

Mary Micco
**Control de Autoridad de Materia
en el Mundo de Internet**

Este artículo está dividido en dos partes. La primera se refirió al problema en general del control de autoridad en Internet y las dificultades involucradas. La segunda habla del uso de la clasificación en la autoridad de materia a través del mundo de Internet, para mejorar el filtrado y la precisión.

Resumen

Hoy en día, debemos analizar el control de autoridad de materia como un sistema que soportará búsquedas a través del vasto dominio de Internet. Si bien reconocemos que existen muchas dificultades, es hora de darle otra mirada y buscar nuevas soluciones. Como herramientas de navegación, se deben diseñar presentaciones gráficas generadas en forma automática, que identifiquen los objetos de la información en muchos niveles diferentes, pero que estén organizadas por un área temática amplia. Esto significa que necesitamos sacar ventaja de los nuevos métodos de producción de documentos, incentivando a los autores a agregar descriptores y, más importante aún, números de clasificación generales con la ayuda de software de sistemas expertos. Más adelante, esta catalogación y clasificación inicial puede ser perfeccionada por profesionales, pero al menos sentará la base para una manipulación e indexación más sofisticada en el punto de ingreso al sistema. En la actualidad, el mayor problema es cómo filtrar lo que no queremos y restringir la búsqueda a los documentos que tal vez nos sean de utilidad.

II Parte. El uso de la clasificación en la autoridad de materia a través del mundo de Internet para mejorar el filtrado y la precisión.

I. Crear filtros efectivos (autoridad de materia)

Debemos reconocer que uno de nuestros mayores problemas es desarrollar un sistema filtrante efectivo, que restrinja el dominio de la búsqueda a las áreas generales apropiadas y, a la vez, aumente la profundidad y variedad de los descriptores dentro de ese grupo. Para ello, se pueden seguir varios pasos que mejoren el proceso de búsqueda. El primero es asegurar la inclusión de un número de clasificación rudimentario que identifique la "naturaleza" del documento, el tipo de documento (paquete de información) y los conceptos claves involucrados, en la información de los ficheros para cada documento que se ingrese a Internet en el punto de entrada. Hay muchas clases de filtros que se pueden crear con una dificultad mínima.

A. Determinar la "naturaleza" o tema del documento.

La mayoría de los autores suelen crear cadenas de los autores suelen crear cadenas o títulos para describir el tema de su documento. Mejorar la calidad de esta indexación, fomentando el uso de palabras clave y subtítulos significativos, no tendría que ser una tarea difícil. Junto con

los principales descriptores extraídos del documento, un software de búsqueda debe ser capaz de comparar este documento con otros del sistema que tratan el mismo tema.

B. Identificar el paquete apropiado de información.

Con el tiempo, los conceptos y las ideas evolucionan y se convierten en conocimientos fijos, formando parte de la reserva de información o legado que se transmite a la próxima generación. En el camino, gran parte de esos datos se descartan, consolidan, modifican o se utilizan de otro modo. La información se empaqueta varias veces. Comienza como un ítem en un servidor de listas, después se convierte en un documento via ftp y, más adelante, en un artículo publicado en una revista. En algunos casos, se le asignan nuevos términos al concepto y se convierte en parte del lenguaje técnico de una disciplina. Pueden pasar muchos años antes de que finalmente aparezca como un capítulo en un libro, un artículo en una enciclopedia o una definición en un diccionario. Sólo si es muy significativo será digno de un libro completo. El paso final llega cuando se lo incorpora a un libro de texto, pero los conceptos que alcanzan este punto son relativamente pocos. La mayoría desaparece en el camino. Cuando se analiza el control de vocabulario, la pregunta es cuándo llega a estar lo suficientemente maduro un concepto como para asegurar el esfuerzo y el costo de procesamiento a cargo de un catalogador capacitado, y cómo podemos utilizar nuestra tecnología para proveer en forma automática algún nivel de control para materiales efímeros.

Como la información puede estar en muchos paquetes diferentes, según su etapa evolutiva, y como cada uno de ellos tiene importancia, filtrar la información es una vía esencial: limitar la búsqueda a revistas especializadas, buscar en artículos enciclopédicos, Usenet, actas de sesiones, memorias de empresas... Igualmente importante es comprender que tenemos la unidad básica -el documento en muchos paquetes-, pero también tenemos paquetes compuestos que contienen documentos múltiples, como actas de sesiones y bases de datos periódicas. Además, contamos con herramientas de búsqueda, como guías telefónicas, libros de referencia, índices, diccionarios y tesauros, paquetes de información que nos asisten en el proceso de localización de la información requerida. Hay que codificar cada documento por tipo de paquete para mejorar el proceso de filtrado.

B. Nombrar a los futuros usuarios en el campo "A:"

Si bien desde hace años los bibliotecarios proveen una etiqueta para clasificar el nivel intelectual del documento, hay bastante resistencia al respecto. En su lugar, proponemos utilizar el campo "A:". La mayoría de la información está dirigida a un público específico. Puede ser un informe científico sobre una nueva terapia con fármacos para el infarto de miocardio, destinado a todos los profesionales de la medicina, o una guía para pacientes sobre cómo prevenir arterias obstruidas mediante ejercicios y dieta, que utiliza términos no especializados. Contar con sólo 5 o más categorías generales -como Juvenil, No profesional, Profesional, Académica, Técnica-, permitiría filtrar el material no deseado.

C. Extraer los conceptos principales: Qué y Por qué

La mayoría de los autores pueden hacer un resumen explicando qué escribieron y por qué. Estos descriptores deben estar colocados en la información identificadora.

D. Utilizar facetas como Cuándo, Dónde, Quién y Cómo

Las respuestas a las preguntas cuándo, dónde, quién y cómo también son útiles para filtrar. De establecerlas como filtros, se pueden utilizar junto con los descriptores de materia para ayudar a mejorar la precisión de la búsqueda. Si existiera un sistema de control de autoridad de nombres en Internet, los autores podrían verificar si utilizan la forma correcta del nombre con un sistema experto. Si no estuviera el nombre en la base de datos, se lo podría presentar para su verificación y posible inclusión.

