Extracción de palabras claves de la tabla de contenido e índices analíticos: una técnica rápida para la indización de monografías en bibliotecas especializadas

Ana M. Martínez Tamayo

Cátedra de Organización del Conocimiento I, Departamento de Bibliotecología, Facultad de Ciencias Exactas, Universidad Nacional de La Plata. Calle 48 e/6 y 7, 1900 La Plata, Argentina. E-mail: ammarti@speedy.com.ar

Resumen

Se describe una técnica para la extracción de palabras claves de la tabla de contenido y el índice analítico de monografías. En una prueba realizada sobre cinco libros de Bibliotecología, se demostró que se pudo obtener un promedio de 26,4 palabras claves por libro en solo 2,1 minutos. Estas palabras claves pueden incluirse en el registro bibliográfico como términos no controlados o como un resumen y pueden servir de base para la traducción a descriptores de un tesauro. La técnica es sencilla y rápida y permite aumentar la exhaustividad y la especificidad de la indización.

Palabras claves

Indización, extracción de palabras claves, monografías

Abstract

A technique is described for extracting keywords from the table of contents and the back-of-the-book index of monographs. A test carried out on five books of library science demonstrated that an average of 26,4 keywords were obtained in only 2.1 minutes. The keywords can be included in the bibliographic record as uncontrolled terms or as an abstract, and also can be the basis for the translation to the thesaurus descriptors. The technique is simple and rapid, and increases both exhaustivity and specificity of indexing.

Keywords

Indexing, keyword extraction, monographs

Introducción

Numerosos estudios, realizados principalmente en Estados Unidos entre 1983 y 1996, permitieron establecer que la búsqueda por materia en los catálogos en línea podía presentar dos dificultades muy serias. Por un lado, las búsquedas fracasadas (0 registros recuperados) ascendían al 40% de todas las

búsqueda, mientras que el 47% presentaba un alto índice de sobrerrecuperación, es decir búsquedas con más de 50 registros recuperados. La búsqueda fracasada se ha relacionado entre otras cosas con la falta de coincidencia entre los términos del usuario y los términos de indización de la biblioteca, sobre todo por la escasa cantidad de estos últimos que suelen incluirse en un registro bibliográfico. Por su parte, la sobrerrecuperación se ha relacionado con el desconocimiento del usuario respecto a las técnicas de búsqueda y cierto rechazo al uso de operadores booleanos (Borgman, 1996; Holley, 1989; Larson, 1991; Ríos Garcia, 1991).

Una de las recomendaciones para solucionar el problema de la búsqueda fracasada ha sido el llamado *registro enriquecido*, es decir un registro bibliográfico al que se le agregan numerosos términos descriptivos de materia, ya sean controlados o no controlados. En este sentido, las bibliotecarias estadounidenses Barbara Settel y Pauline Cochrane (1982), de la Syracuse University School of Information Studies, propusieron una técnica sencilla para extraer términos significativos de materia de las tablas de contenido e índices analíticos de monografías (tratados, manuales, libros de texto, etc.) para incluirlos en el campo *653 Términos de indización no controlados* del formato MARC. Para evaluar la metodología, compararon un catálogo de 1979 monografías sobre humanidades y ciencias sociales, cuya descripción se hizo con registros de Library of Congress (promedio de 6 términos controlados) y otro catálogo confeccionado con la metodología en estudio, que se denominó *Books*, y cuyos registros contenían un promedio de 32 términos (los controlados asignados por Library of Congress y los no controlados extraídos de las tablas de contenido y los índices analíticos). En 90 búsquedas de prueba, la recuperación con los registros de Library of Congress fue de 56 registros en 8 minutos, mientras que con Books fue de 130 registros en 4 minutos, es decir más del doble de registros en la mitad del tiempo.

