MEDIANTE LA LA CDU CON EL PROCEDIMIENTO EN CADENA

Rosa San Segundo Manuel
Universidad Carlos III de Madrid
España
rsan@bib.uc3m.es

Resumen Se entiende por clasificación automática el proceso de agrupar según el contenido las referencias de los documentos o bien los propios documentos electrónicos. Este proceso se realiza mediante programas capaces de comparar términos empleados utilizados en el documento. E incluso hay otras formas automáticas de clasificación que emplean procedimientos automáticos para generar clases de documentos También existen los sistemas automáticos de que extraen términos de indización, sin embargo la clasificación automática empleando sistemas de clasificación va a resultar mas compleja.
Con el fin de clasificar automáticamente usando sistemas de clasificación como la CDU es imprescindible un índice alfabético de materias y términos del lenguaje natural asociados a cada notación de la CDU, de forma tal que todas las notaciones tendrían vinculadas términos del lenguaje natural y encabezamientos de materias. Este método empleado para la clasificación automática es el procedimiento de cadena ideado por Ranganathan. El procedimiento en cadena supone la construcción de encabezamientos alfabéticos y términos del lenguaje natural, asociados a una Clasificación sistemática de estructura jerárquica y con notación preferentemente numérica. La clasificación automática mediante el procedimiento en cadena supone el proceso donde de los términos alfabéticos asociados a un documento obtendríamos, de forma automática, la asignación de su notación correspondiente y también el proceso inverso. El sistema de clasificación que mas se ha aplicado a esta forma de clasificación automática es la Clasificación Decimal de Dewey electrónica la 21ª ed. que incluye asociados encabezamientos de materia (LCSH) con los números clasifícatenos. Sistemas de clasificación muy relevantes, como la CDU, carecen de la edición electrónica de sus tablas con el procedimiento en cadena. Sin embargo ya hay varios ejemplos prácticos de clasificación automática empleando la Clasificación Decimal Universal y además se emplea para la clasificación automática de los recursos de Internet, se trata de sistemas de indización y clasificación automáticos para Internet que permiten la búsqueda y la navegación integradas, y contienen términos de índice, adjuntos a números de clasificación, en alemán, inglés y francés. Un hecho muy importante para este desarrollo de la CDU es el elemento del Dublin Core titulado *Materia,* que sugiere la utilización, de forma conjunta, de encabezamientos de materias y sistemas de clasificación y ha incluido la *Clasificación Decimal Universal.* Así un instrumento clásico como la CDU se presenta como un novedoso instrumento para la clasificación y organización de la información electrónica.

1. Distintas concepciones de la Clasificación automática

En un sentido general se entiende por clasificación automática el proceso de agrupar según el contenido las referencias de los documentos o bien los propios documentos electrónicos. Este proceso se realiza mediante programas capaces de comparar términos empleados utilizados en el documento. A partir de éstos, calculan el grado de semejanza de las palabras claves utilizadas para indizar el documento o el

grado de semejanza de los términos. La clasificación automática presenta algunas semejanzas con la indización automática ya que, dependiendo de la sofisticación del sistema, ambas utilizan métodos similares de procesamiento del lenguaje natural, con el fin de aislar términos de documentos con texto completo, y pueden recoger y analizar la aparición y la frecuencia de términos utilizados en un documento en particular, contrastándolas con la aparición y frecuencia de los mismos dentro del total de la base de datos. E incluso hay otras formas automáticas de clasificación que emplean procedimientos automáticos para generar clases de documentos o clases de términos mediante técnicas denominadas de cluster, que, de forma automática, crean conjuntos de documentos que tienen términos, autores, citas y otros aspectos en común.