E. Asignar el número de clasificación: qué lugar ocupa en el árbol de conocimientos existente.

Esta es, tal vez, la tarea intelectual más difícil para el autor en el momento de describir el documento. En muchos casos puede ser un foro de discusión y en otros casos puede ser un foro de discusión. Todas las publicaciones periódicas ya están clasificadas, así que los artículos presentados pueden adoptar automáticamente este número de clasificación más general. Los sitios FTP son bastante más difíciles porque son depósitos de basura. De todos modos, se los debe incentivar a que agrupen los materiales similares en directorios. Los sitios bien mantenidos podrían proveer números de clasificación para los directorios sin un esfuerzo sobrehumano. Otra posibilidad es que los autores utilicen el mismo software diseñado para ayudar a los usuarios a identificar registros que coincidan con sus pedidos, con el fin de localizar materiales similares y obtener los números de clasificación de los registros más aproximados.

F. Ponderación automática de términos en base a etiquetas SGML.

Resta solucionar el problema de indexar el texto completo del documento. Hay un gran interés por confeccionar bases de datos en texto completo, y muchas instituciones ya están escaneando grandes colecciones de informes técnicos y expedientes judiciales con el propósito de mejorar sus capacidades de búsqueda. Todavía se necesita como guía la descripción básica del documento, que permitirá la indexación de grandes grupos de documentos en texto completo para que los usuarios puedan ubicar rápidamente los textos de interés. En los futuros sistemas, el usuario quizá pueda comenzar por buscar sólo las descripciones de los ficheros, y luego pasar a una búsqueda de texto completo una vez seleccionados los grupos adecuados de registros. Aún así, es probable que recuperen más información de la que puedan manejar, a menos que se aplique alguna forma de ponderación. Por ejemplo, si una búsqueda en texto completo de “infarto de miocardio” da como resultado 3.000 registros en una base de datos de medicina, el usuario necesitará utilizar la ponderación para obtener lo mejor de esos resultados. Si el término aparece en la descripción de los ficheros (ponderación 1) o en el título o resumen (ponderación 2), es evidente que será más apropiado que en un encabezamiento de párrafo (ponderación 3) o una referencia de paso en el texto de un párrafo (ponderación 4). Esta ponderación se puede automatizar por completo con la etiqueta SGML asociada a cada término. En nuestro sistema prototipo, clasificamos los términos ponderados y seleccionamos sólo el de valor más alto si había duplicados. Una vez que esté todo ordenado y que los usuarios se acostumbren a incorporar esta clase de información a los encabezamientos de sus documentos, se podrían indexar en forma automática los nuevos ítems ingresados al sistema.

II. Desarrollar herramientas de gestión de la información y sistemas expertos

El problema del control de vocabulario es muy complejo. va que hablamos de un sistema que funcionará independientemente de los

objetos de la información del sistema y que ayudará a los usuarios a acceder al tipo de material que buscan, en el área temática de su elección, en el nivel correcto de especificidad y en las bases de datos pertinentes. Será necesario crear mapas de áreas temáticas, vinculados entre sí en una estructura de árbol muy parecida a nuestros mapas de rutas. Si queremos ir de California a la ciudad de Nueva York, usamos mapas que muestran esencialmente las principales autopistas interestatales. Cuando llegamos al estado de Nueva York, necesitamos conocer las autopistas locales y estatales. Luego, en la ciudad de Nueva York, vamos a necesitar un mapa detallado de calles que nos ayude a llegar a nuestro destino. Lo mismo ocurre con las búsquedas. Los usuarios deben ser capaces de aproximarse o alejarse a voluntad, utilizando vínculos de hipertexto. Necesitaremos contar con sistemas expertos que construyan y administren estos mapas como pantallas dinámicas, que reflejen nuestras bases evolutivas de conocimientos en forma precisa y asistan a los usuarios a localizar las áreas de interés.

A. Usar un esquema de clasificación para controlar el vocabulario.

La clasificación es una herramienta importante en el control de vocabulario, porque representa una organización del conocimiento y se puede utilizar para determinar el lugar que le corresponde a una materia dentro del contexto más general. Provee un marco para una infraestructura. Muestra las relaciones entre términos a diferencia de las listas alfabéticas de términos o la búsqueda de “fuerza bruta” a vuelo de pájaro. En las bibliotecas se utilizan varios sistemas de clasificación importantes para organizar el total del conocimiento humano. El esquema de clasificación de la Biblioteca del Congreso de los EE.UU. se usa ampliamente como un sistema de almacenamiento de libros en estanterías, pero no está bien diseñado para la manipulación mecánica y no es jerárquico, por lo tanto no soporta ampliar o refinar un tema, así como tampoco se pueden aprovechar los números de clasificación para obtener los conceptos más aproximados. Tanto el Esquema de Clasificación de Dewey (R) como la Clasificación Decimal Universal (UDC) son ideales para la manipulación mecánica. La UDC ya cuenta con una versión legible por máquina que los usuarios pueden ver y manejar. Utiliza la subdivisión decimal, la notación numérica y las subdivisiones estándar con características mnemotécnicas. Las cadenas de texto sirven como guía para los no profesionales y proveen más descriptores para las búsquedas. Se utiliza en Europa, pero no parece tener el reconocimiento y la difusión que tiene el esquema de Dewey. Existe una propuesta para seleccionar uno de estos dos esquemas como base para un sistema amplio de clasificación de red. Los experimentos realizados con el proyecto ILSA (Micco) revelaron que hasta un número de clasificación de 3 dígitos sirve para filtrar el material no deseado, con el beneficio adicional de que los usuarios pueden determinar qué aspecto del tema se está tratando, si bien se prefieren 5 o más dígitos.

