CONFECCIÓN DE CATEGORÍAS Y RECUPERACIÓN DE LA INFORMACIÓN EN INTERNET

Autores: Mònica Bechini i Tristany
Documentalista
mbechini@teleline.es

Ferran Burguillos Martínez
Universitat de Barcelona.
Facultad de Biblioteconomía y Documentación
Universitat de Vic.
Facultad de Ciencias humanas, traducción y documentación.
Departamento de Biblioteconomía y Documentación
fburg@fbd.ub.es

Albert Díaz Mota
Documentalista *daVinci Consulting Tecnológico*
Albert.Diaz@dvc.es

Resumen: Se describen las características de los sistemas de categorías en el entorno web, así como su elaboración y las técnicas de categorización. Los sistemas de categorías son herramientas de visualización y búsqueda de información en el medio electrónico mediante el *browsing*. En el proceso de categorización se proponen dos métodos. En el primero, la estructura del sistema diseñado será la misma que utilizaremos para clasificar los ítems y la que verá el usuario. El segundo método, requiere el uso de estructuras jerárquicas elaboradas *a priori*, y haber clasificado con ellas los ítems; la categorización, entonces, se realiza mediante una tabla de correspondencia entre códigos de clasificación y categorías. Se detalla el proceso y se enumeran las ventajas e inconvenientes de ambos. Por último, se describe la elaboración automatizada de categorías y asignación automática de los ítems. Palabras-clave: categorías; categorización manual; categorización automatizada; Internet; *browsing*; recuperación de la información

Abstract: The characteristics of the category systems in a web environment are described, as well as its making process and the categorization techniques used. Category systems in an electronic environment are tools of viewing and information searching by means of browsing. In the process of categorization two methods are proposed. In the first one, the structure of the system to design will be the same that will be used to classify the items and the one the user will see. The second one requires the use of hierarchical structures previously created, and also requires to have classified the items with the already mentioned hierarchical structures; the categorization then is made using an equivalence table between classification codes and categories. The process is detailed and advantages and disadvantages of both methods are explained. Finally, the automatic creation of categories and the automatic categorization of items are described.

Keywords: categories; manual categorization; automatic categorization; Internet; browsing; information retrieval

Características de los sistemas de categorías en el medio electrónico

Desde la aparición de Internet, uno de los mayores problemas de la navegación en documentos hipertextuales y en bases de datos de texto completo ha sido el fenómeno que se conoce por "perdido en el hiperespacio". La estructuración jerarquizada de la información en una secuencia de categorías de conceptos, progresivamente más concretas, donde cada categoría más amplia incluye todas las particulares, ha sido uno de los métodos más extendidos como soporte a la recuperación de la información a través del proceso de hojeo o *browsing*. Este tipo de estructura, que permite visualizar de forma global el conocimiento almacenado en el sistema y potencia la exploración de las relaciones entre los ítems de información, está siendo aplicado tanto en directorios temáticos generalistas como en webs de información especializada.

Entre las ventajas mencionadas a lo largo de la bibliografía existente sobre el uso de los sistemas de clasificación jerárquicos y facetados o de estructuras arbóreas *ad hoc* en Internet, resulta notable tanto la facilidad de utilización por parte de usuarios no expertos, como los beneficios del *feedback* durante el proceso de *browsing*. Otra de las ventajas es la posibilidad de ampliar o delimitar las búsquedas según se precise, sin necesidad de tener que conocer las complejas instrucciones de búsqueda ni las formalidades de la nomenclatura propia de un tesauro o de un sistema de clasificación tradicional. Resulta evidente que en todo sistema de categorías los usuarios no requieran ver la notación, puesto que el concepto se entenderá mejor si se le da nombre. Por otro lado, si bien en la ubicación física la notación debe ser única, en el entorno electrónico los ítems pueden aparecer en categorías distintas. Cabe destacar asimismo la posibilidad de realizar la búsqueda por término específico, una vez delimitado el campo semántico. Esto permite contextualizar dicha búsqueda y disminuir el ruido provocado en la recuperación.