En este trabajo vamos a diferenciar entre la clasificación automática y una parte de ella que seria la asignación automática de notaciones, y en concreto en relación con la CDU Mientras que para la mayoría de los sistemas automáticos de indización puede ser suficiente extraer los pertinentes términos de indización, la clasificación automática va resultar mas compleja. La asignación automática de notaciones consistiría en una vez que tenemos los documentos indizados manual o automáticamente, para clasificar o asignar una notación de forma automática sería necesario emplear las técnicas del procedimiento en cadena. Así, con el fin de clasificar automáticamente usando sistemas de clasificación como la CDU resultaría impres-cindible un índice alfabético de materias y términos del lenguaje natural asociados, cada notación de la CDU, de forma tal que todas las notaciones tendrían vinculados términos del lenguaje natural y encabezamientos de mate-rías.

2. Técnicas del procedimiento en cadena para asignación automática de notaciones y de encabezamientos de materias

Un método muy eficiente para la clasificación automática es mediante el empleo del procedimiento de cadena ideado por Ranganathan, donde el índice elaborado de esta forma recibe el nombre de índice de cadena. Si bien es teóricamente posible generar semejante índice de forma aitomática, generalmente, se realiza con intervención humana con el fin de evitar posibles equívocos. El procedimiento en cadena fue un concepto ideado por Ranganathan en su obra [1] *"Classified catalogue with additional rules for dictionaty catalogue code"*. Este procedimiento en cadena supone la construcción de encabezamientos alfabéticos y términos del lenguaje natural, asociados a una Clasificación sistemática de estructura jerárquica y con notación preferentemente numérica. O sea, se trata de asociar y derivar encabezamientos de materia de los números clasificatorios y viceversa. De forma que todos y cada uno de los números y notaciones de un sistema, van a aparecer ligados o encadenados a uno o varios encabezamientos de materias. El procedimiento en cadena puede aplicarse mejor en cualquier sistema de clasificación estructurado jerárquicamente y constituido mediante notación sistemática.

Puede decirse que, los prolegómenos de esta metodología se encuentran y provienen del índice temático que Dewey introdujo, por vez primera, en su sistema clasificatorio. Posteriormente se ha abordado esta problemática en diversos momentos, como la propuesta de Cutter en su *"Rules for dictionary Catalogue"* [2], por Kaiser, Coates y finalmente Ranganathan conceptualizó el procedimiento en cadena como se entiende en la actualidad.

La clasificación automática es un proceso diametralmente opuesto al realizado de forma manual, y requiere, previamente, la indización manual o automática. Si esta se realiza de forma automatizada su método mas general es por extracción automática o mecánica de términos, operando los siguientes pasos en esta extracción automática [3]:

1. Selección previa y señalización, en los documentos digitales, de las partes del texto con mayor contenido semántico o segmentos ricos de información (como la primera y última frase, las conclusiones, el resumen, el índice y otros).
2. Eliminación de palabras y términos no significativos o vacíos como artículos, preposiciones, conjunciones y otros.
3. Selección de raíces de palabras (Ejemplos: biblio-; españ-...) conver-

giendo todos los términos derivados a su raíz común.
4. Selección de frases con frecuencia significativa.
5. Ordenación de los términos extraídos mas empleados hasta los menos empleados.
6. Determinación de un punto de corte de los términos más empleados desechando los poco empleados.
7. Análisis de la frecuencia de aparición de términos empleados, o sea relacionar los términos seleccionados anteriormente considerando la frecuencia en que dichos términos aparecen la base de datos.
8. Selección de cadenas de palabras o términos que siempre aparecen asociados (por ejemplo, Organización del conocimiento)
9. Análisis de la frecuencia absoluta y de la frecuencia relativa de los términos y la combinación de los dos criterios añadiendo además criterios posicionales y sintácticos.
10. Revisión manual posterior de la indización automatizada.

Este proceso de extracción automática es un proceso que los ordenadores ejecutan con bastante acierto y, en la generalidad de los casos, suele ser bastante coherente.