B. Características esperadas de un esquema de clasificación

1. Esquema de numeración jerárquica. Si en un sistema decimal se reserva cada serie de 0 a 9 para un tema, después se puede subdividir cada número por 10, que a su vez también se puede subdividir. Esto concuerda con el hecho de que sólo podemos manejar alrededor de 7 ítems en nuestra memoria a corto plazo. Esto significa que cada tema tiene una escala de números conocida que se adapta a las subdivisiones estándar.

2. Característica esquemática. Si se quiere hacer una búsqueda sobre cáncer (melanomas), se puede localizar el número de clasificación

general y aprovecharlo para obtener todos los encabezamientos (representados por subdivisiones decimales), que quizá sean más específicos. Esto no es posible en una lista alfabética de materias.

3. Niveles de especificidad. Se puede seleccionar el nivel de especificidad que se ajuste a los contenidos del artículo. En Dewey, por ejemplo, una enciclopedia general estaría clasificada bajo el “000”, mientras que un trabajo sobre lugares de pesca en Newfoundland estaría bajo “Pesca 612.46”.

4. Filtrado efectivo. Si sólo se quieren ver artículos sobre Pesca, sencillamente se restringe el filtro a “612.46” y se eliminan todos los demás.

5. Control de sinónimos. Tal vez existan muchas otras palabras para describir estos lugares, pero el número de clasificación trabajará con todas las posibles variaciones, incluidas las traducciones.

6. Se pueden incorporar características mnemotécnicas. Por ejemplo, se puede acordar que los dígitos “0” representen los “tratamientos generales” y los “9” los “tratamientos históricos”.

7. Subdivisiones estándar. Las subdivisiones estándar se pueden generar y utilizar en todas partes; por ejemplo, las subdivisiones geográficas para lugar y período.

8. Manipulación eficiente de computadoras. Los números, aún los que son muy largos, se pueden manipular por computadora de una manera muy eficiente. Sin embargo, a los seres humanos les resulta difícil comprenderlos. La solución adoptada en Medlars es representar los números con frases descriptivas, diseñadas para el hombre, que contienen palabras clave. Otra ventaja es que estas frases proveen descriptores adicionales para indexar.

9. Indexación semiautomática. Utilizando un índice Wais, se pueden obtener documentos similares y luego extraer la información a indexar de los mejores registros coincidentes.

C. Desarrollar un sistema experto que organice la infraestructura de la información.

Más de 22 millones de libros y publicaciones periódicas ya tienen asignados números de clasificación, de manera que tenemos una base significativa de conocimientos en forma legible por máquina desde la cual comenzar. Cada registro también contiene materias que representan los términos controlados recopilados y mantenidos por la Biblioteca del Congreso durante más de 100 años. Utilizando la información ya incluida en los registros Marc podríamos generar tres índices útiles.

1. Número de clasificación y encabezamientos de materia y descriptores asociados. Si los términos controlados o cualquier otro término del lenguaje natural que está en los registros (que proviene de títulos, notas o resúmenes) estuvieran conectados por máquina con los números de clasificación, podríamos clasificar automáticamente las listas de términos por frecuencia y asociarlas con cada número de clasificación. Esta distribución se podría representar en mapas de áreas temáticas.

2. Término índice y números de clasificación asociados. También se podría crear un índice invertido con todos los términos controlados y no controlados, que muestre con qué números de clasificación están asociados. El usuario ingresa un término y se fija qué números de

clasificación contienen enlaces a ese término. El resultado de una experiencia con el proyecto experimental ILSA (Micco), conyector experimental ILSA (Micco), con 100.000 registros Marc, fue un promedio de 3,7 números clasificadores por término. Si a estos números se les asignan frases descriptivas como Historia, Europa Oriental, Período de Guerra Fría o Muro de Berlín, el usuario podrá observar qué aspecto del tema se está tratando y seleccionar el área temática que le resulte más útil.

3. En un tercer índice, los términos del lenguaje natural deben estar conectados con los encabezamientos temáticos controlados cada vez que aparecen en el mismo registro, ofreciendo de esta manera una forma de control automatizado de sinónimos. Cuando el usuario busca un término, aparecen enlaces a otros términos. Con una base de datos así, desde la cual trabajar, hay una buena probabilidad de que un usuario obtenga un registro que coincida con su término y que sea capaz de determinar, desde este punto de entrada, qué sinónimos se utilizaron y localizar términos relacionados, más aproximados y generales. Si bibliotecarios capacitados controlaran esta base de datos evolutiva con la habilidad de agregar, borrar y hacer cambios globales rápidamente, el control de autoridad estaría íntimamente vinculado con la recuperación; mamente vinculado con la recuperación y representación de información, y obtendríamos un beneficio enorme con una mínima intervención del hombre. Si identificáramos a todos los objetos reconocidos como herramientas de búsqueda (tesauros, diccionarios, enciclopedias) a través de códigos en esta base de datos primaria, podríamos incluir mapas temáticos más detallados y herramientas de búsqueda especializadas, como los tesauros, para realizar una búsqueda más a fondo.

D. Bases de datos periódicas: Se pueden manipular de una manera bastante parecida. Aquí las publicaciones periódicas ya fueron clasificadas. Podríamos asignar el número de clasificación de una publicación a cada artículo que proviene de ella. Una vez más, sugerimos vincular los términos del título y del resumen con el número de clasificación y viceversa, pero aquí la ponderación va a ser más significativa como herramienta para filtrar. Para cada base de datos, el tesoro con sus términos controlados y red de referencias tendría que pasar a formar una parte integral del software de búsqueda.