Revisando algunos catálogos de bibliotecas especializadas argentinas disponibles en línea, hemos observado cierta tendencia a asignar muy pocos descriptores, que obviamente representan conceptos muy generales. Sin embargo, los usuarios de estas bibliotecas suelen solicitar información por temas muy específicos que no están presentes en los registros bibliográficos, de modo que no es aventurado asumir un alto porcentaje de búsquedas fracasadas, debido justamente a la falta de coincidencia entre los términos del usuario y los que ofrece el catálogo.

Teniendo en cuenta lo anterior, el propósito de este trabajo es describir la técnica de Settel y Cochrane y mostrar sus ventajas para la indización de monografías en bibliotecas especializadas.

Metodología

a. Selección de palabras claves

En este trabajo consideramos como *palabra clave* un término no controlado extraído del título, la tabla de contenido o el índice analítico de una monografía, que se ajusta a las siguientes formas gramaticales:

- Un sustantivo sin artículo, por ejemplo *catalogación, catálogos*.
- Un sustantivo más uno o dos adjetivos, por ejemplo *tesauros multilingües, publicaciones periódicas especializadas.*
- Una de las anteriores más una frase preposicional, por ejemplo *normas de catalogación, sistemas de clasificación facetados.*

La técnica de Settel y Cochrane consiste en elegir las palabras claves de la tabla de contenido o del índice analítico de una monografía, de acuerdo con una serie de reglas que se indican a continuación.

El primer paso es analizar tanto la tabla de contenido como el índice analítico y a continuación tomar una de las siguientes decisiones:

a. Si la tabla de contenido por sí sola aporta la suficiente cantidad de palabras claves para la recuperación de la información, no se examina el índice analítico.

b. Si la tabla de contenido aporta algunas palabras claves, pero no las suficientes, se examina además el índice analítico.

c. Si la tabla de contenido no aporta ninguna palabra clave, se examina exclusivamente el índice analítico.

Si se ha tomado la decisión de extraer las palabras claves de la tabla de contenido, se procede de la siguiente manera:

1. En primer lugar, hay que decidir qué títulos de los que aparecen en la tabla de contenido se usarán como fuente para la selección de palabras claves: los títulos de tomo, parte, capítulo, sección o apartado. Por lo general, se utilizan los títulos de capítulo, pero es posible que una monografía contenga tantos capítulos que la cantidad de palabras claves resulte excesiva; entonces se puede optar por los títulos de tomos o partes. Por el contrario, si la monografía cuenta con un número escaso de capítulos, tal vez sea conveniente añadir las palabras claves de las secciones o apartados, siempre y cuando aparezcan en la tabla de contenido.

2. Seleccione solamente los términos significativos y subráyelos levemente con lápiz. Para facilitar la tarea del indizador, la biblioteca puede hacer un *stopword* o lista de palabras que no deben seleccionarse, como *prólogo, prefacio, apéndice, anexo*, etc.

3. Seleccione cada término una sola vez. Si ya lo eligió no lo repita, aunque en la tabla de contenido aparezca varias veces.

4. No seleccione los términos que requieran palabras adicionales para su comprensión

En la Fig. 1 se muestra un ejemplo de selección de palabras claves de una tabla de contenido.

Figura 1. Fragmento de una tabla de contenido en la que se han subrayado las palabras claves seleccionadas para la indización.

Si se ha tomado la decisión de extraer las palabras claves del índice analítico, se procede como sigue:

1. Seleccione los términos que tengan cinco o más páginas consecutivas.

2. Seleccione los términos con diez o más páginas no consecutivas.

3. Seleccione los términos con cinco o más subdivisiones.

En todos los casos, se subrayan los términos levemente con un lápiz.

Debe tenerse en cuenta que la cantidad de páginas y subdivisiones que se mencionan más arriba son ilustrativos, ya que cada biblioteca puede aumentar o disminuir esas cantidades. Si se aumentan las páginas y subdivisiones, la cantidad de palabras claves será menor; si por el contrario se disminuyen las páginas y subdivisiones, se obtendrá una mayor cantidad de palabras claves.