La indización y clasificación automáticas tratarían de asignar un perfil de palabras, frases y de estos términos se asocia, de forma automática, encabezamientos de materias, para finalmente asignar notaciones, que estarían asociadas a toda la terminología alfabética y a los encabezamientos de materias. Es decir, se trata de asociar a los términos obtenidos una notación, o sea un vocabulario controlado. La remisión de un término del texto a un término del vocabulario controlado puede realizarse mediante un vocabulario puente. La clasificación automática mediante el procedimiento en cadena supondría el proceso donde de los términos alfabéticos asociados a un documento obtendríamos, de forma automática, la asignación de su notación correspondiente. De esta forma podremos consultar al mismo tiempo una enumeración exhaustiva de la notación del sistema clasificatorio y de los términos que se relacionan temáticamente con esta notación. Así se posibilita la creación de un catálogo con referencias cruzadas entre la clasificación sistemática y la lista de encabezamientos de materia, o sea, se puede crear una organización sistemático-alfabética, donde el primer orden de presentación será el sistemático y el segundo el alfabético.

E incluso, también pude realizarse el proceso de forma inversa, la metodología práctica de la indización automática mediante el procedimiento en cadena supone que tenemos un libro o documento que tratamos de clasificar. En un primer momento ubicaremos la temática del libro dentro de la estructura jerárquica de un sistema clasificatorio, insertándolo dentro de un número de clase con notación del tipo que fuere, al mismo tiempo esta notación tendrá asociados varios términos de alfabéticos que se asocian de forma automática.

Hay dos Sistemas de clasificación, o tablas clasificatorias, que se han publicado en versión electrónica donde cada notación del sistema lleva asociada varios encabezamientos de materias precoordinados y también toda la terminología alfabética relativa a esa notación, o sea mediante el procedimiento en cadena: La Clasificación Decimal de Dewey DDC electrónica o DDC 21 que incluye la conjunción electrónica de la DDC y la LCSH (Subjet Headings Libray of Congress, Lista de encabezamientos de Materias de la Biblioteca del Congreso de Washigton), y solventa muchas deficiencias de la DDC y de la LCSH. Su plasmación emplea el hipertexto como base material para los términos encadenados o asociados. No emplea solo el índice alfabético de la Clasificación de Dewey, pues este índice, así como el de la CDU no han sido ideados como términos de acceso, por ello hace uso de una lista de encabezamientos conformada como tal. Esta Clasificación Deci-mal de Dewey electrónica trata de unificar la 20ª ed. de la Clasificación de Dewey con los LCSH, asociando estos encabezamientos de materia con los números clasificatorios. Igualmente la ultima edición de las tablas de Clasificación de la Biblioteca del Congreso de Washington, se han editado de forma electrónica con el procedimiento en cadena, o sea cada notación de las tablas lleva asociada varios encabezamientos de materias de la Biblioteca del Congreso (LCSH) y la terminología alfabética relativa a esa notación

que está en las propias tablas. Sin embargo otros sistemas de clasificación muy relevantes, como la CDU, carecen de la edición electrónica de sus tablas con el procedimiento en cadena.

Hay varios ejemplos prácticos de clasificación automática empleando la Clasificación Decimal Dewey para Windows [4]. Destacan los proyectos actuales de indización y clasificación automática, auspiciados por OCLC, donde a un término extraído del documento, se le asocia una notación o un encabezamiento de materia precoordinado, estos términos alfabéticos se han de relacionar con las notaciones jerárquicas, y van a ser de mayor utilidad si emplean lenguajes precoordinados frente a los postcoordinados, como el proyecto Scorpio, o el proyecto Mantis que necesitan de la intervención humana en su revisión final. El procedimento en cadena también se emplea en el Sistema DORS, (Dewey Online Retrival System), este sistema DORS trata de hacerse un espacio dentro de OCLC y comparar su eficacia frente al catálogo tradicional, ya que ofrece encabezamientos encadenados a la Clasificación Decimal de Dewey asignados automáticamente.