<

E. Las bases de datos en texto completo son bastante más difíciles, ya que varían considerablemente en granularidad. Por un lado, tenemos un CD-ROM con el texto completo de la Enciclopedia Grolier; por el otro, podríamos tener una base de datos con 100.000 informes técnicos del Ministerio de Defensa, con términos controlados asignados a cada documento tomados de un tesoro del Ministerio. Con cada una de estas bases de datos, se deberá evaluar en forma profesional el grado y la profundidad de la indexación y la clasificación a utilizar. Las herramientas de indexación y control de autoridad, utilizadas por los catalogadores, se deben examinar y organizar minuciosamente para integrar un conjunto completo de herramientas, que después estarán en línea y a disposición de los usuarios y catalogadores, con la guía de normas de sistemas expertos, para ayudar a los usuarios a seleccionar el nivel apropiado.

F. Integrar todo: Para facilitar el problema de la navegación y asistir a los usuarios, necesitaremos otro nivel de organización basado en la clasificación. Si a cada objeto de la información en Internet se le asignara un número de clasificación, se podrían generar automáticamente, se podrían generar mapas que indiquen qué recursos están disponibles en un área determinada, con datos adicionales que muestren cada objeto y señalen si se lo puede oír o consultar. Esto, por supuesto, se puede generar por computadora con los datos ya reunidos en la

información de los ficheros.

G. Sistemas expertos. Para administrar la complejidad de una red de fuentes de este tipo y ayudar a los usuarios, se deben diseñar sistemas expertos que primero recaben el perfil del usuario, luego verifiquen la necesidad de información y, por último, lo ayuden a navegar a través de Internet, ubicando los paquetes de información más adecuados y, lo más importante, filtrando el material considerado irrelevante. Una de las tareas más difíciles será exponer los datos recuperados de maneras que eduquen, y no confundan, al que realmente busca información. El sistema siempre debe dejar que el usuario tenga el control y poder interpretar sus decisiones. Los usuarios deben poder sacar ventaja de un sistema de vínculos bien diseñados, entre las muchas herramientas de búsqueda diferentes, para conocer la variedad de recursos que tienen a su disposición y refinar sus selecciones sin esfuerzo. El objetivo final del sistema debe ser realizar la mejor comparación y selección posible entre la necesidad de información del usuario y los recursos del sistema.

III. Conclusión

Cada artículo debe tener un rótulo estándar para mejorar el manejo de la información y, en particular, el filtrado. Si vamos a construir sistemas expertos efectivos que nos ayuden a recuperar información, éstos deben contar con una ubicación uniforme de la información crítica y un uso uniforme del lenguaje. El autor del documento es el que puede realizar mejor el trabajo de crear el rótulo estándar, siempre que tenga a su disposición las herramientas necesarias para determinar el número de clasificación y los encabezamientos de materia.

Damos por sentado el uso de etiquetas SGML para ponderar los términos del sistema.

Se debe diseñar una infraestructura de la información fundada en una estructura de clasificación de base, compuesta por mapas con punteros a herramientas de búsqueda, guías, libros de referencia, bases de datos, herramientas de control de vocabulario, bases de datos peri de vocabulario, bases de datos periódicas, sitios FTP, servidores de listas, etc. En vez de un fichero horizontal gigante de descriptores, necesitamos ofrecer un conjunto de herramientas de acceso, incluyendo índices y tesauros, tan completo como sea posible, para satisfacer las diferentes necesidades y expectativas, pero éstas deben estar identificadas y señaladas en los mapas de áreas temáticas, para que los usuarios pueden pasar de una a otra fácilmente.

Existen divergencias, aún dentro del mundo bibliotecario, con respecto al uso de la clasificación en los sistemas en línea. Sin embargo, el sistema Medlars, de gran éxito, demostró sin lugar a dudas el valor de tener un sistema numérico de base estructurado en forma jerárquica. Nuestra propuesta es perfeccionar la clasificación y utilizarla para poner un poco de orden en el caos de Internet.

Bibliografía

1. Bowman, C. Mic y otros. "Scalable Internet Resource Discovery: Research Problems and Approaches". Communications of the ACM. Agosto 1994. Vol. 37. No. 8. pp98-114. "Taxonomies allow a more uniform search space than is possible solely by content-indexing of documents". Eby content-indexing of documents". Expert system technology could be developed to match the key terms in the document against similar documents. The author could then select the class number and index terms that seemed most relevant adding more as needed. "We believe tools should be developed that allow authors to mark up their documents with classification terms from some selected set of taxonomies."

2. Boynton, G.R. y Sheila D. Creth, eds. *New Technologies and New Directions*. Westport, CT: Meckler Publishing, 1993. 118p. "Nine articles from a symposium of University "scholarly" publishing, learning, creating and management of information using computers."
3. Broad, William J. "Doing Science on the Network: A Long Way from Gutenberg." *New York Times*, martes 18 de mayo de 1993, p.B10, col 1. "Much of the beauty and wonder of Internet and its resources could become a horrific problem. Systems and people will shut down. I know people who have stopped using Internet because they get 500 messages a day." Susan K. Kubany, Presidente de Omnet, Inc.
4. Chan, Lois Mai. "Part II: Library of Congress Classification System. The Library of Congress in an Online Environment." *Cataloging and Classification Quarterly* 11 (1):7-25. "One of the great adv11 (1):7-25. "One of the great advantages of online retrieval systems is that access provisions need no longer be either/or propositions. The question now is, how can we make the best of our battery of bibliographical access tools, with classification and the alphabetical approach used together to complement each other?" p 25.
5. Cochrane, Pauline A. y Karen Markey. "Preparing for the Use of Classification in Online Cataloging Systems and in Online Catalogs." *Information Technology and Libraries* 4 (Junio 1985) (91-111). For the online catalog user, library classification becomes a tool for augmenting subject access, providing browsing capabilities through the classed approach to subject searching in the schedules and the alphabetical approach in the index and enhancing the display of library materials' subject matter. 109
6. McMillan, Marilyn and Gregory Anderson. "The Prototyping Tank at MIT: "Come On In, the Water's Fine". *Cause/Effect*. Vol.17. No.3. Otoño 1994. pp51-54.
7. Micco, Mary y Rich Popp. "The ILSA Project: an Investigation into Techniques for Improving Subject Access: A Theory of Clustering based in Classification". *Library High Technology*. Junio, 1994.