En la Fig. 2 se muestra un ejemplo de índice analítico con la selección de palabras claves de acuerdo con las instrucciones anteriores.

Diccionario literario de obras y personajes, 339, 341
Diccionarios, 85-129
alemanes, 129-122
bibliografía, 99
bilingües, 124-129
de americanismos, 106-108
de argentinismos, 108-111
españoles, 99-106
franceses, 111-115
historia, 85-88
[…]
Estados Unidos. Congress. Library, 34, 99, 160, 162, 167, 168, 194, 226, 227, 229, 230, 237, 29, 250, 251, 253, 259, 313.

Figura 2. Fragmento de un índice analítico en el que se han subrayado las palabras claves seleccionadas. las subdivisiones deben registrarse en el catálogo junto con el encabezamiento, pues de lo contrario perderían su significado, por ejemplo: *diccionarios bilingües, diccionarios españoles, diccionarios franceses.*

Las palabras claves seleccionadas deben respetarse tal cual las ha escrito el autor, pero es conveniente que se ajusten a las formas gramaticales mencionadas anteriormente. Se debe recordar que mientras mayor sea la manipulación de los términos por el bibliotecario, habrá más posibilidades de cometer errores. Se pueden permitir dos excepciones:

1. Cuando en la tabla de contenido o en el índice analítico se encuentra un término, compuesto por un sustantivo con dos o más adjetivos o viceversa, que en realidad representan varios conceptos y pueden afectar la recuperación, por ejemplo: *bibliotecarios uruguayos y argentinos* debe sustituirse por dos términos: *bibliotecarios uruguayos* y *bibliotecarios argentinos*. O bien *catálogos y bibliografías en línea*, que debe sustituirse por *catálogos en línea* y *bibliografías en línea*. Pero deben mantenerse unidos los términos compuestos que se refieren a un único concepto, aunque sean demasiado extensos, por ejemplo *listas de encabezamientos de materia*.

2. Cuando la tabla de contenido o el índice analítico se encuentran en otro idioma, las palabras claves pueden traducirse al español. En este caso, es indispensable que el indizador tenga conocimientos suficientes del idioma extranjero, pues de lo contrario la tarea se demoraría demasiado y se perdería una de las ventajas de la técnica que es su rapidez.

b) Registro de las palabras claves en el catálogo

Una vez que se han seleccionado las palabras claves, se pueden hacer con ellas tres cosas:

1. Registrar las palabras claves como términos no controlados, en un campo repetible. Si la biblioteca usa formato Marc21, podrá registrar las palabras claves en el campo *653 Términos de indización no controlados*. Si utiliza el formato Bibun, podrá hacerlo en el campo *62 Palabras claves*. Los formatos Focad y Cepal no permiten registrar términos no controlados, de modo que la biblioteca tendría que

agregar a su base de datos un campo fuera de formato. Más adelante se presenta un ejemplo en la Fig. 3.

2. Redactar un resumen con las palabras claves. Para la redacción del resumen, es conveniente comenzar con una frase introductoria, que puede incluir una referencia a la forma del documento, por ejemplo: *este manual trata sobre [...]* y agregar a continuación las palabras claves separadas por comas. El resumen se registra como sigue: Marc21 campo 520, Bibun campo 69, Focad campo 69, Cepal campo 72. Más adelante se presenta un ejemplo en la Fig. 4.