El procedimiento en cadena tiene otras distintas vertientes y matices en sus diversas formulaciones así Bhattachayya propuso el sistema POPSI (Postulated-based Permuted Subject Indexing), proponiendo casi una clasificación verbal que contiene términos muy específicos en el ultimo eslabón jerárquico. En la India se aplica el procedimiento en cadena en múltiples repertorios, como el de Publicaciones periódicas, índice de prensa, Literatura en las bibliotecas indias. El procedimiento en cadena tiene gran actualidad, pues son numerosos los repertorios bibliográficos que ya están así organizados, como la Bibliografía Nacional Inglesa desde 1950, este sistema es denominado PRECIS .El procedimiento en cadena también tiene su aplicación práctica en otros contextos distintos. Así en la antigua Unión soviética emplearon ya el procedimiento en cadena en la década de los años 70 cuando la antigua Biblioteca Lenin de Moscú, hoy denominada Biblioteca Central Estatal, asoció un índice alfabético a, la antes denominada Clasificación Biblioteco-blibliográfica la BBK, hoy LBC.

La automatización de las bibliotecas ha transformado muchas tareas tradicionales. Los catálogos en línea son más eficientes en cuanto a la función y mejores en cuanto a rendimiento en multitud de aspectos, sí los comparamos con sus predecesores manuales y en lo que hace referencia a la clasificación automática las mejoras son notables ya que el promedio de encabezamientos de materias asignados de forma manual es uno o bien dos, sin embargo en una indización automatizada el número de encabezamientos asignados puede resultar 30 ó 40 y además ser términos pertinentes.

3. Propuesta del procedimiento en cadena para la CDU

Los sistemas de clasificación están reconocidos como una extraordinaria herramienta para la indización, clasificación y recuperación temáticas ya que pueden resumir todo el contenido temático de un documento en un símbolo de clasificación: una secuencia de números o letras o combinaciones de ambos que no dependen de un idioma y que, normalmente, son fácil y rápidamente manejados (indexar, clasificar, recuperar) por un ordenador. Estos símbolos puede ser utilizados como un apoyo al que pueden ir anexos vocabularios controlados en diferentes idiomas y de diferentes clases con el procedimiento en cadena, y así convertir la clasificación en un lenguaje de permuta. Además la notación de la clasificación podría utilizarse para hacer mas fácil la visualización y navegación del contexto jerárquico.

La elaboración de un vocabulario con los términos alfabéticos de la CDU es una tarea que no es fácil de realizar porque numerosos conceptos no están explicados con detalle en las propias tablas de la clasificación, por más que puedan ser expresados con notaciones de la CDU, y en nuestro ámbito concreto se complica por la inexistencia de una lista completa de encabezamientos de materia en castellano.

La CDU es idónea para formular con sus tablas el procedimiento en cadena, y editarse en versión electrónica como la CDD o la CBC, ya que tiene una estructura jerárquica y sistemática y comprende y una notación capaz de saltar las barreras lingüísticas. Además puede servir como fuente de vocabulario de lenguaje natu-

ral. Puede utilizarse como herramienta para formar un listado alfabético, una vez completado el listado, deberán asociarse un listado de encabezamientos de materias percoordinados, al igual que la Clasificación de Dewey o la Clasificación de la Biblioteca del Congreso de Washington en sus ultimas ediciones electrónicas. Estos sistemas de clasificación con asociaciones a las listas de encabezamientos de materias (LCSH - Encabezamientos de Materias de la Biblioteca del Congreso, en inglés; RAMEAU - Repertoire d'Autorité-Matiére Encyclopédique et Alphabetique Unifié, en francés, etc.) y tesauros como MeSH (Encabezamientos de Materias Médicas) son un buen ejemplo a imitar, a pesar de la falta de una buena lista de encabezamientos de materias editada en castellano. Realizar este instrumento sería fundamental, pese a hacerlo en nuestra lengua no se perdería la característica terminológica multilingüe de la CDU ya que se perdería que se comprendería también de repre-sentación numérica con el fin de preservar el significado original de cada notación.