Desde la perspectiva cognitiva clásica, el proceso humano de categorización se basa en principios que van más allá de los que propone un sistema jerárquico de clasificación tradicional. Una de las habilidades del razonamiento humano consiste en pensar sobre una misma entidad o acontecimiento de diferentes maneras, lo que nos brinda la posibilidad de acceder a diferentes tipos de conocimiento sobre ello. Esta flexibilidad presenta la característica de que cada objeto se almacenará en una categoría en nuestra memoria, de tal modo que podremos utilizar la que más nos convenga según el momento. La categorización y el análisis discriminatorio están íntimamente ligados en la representación de los conceptos, ya que los límites entre una y otra categoría deben ser claros con el propósito de que un objeto pueda ser descrito a partir de un número de parámetros.

La categorización en el entorno web consiste en una clasificación en clases preexistentes. Estas clases constituyen un entorno dinámico de aprendizaje, que permite inferir más información de la que nos proporciona una aproximación al objeto individual y el acceso a documentos u objetos que

comparten características esenciales similares o idénticas a un sujeto de búsqueda considerado pertinente, en tanto que previamente o en el momento de la representación, se ha hecho una división de un conjunto de objetos en un conjunto de clases lo más homogéneas posible.

En la visualización de la mayoría de los sistemas de categorías, uno de los factores a tener en cuenta es que, al no presentar divisiones mutuamente excluyentes, un objeto puede aparecer en más de una categoría al mismo tiempo. Cuando la asignación es manual, esto puede presentar serios problemas de objetividad. En el caso de que sea lógico mostrar un objeto en más de una categoría, teniendo en cuenta el modo natural de búsqueda de la información por parte del usuario, deben seguirse criterios coherentes que queden formalizados en el momento del diseño del sistema. En tanto que las interferencias entre categorías no son deseables, debemos tener en cuenta que una jerarquía bien hecha es la que tiene objetos similares dentro de cada división. Para identificar dichas interferencias, conviene evaluar cuántas categorías diferentes consulta el usuario para llegar a una misma información. Si los usuarios visitan distintas categorías para llegar a un mismo objeto, puede significar que la distinción entre ellas no es clara. No obstante, esto presenta la ventaja de encontrar una misma información bajo distintos puntos de acceso.

Por otro lado, la mayor parte de los directorios temáticos generalistas, suelen presentar estructuras arbóreas asimétricas, sin niveles consistentes de jerarquía formalmente lógica, lo que conduce a estructuras descompensadas, difíciles de entender y de retener en la memoria. A medida que vamos avanzando, se hace difícil prever cuáles serán las subcategorías y, a la vez, vamos dependiendo cada vez más de la terminología utilizada para representarlas. En sitios web con un número elevado de usuarios y un alto incremento de la colección, se echa de menos el uso de clasificaciones ya existentes, sean o no documentales, que garanticen una coherencia estructural mínima. En este sentido, deberemos tener en cuenta que uno de los condicionantes de la visualización vendrá dado por el colectivo de usuarios del web, que podrá considerarse genérico o bien con características claramente definidas que nos permitan establecer distintas tipologías.

Otro de los inconvenientes más destacables en cuanto al uso de este tipo de sistemas es la exploración de un parte limitada del mismo, inherente al proceso de *browsing*, en el que, basándonos en el reconocimiento de lo que ya sabemos, excluimos partes de la colección que no nos interesa recuperar. Como posible solución parcial a este problema, podemos complementar una jerarquía con una estructura facetada o una visualización dinámica en forma de árbol hiperbólico. La geometría hiperbólica permite visualizar una gran cantidad de información particular (foco) sin perder el contexto.

Metodología para la construcción de un sistema de categorías

Análisis del contexto

a) El primer paso consiste en definir el alcance y la organización del web para poder crear unas categorías adecuadas a sus características y objetivos.

Normalmente podremos diferenciar entre las páginas web de organizaciones con una finalidad lucrativa clara, es decir que se propongan comercializar un producto o servicio (librerías y tiendas virtuales, empresas constructoras, operadores de telecomunicaciones), y las que ofrecen una información a un colectivo de usuarios (enciclopedias, publicaciones oficiales, bibliotecas digitales, medios de comunicación). En cuanto al alcance, diferenciaremos los webs centrados en un ámbito temático de los que ofrecen información general.