8. Reinhart, Andy. "Managing the New Document" *Byte*. Vol. 19. No. 8. Agosto, 1994. "A document will no longer be a single file but rather a book of pointers to text objects, data objects, images, fonts, and so on".... p.93

=====
Copyright Mary Micco 1996.

Este artículo apareció originalmente en *LIBRES: Library and Information Science Electronic Journal* (ISSN 1058-6768). Septiembre 1996. Vol. 6, Núm. 3.

Traducido con la correspondiente autorización de los autores.
Departamento de Informática y Sistemas.

Mary Micco

Control de Autoridad de Materia en el Mundo de Internet

Este artículo está dividido en dos partes. La primera se refiere al problema en general del control de autoridad en Internet y las dificultades involucradas. La segunda habla del uso de la clasificación en la autoridad de materia a través del mundo de Internet, para mejorar el filtrado y la precisión.

Resumen

Hoy en día, debemos analizar el control de autoridad de materia como un sistema que strol de autoridad de materia como un sistema que soportará búsquedas a través del vasto dominio de Internet. Si bien reconocemos que existen muchas dificultades, es hora de darle otra mirada y buscar nuevas soluciones. Como herramientas de navegación, se deben diseñar presentaciones gráficas generadas en forma automática, que identifiquen los objetos de la información en muchos niveles diferentes, pero que estén organizadas por un área temática amplia. Esto significa que necesitamos sacar ventaja de los nuevos métodos de producción de documentos, incentivando a los autores a agregar descriptores y, más importante aún, números de clasificación generales con la ayuda de software de sistemas expertos. Más adelante, esta catalogación y clasificación inicial puede ser perfeccionada por profesionales, pero al menos sentará la base para una manipulación e indexación más sofisticada en el punto de

ingreso al sistema. En la actualidad, el mayor problema es cómo filtrar lo que no queremos y restringir la búsqueda a los documentos que tal vez nos sean de utilidad.

=====

I Parte. El control de autoridad de materia se debe analizar en un contexto más amplio

A. Objetivo: Construir sistemas de computación para soportar una recuperación efectiva de la información. Hacerlo con una mínima intervención del hombre. Para los fines de este análisis, preferí limitar mis comentarios al control de autoridad de materia, que analizaré no en función del catálogo bibliotecario para libros y materiales no impresos, sino desde el punto de vista del usuario que tiene una determinada necesidad de información. El usuario que busca información vía Internet, un medio que provee acceso no sólo a los recursos bibliotecarios sino también a todo un conjunto de recursos menos tradicionales. El verdadero fin del control de autoridad debe ser ayudar al usuario a pasar, sin esfuerzos, de su terminología (idioma natural) a los términos en uso del sistema (vocabulario controlado), y ubicar todos los materiales (objetos) pertinentes sin considerar en qué base de datos están almacenados o la forma en la que están presentados. Si busca “ataque cardíaco”, el sistema de control de autoridad debe vincularlo con “infarto de miocardio” u otros sinónimos, así como también con todas las fuentes de información apropiadas. Un requisito igualmente importante para un sistema de control de autoridad de materia es que debe estar automatizado y mantenerse solo, con una mínima intervención del hombre.

Si todas las bases de datos periódicas, listas y revistas electrónicas, CD-ROM, directorios ftp, menús gopher y páginas iniciales (“home pages”) de la WWW se clasificaran como objetos, se les asignara un número de clasificación (por materia) e identificara por tipo, sería relativamente simple concentrar la búsqueda de un usuario en el área temática de su interés, en vez de hacer una búsqueda de “fuerza bruta” de los recursos del mundo.

B. Dominio: Los conocimientos registrados de la humanidad. Por lo general, el usuario tiene interés por todo lo que está disponible y desea comenzar su búsqueda desde una computadora de escritorio. Le gustaría tener acceso a todos los conocimientos registrados de la humanidad, pero sólo en lo que respecta a su interés. En otras palabras, quiere hacer una búsqueda generalizada, pero filtrar lo que no se relaciona con su necesidad particular. Quiere lo mejor de lo que está disponible, en vez de los primeros 200 resultados. Necesitamos herramientas que ordenen el perfil de un usuario que ordenen el perfil de un usuario, para ayudarlo a filtrar lo que no es apropiado. Esto indicaría que también necesitamos implementar una asignación automatizada de valores para ayudar a clasificar los resultados. Es evidente que si el resultado aparece en el título o en las palabras clave controladas, se le debe asignar un valor más alto que si aparece simplemente en el texto de un párrafo. El sentido es que el usuario sea

capaz de utilizar una clasificación jerárquica para ingresar al sistema, en el nivel deseado de especificidad en el tema de su elección, con la opción de ampliar o acotar una búsqueda que no es fructífera. Las herramientas actuales no lo permiten, y no lo permitirán a menos que se implemente alguna forma de clasificación con organización jerárquica. Es interesante destacar cuántos catálogos clasificados están apareciendo en Internet para guiar al usuario a través del laberinto. El éxito del servidor Yahoo (www.yahoo.com) muestra claramente que la gente aprecia una organización jerárquica por materia.

C. Materiales: Todos los disponibles. Internet expandió en gran medida las formas de transmisión de información. El correo electrónico, los servidores de listas, tableros de los servidores de listas, tableros de anuncios y servicios usenet proveen fácil acceso a un caudal de datos actuales e informales sobre cualquier tema que se pueda imaginar. El FTP hizo posible la publicación y distribución informal de materiales, en papel y multimedia, muy rápidamente y a bajo costo. Los servidores Gopher nos posibilitaron el acceso a toda la información en línea de las ciudades universitarias, organizaciones gubernamentales y empresas en una red vinculada de sitios. La World Wide Web expandió aún más esta capacidad con los enlaces de hipertexto para multimedia, para que podamos navegar por imágenes gráficas, videoclips, animaciones y una infinidad de recursos en forma rápida y fácil. Todavía más poderosa es nuestra habilidad de reunir, en una página inicial, referencias a recursos de cualquier parte del mundo respecto de un tema o interés particular. Incrementamos considerablemente el número de lugares en los que podemos buscar información, sin un aumento proporcionado en la sofisticación de nuestras herramientas de búsqueda. Los usuarios deben ser capaces de especificar un paquete de información particular, y nosotros debemos empezar a considerar el hecho de que ciertos paquetes representan estructuras compuestas coepresentan estructuras compuestas con documentos múltiples; por ejemplo, las páginas iniciales de la Web.