Pese a no existir esta herramienta en castellano, si se ha trabajado, en otras lenguas, con la asociación de notaciones a terminología del lenguaje natural, por ejemplo como el sistema alemán GERHARD (Recopilación, Recuperación y Directorio Automatizado Alemanes http://www.gerhard.de), donde las notaciones de la CDU representan disciplinas, materias o conceptos, y se han completado con las descripciones verbales que aparecen en las propias tablas y con una lista de encabezamientos de materias.

4. Futuro de la CDU para Clasificación automática

La clasificación automática de los recursos de Internet es cada vez más importante La aplicación práctica, en la actualidad, del procedimiento en cadena tiene, también, su plasmación en Internet. Proyectos como DESIRE, NORDIC, SCORPION, CORC, GERHARD, etc., han facilitado el contexto en el que los sistemas de clasificación pueden desempeñar un papel muy relevante en Internet. El proyecto Scorpion de OCLC se basa en la asignación automática de encabezamientos de materia mediante la Clasificación Decimal Dewey, donde toda la terminología alfabética que se alberga en las tablas clasificatorias, junto con los encabezamientos de materias de la Biblioteca del Congreso de Washington ligados a cada notación, lo que implica la posibilidad de asignar encabezamientos de materia y notaciones de forma automática. Este proyecto no ha sustituido totalmente al clasificador sino que se presenta como una herramienta de gran ayuda para este. (http://pur1 .oc1c.org/scoTpion)

También existen en Internet servicios o portales de acceso a la información electrónica clasificados por medio de la CDU como SOSIG (Portal de Acceso a la Información de Ciencias Sociales) y NISS (Información Nacional sobre Software y Servicios). El WWW es portal de acceso que ha experimentado con la clasificación automática, Subject Tree'de bases de datos de WAIS (Servidor de Información de Área Ancha), forma parte del Nordic WAIS/World Wide Web Project
http://www.ub2.lu. se/auto newfl JDC.html. E incluso el sistema GERHARD, es un sistema de indización y clasificación completamente automático para Internet que permite la búsqueda y la navegación integradas, y contiene términos de índice, adjuntos a números de clasificación, en alemán, inglés y francés.

Asimismo, destaca el empleo de la CDU para clasificación automática el "UDCZ-Lexicon" del sistema de la ETH, que se crea automáticamente desde el texto de la base de datos de la ETH, o sea se trata de páginas de Internet en lengua alemana recopiladas en una base de datos, tomando expresiones lingüísticas que normalmente aparecen en páginas de Internet que van a ser clasificadas. Ello comprende un análisis morfológico previo de cada palabra de la entrada de la CDU y la reducción de cada una a su raíz, como antes vimos. El texto que ha de ser clasificado es objeto de un análisis de texto antes de que las palabras que contiene se cotejen con el UDCZ-Lexicon y se les asigne una notación.

Finalmente, un hecho muy importante para este desarrollo de la CDU es el elemento del Dublin Core titulado *Materia,* que sugiere la utilización, de forma conjunta, de encabeza-

mientes de materias y sistemas de clasificación, como los *Encabezamientos de Materias de la Biblioteca del Congreso,* los *Encabezamientos de Materias Médicas,* la *Clasificación de la Biblioteca del Congreso,* la *Clasificación Decimal de Dewey* y, también se ha incluido, la *Clasificación Decimal Universal.* Esta recomendación se hizo oficial en junio de 2000 [5].

Nos encontramos en un momento de gran incidencia de las tecnologías de la información electrónica donde la clasificación automática va a presentarse como un proceso fundamental. La aplicación de una herramienta como la CDU electrónica que incluya encabezamientos de materias asociados se presenta como un instrumento fundamental, en tanto que es un sistema accesible, en todo lugar, con listas de encabezamientos de materia con carácter multilingüe. Se trata, en definitiva, de aplicar uno de los tradicionales sistemas de clasificación que han tenido, y tienen, una importancia fundamental. Hoy, este tradicional sistema de clasificación se presenta como un novedoso instrumento para la clasificación y organización de la información electrónica. Instrumentos clásicos de la clasificación documental son ahora la base teórica y metodológica de la clasificación automática. Viejos instrumentos, por su conveniente y peculiares características, son imprescindibles para la implantación de nuevos conceptos, normas y formatos.