3. Traducir las palabras claves a descriptores de un tesauro. En este caso, la técnica de Settel y Cochrane equivaldría al análisis conceptual. La forma de registrarlos sería en los campos: 650 de Marc21, 65 de Bibun, 65 de Focad, 76 y 77 de Cepal. Más adelante se presenta un ejemplo en la Fig. 5.

c) Utilidad de la técnica de Settel y Cochrane

Para mostrar la utilidad de esta técnica, se identificaron cinco libros en los catálogos de siete bibliotecas especializadas argentinas, disponibles en línea (Lancaster, 1995a y b; Sabor, 1978 y 1984; Dobra, 1997). Se verificó la cantidad de descriptores asignados a dichos libros en cada una de las bibliotecas:

Por otro lado, un bibliotecario entrenado en la técnica de Settel y Cochrane y con experiencia en su aplicación, extrajo las palabras claves de la tabla de contenido de los cinco libros, midiendo el tiempo que demoró esta selección. Además, basándose en el libro de Lancaster (1995b), completó los siguientes ejemplos: a) registro de palabras claves en el catálogo (Fig. 3), b) redacción y registro de un resumen en el catálogo (Fig. 4) y c) traducción de las palabras claves seleccionadas a términos controlados del tesauro *Tesauro de bibliotecología y documentación* (Mochón y Sorli, 2005) (Fig. 5). En todos los ejemplos mencionados se asumió que la biblioteca usaba formato Marc21.

Resultados y discusión

En la Tabla 1 se presenta una comparación entre la cantidad de descriptores asignados por las bibliotecas estudiadas y la cantidad de palabras claves obtenidas con la técnica de Settel y Cochrane, así como el tiempo empleado para la selección de estas palabras claves. De dicha tabla se obtuvieron los siguientes promedios: la demora en la selección de palabras claves fue de 2.1 minutos por libro y se extrajeron 26,4 palabras claves por libro, frente a 2,8 descriptores asignados por las bibliotecas estudiadas, cifras que equivalen a decir que en aproximadamente 2 minutos se registraron 9 veces más términos descriptivos de materia por registro.

Tabla 1. Comparación de la cantidad de descriptores asignados a una muestra de libros y las palabras claves extraídas de las tablas de contenido correspondientes.

Palabras claves extraídas de la tabla de contenido			Descriptores asignados	
Libros	Demora en minutos	Cantidad de palabras claves	Biblioteca	Cantidad de descriptores
Lancaster (1995a)	1½	14	B D F G	2 3 4 1
Lancaster (1995b)	2	26	C E F G	5 3 7 1
Sabor (1978)	2	15	A B D G	1 2 4 1
Sabor (1984)	2½	33	B F G	5 4 1
Dobra (1997)	2½	23	G	1

Dos bibliotecas (A y G) sólo asignaron un término descriptivo de materia por libro; 4 bibliotecas (B, C, D y E) asignaron entre 2 y 5 términos y una biblioteca (F) asignó 7 términos en uno de los tres libros que indizó.

En la Tabla 2 (véase al final del trabajo) se muestra la diferencia de términos disponibles para la búsqueda entre a) los descriptores asignados según la práctica habitual de las bibliotecas estudiadas, b) las palabras claves aportadas por la técnica de Settel y Cochrane, usando solamente la tabla de contenido y c) los descriptores obtenidos al traducir las palabras claves a los términos del tesauro. Como se puede observar, frente a la suma de 10 descriptores asignados por todas las bibliotecas estudiadas para el libro de Lancaster (1995b), se extrajeron 26 palabras claves, de las cuales 19 (73%) pudieron traducirse exactamente a descriptores del *Tesauro de bibliotecología y documentación*, con lo cual el registro bibliográfico de la Fig. 5 presenta prácticamente el doble de descriptores que los asignados por todas las bibliotecas estudiadas.

Más importante que la cantidad resulta la calidad de los términos. De los 10 descriptores asignados por las bibliotecas, 6 no están contemplados en la tabla de contenido del libro (Bibliotecología, Bibliotecología y ciencia de la información, control terminológico, descripción de documentos, informática y lingüística). Cabría preguntarse entonces si están verdaderamente tratados por el autor Algunos de ellos son muy generales y otros francamente incorrectos como *descripción de documentos*, ya que el libro no trata sobre este tema, sino sobre la construcción de vocabularios controlados.