Con el objetivo de evitar categorías descompensadas en el desarrollo de los niveles de especificidad, sería recomendable disponer de los datos sobre la cantidad total de ítems a categorizar, y a ser posible, distribuidos por ámbitos temáticos. Si la información no tiene ningún tipo de tratamiento documental anterior sería conveniente llevar a cabo un estudio previo con una muestra de la colección para obtener cantidades aproximadas.

b) Tipología de información que deben recoger las categorías.
Diferenciamos entre materiales individuales o productos concretos, y recursos electrónicos, publicaciones periódicas, compilaciones de objetos, etc., ya que será más complejo categorizar recursos electrónicos que contengan información de diverso tipo, que un ítem, como es un objeto, un informe o un artículo. En el caso de directorios de recursos web, la tendencia es describir únicamente la globalidad del recurso incluido en una categoría, es decir la página de inicio, sin tener en cuenta la información específica que cuelga del sitio web, que podría asignarse a una categoría distinta.

c) Tratamiento documental de la información antes de realizar las categorías.
Si la información ya ha recibido algún tipo de tratamiento documental, se debe valorar en qué medida es reaprovechable esta descripción.

d) Visualización y diseño de las categorías.
Diferenciamos entre estructuras jerárquicas, tipo *Yahoo,* y las tendencias de visualización que corresponden a metáforas visuales, sistemas que representan modelos complejos de información basándose en experiencias visuales con las que el usuario está familiarizado. Existen dos tipos de metáfora: las realistas (mapas, ciudades, cuerpo humano) y las abstractas (árbol hiperbólico). Las metáforas visuales resultan especialmente útiles cuando se desea ofrecer mucha información en una sola pantalla.

Algunas cuestiones que conciernen también a la visualización son:

- Determinación de las categorías principales

Es conveniente no superar la cifra de las diez o doce categorías principales para que el usuario pueda retener la información que ofrece la pantalla. No obstante, y especialmente en webs con contenido generalista, si limitamos todo el conocimiento humano a un número bajo de categorías principales, esto puede provocar que los puntos de entrada resulten confusos para el usuario. En un directorio de recursos web, cabe la posibilidad de crear categorías especiales para colectivos determinados de usuarios (por ejemplo: "Niños", "Jóvenes") o con el objetivo de priorizar cualquier tema de carácter más específico (por ejemplo: crear tres categorías distintas, una para "Arte y literatura", otra para "Medios de comunicación" y una tercera para "Cine y televisión", cuyo contenido podría perfectamente incluirse en las dos anteriores). Al añadir estas categorías especiales a las diez o doce categorías principales que hemos utilizado para representar el conocimiento humano de forma sistemática, superamos la cifra inicial recomendada, pero ganamos en transparencia y en efectividad en la navegación.

- Niveles de especificidad

Es recomendable definir los niveles de especificidad sobre la base de la cantidad de información que la web pueda ofrecer y a la previsión de incremento que tenga.

- Presentación de categorías principales y niveles de especificidad

Una presentación sistemática ayudará al usuario a situar mejor los ámbitos temáticos; por el contrario, una presentación alfabética impedirá una fácil comprensión de las categorías y la distinción entre ellas.

- Combinación de categorías y referencias

Según el sistema de categorías creado, puede ser interesante complementarlo con una estructura facetada que posibilite la combinación de las jerarquías con divisiones de tipo geográfico o cronológico, o bien con facetas propias de una especialidad. Las referencias entre subcategorías pueden plantearse como posible solución a las dificultades que conlleva la indefinición de las necesidades reales de información del usuario en el proceso de *browsing.* Las referencias asimismo permiten restringir el criterio de asignación múltiple, ya que podemos guiar mejor al usuario por la estructura y recomendarle visitar otras categorías relacionadas.

- Terminología

La denominación de las categorías debe ir a la par con las características del diseño y los objetivos generales del web. Según su alcance, y las estrategias de difusión y relación con el usuario, el lenguaje deberá ser especializado o más bien coloquial, aunque no deberíamos perder la perspectiva de un cierto control del vocabulario, utilizando siempre, a lo largo de la estructura, un mismo término para denominar conceptos similares.

Dos posibles métodos para la realización de un sistema de categorías

- Primer método:
Utilizamos para clasificar una estructura única creada a medida, que a su vez, será la estructura que visualice el usuario en el web. De este modo, la herramienta que usa el clasificador es la misma que usa el usuario para recuperar la información.