II. Problemas: En macroescala

A. Explosión de información en Internet. Con más de 3 millones de computadoras conectadas con un número desconocido de usuarios y archivos en estos últimos tres años, no es extraño que nuestra capacidad para manejar y asimilar esta explosión de información se haya quedado atrás.

B. Falta de planeamiento/gobierno cooperativo. La estructura administrativa de Internet se concentra, naturalmente, en la conectividad y los estándares para lograr seguridad e interoperabilidad, más que en el manejo de los problemas de recuperación de información. Hasta este punto, los bibliotecarios y profesionales de tecnología de la información no cooperaron muy activamente en el diseño o administración de los servicios de recuperación de información de Internet, si bien los primeros han hecho un uso activo del “backbone” para distribuir sus propios servicios de bases de datos.

C. Falta de estándares para describir el contenido de información. Serios proyectos esormación. Serios proyectos

están en proceso de ejecución actualmente para redefinir las arquitecturas de manejo de documentos y proveer más información significativa sobre los encabezamientos para archivos. Un documento ya no será más un solo archivo, sino un libro de punteros a objetos de texto, imágenes, fuentes y sonidos, con información concisa que incluya autor, título, palabras clave, número de versión, descripción y estadísticas de archivos, además de los descriptores tradicionales de archivos (Reinhart). Es de vital importancia, para el éxito de un proyecto sobre control de autoridad, integrar a esta arquitectura de documentos la información requerida de una manera estandarizada. La selección de un número de clasificación y de palabras clave temáticas debe hacerla, en primer lugar, el autor del documento cuando lo está creando, y se debe mantener como parte de la información del archivo de una manera bastante parecida a la que se crean los resúmenes para artículos periódicos. Sería excesivamente costoso pagar a intermediarios para hacerlo. Para este proceso, se necesitarán herramientas tales como componentes de sistemas expertos. Ya existe en el mercado un software que extrae registros coincidentes para cadenas de 50 o más palabras clave, y después los clasifica. El usuario sencillamente solicita los registros que coincidan con su documento, y después selecciona los más aproximados aplicando los mismos términos controlados y números de clasificación a su propio documento. Se presume que esta técnica funcionará muy bien para el material efímero, mientras que el material que es más permanente será procesado otra vez en la cadena de la información por profesionales capacitados, que pueden verificar la elección de un número de clasificación y de términos controlados.

D. Falta de una estructura de información que unifique. Dado que ya existen billones de documentos y que se generan más con creciente velocidad, necesitamos una infraestructura sofisticada que ayude a localizar y manejar nuestros recursos de información. En vez de tratar al universo como un servicio monolítico de información, tendría mucho más sentido dividirlo en una serie de mapas temáticos, estructurados de manera jerárquica, mostrando ítems agrupados por tipo de material.

Dentro de cada mapa temático, debemos ser capaces de identificar las colecciones bibliotecarias especiales colecciones bibliotecarias especiales críticas, tesauros, enciclopedias, obras de referencia, publicaciones periódicas, bases de datos de artículos periódicos, servidores de listas, grupos usenet, sitios ftp y páginas iniciales y gophers especializados; todo presentado en una taxonomía general mostrando cómo se subdivide la materia y cómo encaja en el plan mayor. Los usuarios tienen que poder cambiar rápidamente del mapa temático al objeto específico de su elección. Necesitamos un software que pueda construir, en forma dinámica, presentaciones multiescalonadas con las descripciones asociadas con cada objeto del sistema. En lugar de la organización plana actual, donde todas las palabras clave se tratan de igual manera, necesitamos una organización de árbol, un sistema de clasificación o taxonomía que posibilite al usuario ingresar a cualquier nivel temático. y luego ampliar

o acotar su búsqueda a gusto, trasladándose hacia arriba o hacia abajo por el árbol de mapas.

E. Choque de culturas. Tecnócratas vs. académicos vs. vendedores. Es muy evidente que hay varios grupos profesionales diferentes que tienen ideas sobre cómo se deben organizar y manejar los recursos de Internet. Si se quiere brindar al p&u Internet. Si se quiere brindar al público general una buena atención, estos grupos necesitan formar equipos de trabajo cooperativos y escucharse entre sí. Es necesario que se pongan de acuerdo con respecto a los estándares para organizar y recuperar información, que serán de máxima utilidad para todos los grupos a largo plazo. Uno de los problemas claves en todo sistema de manejo de documentos es cómo identificar la información para que pueda ser recuperada al ser solicitada. Hay una creciente necesidad de poder filtrar lo que no se necesita.

III. Problemas: En microescala

A. Herramientas de búsqueda en Internet inadecuadas: Las herramientas de búsqueda actuales se limitan, en general, a la búsqueda de nombres de archivos y sus descriptores, hallados en los directorios de los sitios de computación indexados. Hay varias restricciones obvias en cuanto a la cantidad de información que se puede cargar en los descriptores de archivos. Sin estándares ni normas, menos aún sin un control de autoridad, un nombre de archivo como f-prot puede tener, y de hecho tiene, al menos 7 variaciones posibles. En los sistemas que buscan coincidencias literales de cadenas, esto significa que no se tienen garantías de que se vayan a encontrar todas las de que se vayan a encontrar todas los casos existentes, ni de que el ítem buscado no se halle en el sistema.