A continuación se presentan tres figuras. En la Fig. 3 se muestra un ejemplo de registro en el campo *653 Términos de indización no controlados* del formato Marc21. En la Fig. 4 se puede observar el registro de un resumen en el campo *520 Resumen* del mismo formato. Finalmente, en la Fig. 5 aparece el registro de descriptores en el campo *650 Término de materia* de Marc21.

```
#0$aactualización
#0$acompatibilidad
#0$aconstrucción de un tesauro
#0$acontrol del vocabulario
#0$acoste-beneficio
#0$adirectrices
#0$aestructura del vocabulario
#0$aevaluación de los tesauros
#0$ahomografía
#0$aidentificadores
#0$alenguaje natural
#0$alistas de control
#0$anormas
#0$anotas de aplicación
#0$apresentación del tesauro
#0$aprocesamientos automáticos
#0$arecuperación de información
#0$arelación asociativa
#0$arelación jerárquica
#0$asistemas de recuperación
#0$asistemas postcoordinados
#0$asistemas precoordinados
#0$atérminos
#0$atesauro
#0$avocabulario de entrada
#0$avocabulario postcontrolado
```

Figura 3. Registro de palabras claves extraídas de la tabla de contenido del libro Lancaster (1995b), en el campo 653 de Marc21.

```
3#$aEste manual trata sobre actualización,
compatibilidad, construcción de un tesauro, control del
vocabulario, coste-beneficio, directrices, estructura
del vocabulario, evaluación de los tesauros, homografía,
identificadores, lenguaje natural, listas de control,
normas, notas de aplicación, presentación del tesauro,
procesamientos automáticos, recuperación de información,
relación asociativa, relación jerárquica, sistemas de
recuperación, sistemas postcoordinados, sistemas
precoordinados, términos, tesauro, vocabulario de
entrada, vocabulario postcontrolado.
```

Figura 4. Registro de un resumen redactado a partir de las palabras claves extraídas de la tabla de contenido del libro Lancaster (1995b), en el campo 520 de Marc21.

```
7$aautomatización de tesauros$2tbd
#7$acatálogos de autoridades$2tbd
#7$acontrol del vocabulario$2tbd
#7$adirectrices$2tbd
#7$aelaboración de tesauros$2tbd
#7$aidentificadores$2tbd
#7$alenguajes de recuperación$2tbd
#7$alenguajes naturales$2tbd
#7$alenguajes postcoordinados$2tbd
#7$alenguajes precoordinados$2tbd
#7$anormas$2tbd
#7Sapresentación del tesauro$2tbd
#7$arecuperación de información$2tbd
#7$arelaciones asociativas$2tbd
#7$arelaciones de equivalencia$2tbd
#7$arelaciones jerárquicas$2tbd
#7$asistemas de recuperación de información$2tbd
#7$atérminos$2tbd
#7$atesauros$2tbd
```

Figura 5. Registro de descriptores asignados a partir de las palabras claves seleccionadas de la tabla de contenido del libro Lancaster (1995b), siguiendo el formato Marc21. tbd: *Tesauro de bibliotecología y documentación.*

Conclusiones

La técnica de Settel y Cochrane es útil, porque permite mejorar la exhaustividad de la indización, agregando una cantidad importante de términos descriptivos de materia en muy poco tiempo y a bajo costo. Ha sido utilizada con éxito en diversas experiencias como lo expresan las propias autoras (Settel y Cochrane, 1982).

Es también una técnica versátil, pues permite registrar términos no controlados, redactar un resumen o traducir las palabras claves a los descriptores de un tesauro. Además aporta mayor especificidad a la indización, porque al utilizar la misma terminología del autor se evitan desvíos, confusiones o malas interpretaciones del bibliotecario.