- Segundo método:
Clasificamos los ítems con un sistema de clasificación tradicional, ya sea generalista, como la Clasificación de Dewey, la CDU o la Clasificación de la Library of Congress, ya sea específico de un ámbito concreto, en función del alcance temático del web. Luego, y a partir de este tratamiento documental previo, realizamos una estructura a medida para visualizar los ítems ya clasificados. Por último, creamos las correspondencias entre códigos de clasificación y categorías mediante una tabla.

En el segundo método, el sistema de categorías sólo se utiliza para visualizar la información, ya que previamente habremos clasificado los ítems con otra herramienta: un sistema de clasificación formalmente correcto. El clasificador, por lo tanto, dispone de una herramienta distinta a la del usuario.

En ambos casos, el usuario realiza la búsqueda en un sistema de categorías, elaborado sobre la base de las conclusiones del análisis del contexto.

Ejemplo de aplicación de los métodos

- Primer método
Disponemos de las siguientes subcategorías de deporte: "Baloncesto", "Balonmano", "Fútbol", "Otros deportes". En la última, se incluyen el bádminton, el críquet, y otros deportes minoritarios. El clasificador, asignará los ítems sobre bádminton a la subcategoría “Otros deportes”.

- Segundo método
Se asigna a cada ítem el código propio del sistema de clasificación. Por ejemplo: AA (Baloncesto), AB (Balonmano), AC (Bádminton), AD (Críquet), AF (Fútbol), etc.

Para establecer la correspondencia entre estos códigos y las categorías realizamos una tabla de la siguiente manera:

Baloncesto	incluye	AA
Balonmano	incluye	AB
Fútbol	incluye	AF
Otros deportes	incluye	AD, AC, etc.

El usuario ve el mismo sistema de categorías en ambos casos pero el funcionamiento es diferente. En el segundo caso, el ítem es presentado bajo una categoría pero no pierde su ubicación original en el sistema de clasificación. Con ello, en cualquier momento, podremos desarrollar nuevos

niveles de especificidad bajo la categoría Deportes, sin necesidad de volver a categorizar el ítem original.

Características propias de los métodos

a) Hospitalidad del sistema
En el primer método, deben crearse categorías lo suficientemente hospitalarias para poder añadir nuevos temas, intentando siempre evitar la división de una categoría en dos o más categorías diferentes. En el ejemplo anterior, en el momento en que el bádminton necesite tener una subcategoría específica, se deberán revisar todos los ítems que están bajo “Otros deportes” y reclasificar los de bádminton.

En el segundo método, mucho más flexible, el cambio será rápido: el código de bádminton desaparecerá de “Otros deportes” y formará él sólo una categoría:

Baloncesto	incluye	AA
Balonmano	incluye	AB
Fútbol	incluye	AF
Bádminton	incluye	AD
Otros deportes	incluye	AC, etc.

b) Creación de nuevos niveles de especificidad
En el primer método, debemos prever el incremento de información para así crear suficientes niveles de especificidad. Aún así, será necesario modificar la estructura a medida que se genera más información. Es recomendable que antes de crear nuevos niveles de especificidad se estudie hasta qué punto son necesarios. Recordemos que en este método la herramienta de clasificación es la misma que la de visualización y puede generar fácilmente jerarquías descompensadas que confundan al usuario.
En el segundo método, la creación de nuevos niveles de especificidad no conlleva más problemas que el cambio en la tabla de correspondencia.

c) Integración con otras aplicaciones
Al contrario que en el primer método, si utilizamos clasificaciones tradicionales adaptadas, es posible crear sistemas abiertos, que se puedan integrar con otras aplicaciones.

d) Costes
El inconveniente más importante del segundo método es su coste, tanto económico como temporal, ya que la organización deberá disponer de un equipo de clasificadores que apliquen un nivel más exhaustivo de representación del contenido que en el proceso descrito en el primer método. Los costes pueden disminuir si se utilizan operaciones automatizadas, en lugar de procesos de categorización manual.

Elaboración de categorías y categorización automatizadas

Podemos diferenciar entre aquellos procesos que permiten generar de forma automática el sistema de categorías y aquellos que, partiendo de una estructura predefinida, asignan de manera automática (o semiautomática) los ítems a su ubicación correspondiente dentro del árbol jerárquico.