1. Lista de listas. Listas especializadas como la de Yarnoff. Hasta una tarea aparentemente simple como es localizar un servidor de listas sobre un tema particular, se convierte en una empresa mayor. Hoy en día, una simple búsqueda de cadenas a través del archivo de texto que consta de los nombres y descriptores de las listas que alguien recopiló, es la mejor herramienta disponible.
2. Usenet. Este sistema de foros de discusión ofrece sólo una organización temática jerárquica y primitiva, ya que los boletines están agrupados por amplias categorías de temas con varios niveles de subdivisión cada uno. Pero aún esto es útil. El concepto de hilos ("threads") también es interesante. En un boletín, se puede seguir un tema o hilo particular filtrando otros mensajes. La organización de base todavía es cronológica.
3. Archie [FTP]: El software Archie no ofrece mucha sofisticación o funcionalidad, pero es actualmente la única manera de buscar ítems de interés a través de sitios de archivos de Internet. El software juntarchivos de Internet. El software junta los directorios de archivos de los sitios de archivos todas las noches, y finaliza el círculo completo una vez por mes. Los ingresos obtenidos se desglosan en palabras claves, que luego se clasifican en orden alfabético. La búsqueda por palabra o frase es exacta. Pero se debe buscar cada término o frase por

separado. No se pueden hacer búsquedas con el operador lógico AND. Se limita a sitios FTP.

4. Veronica (Very Easy Rodent oriented Netwide Index to Computerized Archives). Este software solamente busca temas en los Gophers que están en redes conectadas. Otra vez, las palabras clave derivan de nombres de archivos y descriptores ingresados en los directorios. El usuario puede utilizar los operadores booleanos AND/OR/NOT y también hay un truncamiento derecho; por ejemplo, nativo o aborigen*, población* o gente*. Algunos Gophers limitan el número de caracteres en una cadena de búsqueda. Se puede acotar la búsqueda utilizando un limitador "/"; por ejemplo, por tipo de archivo. Por defecto, sólo se obtienen los primeros 200 ítems, pero se puede modificar.

5. WAIS: recuperación basada en la probabilidad. Esta es una herramienta de búsqueda más sofisticada búsqueda más sofisticada, diseñada para la recuperación basada en la probabilidad, con términos que son evaluados dentro de un grupo de bases de datos en texto completo. El software WAIS necesita más refinación y hoy se lo está desaprovechando, ya que se lo utiliza principalmente con los archivos que contienen ingresos de directorios y descriptores en ítems de menús. Se pueden seleccionar varias bases de datos para hacer la consulta, y luego refinar la cadena de búsqueda. Se clasifican los resultados, pero los algoritmos de clasificación también necesitan más refinación.

6. Metaíndices, listas orientadas a materias, páginas iniciales especializadas. Varios grupos estudiaron los problemas de localizar ítems de interés en Internet. Cern ofrece un metaíndice organizado por materia. En la Universidad de Minnesota, los estudiantes de la escuela de bibliotecarios confeccionan listas temáticas de todos los recursos, las cuales están a disposición del público. Lamentablemente, como suele pasar con los proyectos estudiantiles, la cobertura no es amplia y no hay garantía de que la información esté actualizada. Las páginas iniciales especializadas son una promesa, pero también se pra promesa, pero también se presentan problemas de cobertura, control de calidad y mantenimiento.

7. Webcrawlers, Lycos y Aliweb. Varios grupos desarrollaron herramientas automatizadas para buscar URLs e indexar todos los documentos hallados en los sitios Web. Son bastante más poderosas y, probablemente, más actualizadas, porque la actualización es automática. Pero también en este caso, se trata de una búsqueda de palabras de “fuerza bruta” en la base de datos entera, sin clasificación y, prácticamente, sin filtrado.

B. Herramientas bibliotecarias existentes limitadas. En la mayoría de los casos, los sistemas bibliotecarios actuales ofrecen búsquedas de palabras con un método booleano sofisticado, así como con un mínimo acceso a un rastreo de materias. Si bien se trabajó mucho en la asignación de números de clasificación a todos los libros y publicaciones periódicas, esta información todavía no fue explotada ni utilizada, en forma efectiva, en el software OPAC actual.

1. Software OPAC para libros y materiales audiovisuales. La mayoría de los sistemas actuales ofrecen solamente archivos índice invertidos para cada una de las marcas identificadas. Se pueden buscar todas las palabras clave de las materias. O se pueden rastrear las cadenas de materias ordenadas alfabéticamente o combinar palabras clave del título, resúmenes de materias y notas. La única búsqueda es la booleana. Si no se encuentra la palabra, la búsqueda es infructuosa. Si el usuario escribe “ataque cardíaco”, el sistema no lo lleva a “infarto de miocardio” en forma automática. Casi no hay clasificación. En el mejor de los casos, se le ofrece al usuario hojear los autores y títulos en el número de clasificación de interés. No hay títulos para los números de clasificación y no se los puede rastrear como un sistema jerárquico. El único control de autoridad que provee es para las materias.

2. Motores de búsqueda de palabras clave (para bases de datos periódicas). Si bien estos paquetes ofrecen una búsqueda booleana sofisticada, hay muy poca clasificación de términos y cuentan con una habilidad muy limitada para filtrar (salvo por idioma o año de publicación) o para ordenar los resultados. ILSA, un prototipo experimental financiado por el Consejo de Recursos Bibliotecarios, demostro la factibilidad y el valor de organizar los resultados por número de clasificactados por número de clasificación, ofreciendo por lo tanto un desglosamiento muy útil. En una búsqueda de material sobre suicidio, los resultados se relacionaban algunos con la religión, otros con la sociología y otros con la historia.

En la mayoría de los sistemas, sólo se puede consultar una base de datos por vez. Con más de 1.000 bases de datos periódicas en línea, es muy costoso y difícil garantizar que se hallen resultados sobre cualquier tema. Los problemas con la superposición y fragmentación del alcance empeoran la situación.