Referencias

[1] RANGANATHAN, S. R. "Classified catalogue with additional rules for dictionaty catalogue code". —Madras, Londres, 1951.

[2] CUTTER, Ch A. "Rules for dictionary Catalogue". ~4 a ed. -Washingthon, [s.n.], 1904.

[3] LANCASTER, F.W. Indización y reúmenes. -Buenos Aires : E B Publicaciones, 1996.

GIL LEIVA, I. La automatización de la indización de documentos. -Gijón,: Trea, 1990

[4] SAN SEGUNDO MANUEL, Rosa. Indización en cadena y su aplicación práctica. En: IV CONGRESO ISKO-España EOCONSIN 99: Organización y Representación del Conocimiento en sus distintas perspectivas: su influencia en la recuperación de la Información. 53-59.

[5] McILWAINE, I.C. The Universal Decimal Classification. a guide to its use. - The Hague : UDC Consortium, 2000

BIBLIOGRAFÍA

British Catalogue ofMusic, 1957-

CANTOS GÓMEZ, P; MARTÍNEZ MÉNDEZ, F. J.; MOYA MARTÍNEZ, G. *Hipertexto y Documentación.* —Murcia : Universidad, 1994 ; 48.

CDU. *Clasificación Decimal Universal Abreviada I* Adaptada por Rosa San Segundo.--Madrid: Aenor, 2001.

CLASIFICACIÓN Decimal Universal en hipertexto.-Madrid :AENOR: 1999.

COATES. *Subject catalogues. Headings and structure.* London, 1960.

CUTTER, Ch A. *"Rules for dictionary Catalogue".* —4 a ed. —Washingthon, [s.n.], 1904.

DEWEY, Melvil. A classification and subject Índex for cataloging and arranging and pamphlets of a library.- Amherts : Mass, 1876.

DEWEY for windows [Archivo de ordenador] / by OCLC On line Computer Library Center.— OCLC, 1998.

ENCABEZAMIENTOS de materia : normativa para su redacción. —Madrid : Biblioteca Nacional, 1991.

FARRADANE, J.E.L. *A scientific theory of classification and indexing.* En: Journal of Documentation, 6: 83-99. y 8: 73-92.

GIL LEIVA, I. *La automatización de la indización de documentos.* —Gijón, : Trea, 1990

KAISER, J. *Systematic indexing.* —London, 1911.

LISTA de encabezamientos de materias para las Bibliotecas Públicas. —Madrid : Dirección General del Libro, Archivos y Bibliotecas, 1995.

LANCASTER, F.W. *Indización y resúmenes.* -Buenos Aires : E B Publicaciones, 1996.

McILWAINE, I.C. *The Universal Decimal Classification. a guide to its use. —The Hague : UDC Consortium, 2000.*

MISHINAJ. P. *Alphabetic subject index to the classified catalog.* —Moscú : Kniga, 1981.

RANGANATHAN, Shiyali Ramamrita. *"Classified catalogue with additional rules*

for dictionaty catalogue code". —Madras, Londres, 1951.

SAN SEGUNDO MANUEL, Rosa. *Sistemas de Organización del Conocimiento. La Orga-nización del Conocimiento en las bibliotecas españolas*. —Madrid : Universidad Carlos III de Madrid; Boletín Oficial el Estado, 1996.

SAN SEGUNDO MANUEL, Rosa. *Indizacion en cadena y su aplicación práctica*. En: IV CONGRESO ISKO-España EOCONSIN 99: Organización y Representación del Conoci-miento en sus distintas perspectivas: su ínfluen-cia en la recuperación de la Información. 53-59.

SVENONIUS, Elaine; LUÍ, Sonqqiao; SU-RAHMANYAN, Bhagi. *Automation of Chain indexing*.