Observamos diferentes métodos para la generación automatizada de categorías, basados en el contenido o la estructura de la información, según el grado de tratamiento previo necesario:

a) Sistemas generados sobre la base de documentos previamente estructurados en directorios, ya sea importando y analizando la información siguiendo la estructura de un sitio web (directorios virtuales), o examinando la estructura del sistema de ficheros de un servidor (directorios físicos). Este proceso crearía una jerarquía que no sería más que un reflejo de la diseñada para el almacenamiento de la información.
b) Categorías generadas a partir de metainformación, siempre y cuando se haya seguido una política estricta de etiquetado durante el proceso de producción documental o alimentación de la base de datos.
c) Categorías generadas a partir de colecciones de documentos sin ningún tratamiento previo, aplicando técnicas avanzadas de agrupación de palabras-clave, detección lingüística, derivación, normalización, etc.

Por lo general, y con el objetivo de responder a un proceso abierto y sometido a continua revisión, la mayoría de sistemas permiten, una vez finalizado el proceso automático, añadir, suprimir o manipular categorías.

En cuanto a la asignación automática de ítems, ésta puede darse a partir de reglas de categorización basadas en el contenido de los documentos. Estas reglas pueden asignarse manualmente o a través de métodos inductivos. Estos últimos, construyen *clasificadores* a partir de conjuntos iniciales de documentos precategorizados.

En el momento de la asignación, los métodos automáticos compararán las representaciones de documentos y las representaciones de categorías, calculando el grado de semejanza entre ambas. Generalmente, las representaciones de documentos se obtienen durante los procesos de indización, aunque también se han aplicado técnicas de explotación basadas en redes neuronales o en análisis semántico.

Los procesos altamente automatizados dependen de tareas previas de organización o etiquetado de documentos, que requieren una planificación que garantice una correcta clasificación. Por ello, resulta aconsejable, en los casos en que nuestra colección, por sus características, precise de cierta automatización, la combinación entre métodos automáticos y manuales.

Conclusiones

Los dos métodos para la elaboración del sistema de categorías han sido presentados teniendo en cuenta el estudio previo de las necesidades de la organización productora y del colectivo de usuarios, así como la visualización de la estructura. Estos métodos pueden aplicarse tanto en el contexto de la clasificación manual como en el de la automatizada aunque ésta última conlleve una revisión por parte del gestor de la información.

El primer método puede resultar indicado para:
- organizaciones que no planteen la necesidad de añadir cambios frecuentes en el sistema de categorías,
- webs que no tengan un elevado incremento de la colección, siempre y cuando no necesiten potenciar la flexibilidad del sistema. Disponer de las cantidades aproximadas en el momento de realizar las categorías, nos ayudará a realizar un sistema lo suficientemente adecuado para la colección, en cuanto a niveles de especificidad y categorías principales.

Se recomienda el segundo método para:
- webs que vean la necesidad de añadir cambios frecuentes en el sistema de categorías, tanto para adaptarse mejor a las necesidades de sus usuarios, como para poder potenciar temas específicos según el momento,
- webs con un incremento importante de ítems diarios, ya que así podremos ir desarrollando niveles de especificidad a medida que vamos disponiendo de la información. Si utilizamos el primer método debemos controlar que las categorías sean lo suficientemente hospitalarias y que el desarrollo de los niveles de especificidad se haga de forma coherente.

En cuanto a la asignación automática de los ítems, se derivan dos conclusiones principales de los estudios de evaluación realizados sobre técnicas inductivas:
- el promedio de documentos correctamente asignados en una categoría será más o menos alto en función del método en el que se basará nuestro sistema. Las técnicas empleadas hasta ahora ofrecen resultados muy diferentes: desde aproximadamente un 60% hasta un 80-90% de exhaustividad en la recuperación o *recall.*
- incluso en los casos de máxima precisión, el margen de error es todavía lo suficientemente importante como para provocar desajustes notables.

De este modo, resulta aconsejable la implicación de personal especializado, tanto en la construcción y mantenimiento de la estructura (diseño de la jerarquía, asignación de reglas manuales de categorización, selección de conjuntos iniciales de documentos para métodos inductivos, etc.) como durante el proceso de categorización.