C. Falta de mecanismos de control de vocabulario. Si bien los bibliotecarios desarrollaron varias herramientas de control de vocabulario, como las Library of Congress Subject Headings y los tesauros múltiples para publicaciones periódicas, no hay mucho interés, ni tampoco fondos, para proyectos que mejoren o automaticen estas herramientas. De hecho, gran parte de la investigación reveló que los intentos por controlar el vocabulario no mejoraron para nada la precisión en los programas de búsqueda booleana de palabras, dentro de una determinada base de datos. Pese a esto, nadie vaticinó la explosión de bases de datos ocurrida. Ahora necesitamos con urgencia maneras efectivas de filtrar la información, con el fin de limitar el alcance de nuestras búsquedas al subconjunto de interés, antes de iniciar una búsqueda booleana. En la actualidad, no contamos con ninguna herramienta para realizar una búsqueda de “fuerza bruta” en todo el sistema y, si la tuviéramos, recuperaríamos más de lo que posiblemente utilizaríamos. Deberíamos investigar herramientas de búsqueda que tengan múltiples pasos. En el primer paso, se refina el área temática y los objetos de la información. En el segundo paso, se ahonda en una búsqueda en texto completo con los términos evaluados.

1. Los usuarios reciben muy poca ayuda para formular las consultas. Un rastreo alfabético de palabras clave o términos no es de mucha utilidad si no se tiene la palabra correcta o área temática con la cual empezar. Necesitamos tener un panorama general primero, para después aproximarnos al tema que nos interesa. No existen actualmente mapas temáticos que guíen a los usuarios. Sería muy útil si estos mapas pudieran utilizar colores para mostrar la densidad de los resultados para los diferentes términos relacionados y, a la vez, indicar los tipos de materiales. vez, indicar los tipos de materiales. Por ejemplo, habría que distinguir a las bases de datos periódicas como vínculos separados a los que el usuario puede saltar. Hay que enseñarles a los usuarios los tesauros, las enciclopedias y otras herramientas útiles de referencia, y luego ofrecerlas en texto completo a través de accesos directos de hipertexto. A veces están disponibles los tesauros en línea pero, en general, no forman parte del sistema y, en la mayoría de los casos, sólo muestran los términos controlados con términos más amplios, acotados y relacionados. No se vinculan con los números de clasificación o los términos no controlados de los resúmenes o de los documentos en texto completo. Sólo en algunas bases de datos, como Medlars, los términos controlados están conectados significativamente a un sistema de clasificación.
2. Los usuarios no pueden determinar qué términos están en uso. Hojeando un sistema de clasificación que muestre la distribución de la literatura (total de resultados), los usuarios podrían ver rápidamente cómo se diseñó un área temática y qué subtemas se desarrollaron. Podrían tener un panorama general del tema y despu& un panorama general del tema y después ingresar a un subtema particular para más detalles. Los sistemas de hoy en día no poseen estas herramientas.
3. Los usuarios no cuentan con herramientas de navegación. Con la rápida evolución de las interfases gráficas de usuario con enlaces de hipertexto a documentos en texto completo, tendríamos que ser capaces de diseñar pantallas de búsqueda mucho más flexibles, que provean acceso a una serie completa de herramientas útiles de control de vocabulario, con el fin de ayudar a formular búsquedas fructíferas. Actualmente, la mayoría de los programas sólo pide que se ingrese una cadena de búsqueda y luego devuelve el resultado.

Conclusión

Es hora de reconsiderar nuestro enfoque sobre el control de autoridad y, de hecho, nuestro enfoque total sobre el acceso temático. Deberíamos asegurarnos de que cada documento tenga incorporados ciertos descriptores clave, un número de clasificación, un tipo y el público eventual, así como también todos los detalles específicos de los contenidos, para asistir en el filtrado de documentos que es tan necesario cuando estamos saturados con el volumen de informacs saturados con el volumen de información en línea. El autor del documento es el que tendría que hacerlo primero con la ayuda de sistemas expertos. La segunda parte trata sobre los pasos que se pueden seguir para mejorar la búsqueda en Internet incluyendo el uso de la clasificación.

Bibliografía

1. Anon. "A Literature Search for information on Native Americans." Dialog(R) File 648: Trade & Industry ASAP (TM) 1994 15234861. Texto completo. Vol. 17, Núm. 2, Pág. 45 (10) "...subject access to electronic information on this network (Internet) is still primitive compared to the powerful command languages that searchers have become accustomed to with online services."
2. Bowman, C. Michael y otros. "Scalable Internet Resource Discovery: Research Problems and Approaches". Communications of the ACM. Agosto, 1994. Vol. 37, Núm. 8, Págs. 98-114. "Taxonomies allow a more uniform search space than is possible solely by content-indexing of documents". Expert system technology could be developed to match the key terms in the document against similar documents. The author could then select the class number and index terms that seemed most relevant adding more as needed. "We believe tools should be developed that allow authors to mark up their documents with cls to mark up their documents with classification terms from some selected set of taxonomies."
3. Broad, William J. "Doing Science on the Network: A Long Way from Gutenberg." New York Times, Martes, 18 de mayo de 1993, p.B 10, col. 1. "Much of the beauty and wonder of Internet and its resources...could become a horrific problem. Systems and people will shut down. I know people who have stopped using Internet because they get 500 messages a day."
4. Kubany Susan K., President of Omnet, Inc..quoted in Reinhart, Andy. "Managing the New Document" Byte. Vol. 19, Núm. 8. Agosto, 1994. "A document will no longer be a single file but rather a book of pointers to text objects, data objects, images, fonts, and so on." Pág. 93.

=====
Copyright Mary Micco 1996.

Este artículo apareció originalmente en LIBRES: Library and Information Science Electronic Journal (ISSN 1058-6768). Septiembre 1996. Vol. 6, Núm. 3.

Traducido con la correspondiente autorización de los autores.
Departamento de Informática y Sistemas.