Por último, algunas bibliotecas especializadas argentinas están familiarizadas con prácticas similares, aunque no estandarizadas. La técnica de Settel y Cochrane facilita la formalización de los procedimientos de indización y en consecuencia, la capacitación en servicio del bibliotecario y la evaluación de su desempeño.

La técnica puede ser fácilmente adaptada a otros tipos de documentos, como los artículos de revistas científicas, utilizando como fuentes el título y el resumen en vez de la tabla de contenido y el índice analítico.

Referencias bibliográficas

- Borgman, Christine L (1996). Why are online catalogs still hard to use. En: *Journal of the American Society of Information Science*, vol. 47, no. 7, p. 493-503.

- Dobra, Ana (1997). La biblioteca popular, pública y escolar. 2 ed. Buenos Aires: Praxis.

- Holley, Robert R (1989). *Subject access in the online catalog*. New York : Haworth Press.

- Lancaster, Frederic W. (1995a). Indización y resúmenes: teoría y práctica. Buenos Aires: EB Publicaciones.

- Lancaster, Frederic W. (1995b). El cntrol del vocabulario en la recuperación de información. Valencia: Università de Valencia.

- Larson, Ray R. (1991). The decline of subject searching: long term trends and patterns of index use in an online catalog. En: *Journal of the American Society of Information Science*, vol. 42, no. 4, p. 197-215.

- Mochón Bezarez, Gonzalo y Angela Sorli Rojo (1995). Tesauro de biblioteconomía y documentación [en línea]. 2ª ed. Madrid: CINDOC, 1995 [Consulta 27 Oct 2008]. Disponible en la Word Wide Web http://thes.cindoc.csic.es/index_BIBLIO_esp.html.

- Ríos García, Yolanda (1991). Catálogos en línea de acceso público: selección bibliográfica. En: *Revista Española de Documentación Científica*, vol. 14, no. 2, p. 121-141.

- Sabor, Josefa E. (1978). Manual de fuentes de información. 3 ed. Buenos Aires: Marymar, 1978.

- Sabor, Josefa E., coord. (1984). Manual de bibliotecología. México: Kapelusz.

- Settel, Barbara y Pauline A. Cochrane (1982). Augmenting subject descriptions for books in online catalogs. En: *Database*, 1982, vol. 5, no. 4, p. 29-37.

Tabla 2. Comparación entre las palabras claves extraídas de la tabla de contenido del libro de Lancaster (1995b), su traducción a descriptores del Tesauro de Bibliotecología y Documentación y los descriptores asignados por las bibliotecas estudiadas.

Aplicación de la técnica de Settel y Cochrane		Bibliotecas estudiadas	
Palabras claves de la tabla de contenido	Traducción a descriptores	Descriptores asignados	Bibliotecas
actualización	-	bibliotecología	C
compatibilidad	-	bibliotecología y ciencia de la información	F
construcción de un tesauro	elaboración de tesauros	control terminológico	E, F
control del vocabulario	control del vocabulario	descripción de documentos	F
coste-beneficio	-	informática	C
directrices	directrices	lingüística	C
estructura del vocabulario	-	recuperación de información	C, E, F
evaluación de los tesauros	-	términos de indización	F
homografía	-	tesauros	C, E, F
identificadores	identificadores	vocabulario de lenguajes de indización	F
lenguaje natural	lenguajes naturales		
listas de control	catálogos de autoridades		
normas	normas		
notas de aplicación	-		
presentación del tesauro	presentación del tesauro		
procesamientos automáticos	automatización de tesauros		
recuperación de información	recuperación de información		
relación asociativa	relaciones asociativas		
relación jerárquica	relaciones jerárquicas		
sistemas de recuperación	sistemas de recuperación de información		
sistemas postcoordinados	lenguajes postcoordinados		
sistemas precoordinados	lenguajes precoordinados		
términos	términos		
tesauro	tesauros		
vocabulario de entrada	relaciones de equivalencia		
vocabulario postcontrolado	lenguajes de recuperación		