Bibliografía

ALBRECHTSEN, H.; JACOB, E.K. The dynamics of classification systems as boundary objects for cooperation in the electronic library. *Library Trends*, 1998, vol. 47, nº 2, p. 293-312.

CARO CASTRO, C. Sistemas de clasificación y organización de la información en Internet [en línea]. En *Fesabid 98, VI Jornadas Españolas de Documentación. Valencia, 29, 30 y 31 octubre 1998*. <http://www.florida-uni.es/~fesabid98/Comunicaciones/c_caro.htm> [Consulta: 10 enero 2001].

CHEN, H., et al. Internet browsing and searching: user evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 1998, vol. 49, nº 7, p.582-603.

CHOO, C.W.; DETLOR, B.; TURNBULL, D. Information seeking on the web [en línea]: an integrated model of browsing and searching. *First monday*, 2000, vol. 5, nº 2. <http://firstmonday.org/issues/issue5_2/choo/index.html> [Consulta: 12 enero 2001].

DESIRE. The role of classification schemes in Internet resource description and discovery [en línea]. Development of a European Service for Information on Research and Education. Feb. 1997. <http://www.ukoln.ac.uk/metadata/desire/classification/> [Consulta: 27 diciembre 2000].

DUMAIS, S.T. Using SVMs for text categorization [en línea]. *IEEE Intelligent Systems Magazine,* Trends and Controversies, 1998, vol. 13, nº4. <http://www.computer.org/intelligent/ex1998/pdf/x4018.pdf> [Consulta: 2 febrero 2001].

DUMAIS, S.T., et al. Inductive learning algorithms and representations for text categorization [en línea]. *Proceedings of ACM-CIKM98, nov. 1998*, p. 148-155. <http://research.microsoft.com/copyright/accept.asp?path=http://research.microsoft.com/~sdumais/cikm98.pdf&pub=ACM> [Consulta: 2 febrero 2001].

ESPELT, C. Improving subject retrieval [en línea]: user-friendly interfaces and effectiveness. *BiD: textos universitaris de biblioteconomia i documentació*, 1998, nº 1. <http://www.ub.es/bid/01espel1.htm> [Consulta: 10 enero 2001].

GARCÍA MARCO, F.J. Interfaces amigables para la representación de la información bibliográfica. *Scire: representación y organización del conocimiento*, 1995, vol. 1, nº 1, p. 127-148.

IYER, H. Classificatory structures: concepts, relations and representation. Frankfurt: Indeks Verlag, 1995. 229 p. (Textbooks for knowledge organization, 2). ISBN 3-88672-501-4.

KWASNIK, B.H. The role of classification in knowledge representation and discovery. *Library Trends*, 1999, vol. 48, nº 1, p.22-47

LAMPING, J.; RAO, R.; PIROLLI, P. A focus-context technique based on hyperbolic geometry for visualizing large hierarchies. *CHI '95, Conference on Human Factors in Computing Systems, Denver, Colorado, May 7 - 11, 1995*. <http://www.acm.org/sigchi/chi95/proceedings/papers/jl_bdy.htm> [Consulta: 23 enero 2001].

MARCHIONINI, G. Information seeking in electronic environments. Cambridge: Cambridge University Press, 1995. 224 p. ISBN 0-521-58674-7.

MURPHY, G.L..; LASSALINE, M.E. Hierarchical structure in concepts and the basic level of categorization. En *Knowledge, concepts and categories*. Edited. by K. Lamberts and D. Shanks. Cambridge, Mass.: MIT Press, 1997, p. 93-131.

POULTER, A. Browsing the virtual library. En *Encyclopedia of Library & Information Science*. New York: Marcel Dekker, 1998, vol. 62 (sup. 25), p. 54-64.

RISDEN, K. (1999). Toward usable browse hierarchies for the web [en línea]. *Human computer interaction: ergonomics and user interfaces*, nº 1, p.1098-1102.
<http://www.microsoft.com/usability/UEPostings/HCI-kirstenrisden.doc> [Consulta: 19 enero 2001].

WHEATLEY, A. Subject trees on the Internet: a new role for bibliographic classification? *Journal of Internet Cataloging*, 2000, vol. 2, nº 3/4, p. 115-141.