Calidad de la indización en bases de datos de tamaño cualesquiera

Enrique Wulff Barreiro
IEDHC - CSIC
Valencia

Donde se identifica un criterio para establecer la mejor calidad en la indización de bases de datos evaluando tres dimensiones, consistencia, discriminación y exhaustividad.

To determine a criteria for establishing the best quality of indexing for databases three dimensions of indexing were evaluated: consistency, discrimination and exhaustivity.

1. Introducción

De convenir con Schumpeter en que la competencia que cuenta es la que da lugar a superioridad en términos de calidad, cabría afirmar que la amplia distribución electrónica de datos ha creado la demanda de un mayor estándar de calidad que el que se esperaba en los materiales tradicionales impresos. Esta demanda de mejor calidad, al tiempo que la complejidad del proceso de publicación de información electrónica y, junto al valor económico de las decisiones tomadas en base a la información online, ha significado una preocupación significativamente más alta en cuanto a la fiabilidad de la información.[1] Con todo, esta perspectiva frecuentemente no se comprende bien, interpretándola en términos de gestión de sistemas de información (MIS), un terreno donde combaten en pos del dominio, industriales, ingenieros de ordenadores y administradores comerciales.[2] No obstante, en muchos países la comunidad de la información no está plenamente reconocida y debe buscar llegar a convertirse en un grupo de presión real a nivel nacional, capaz de sensibilizar a quiénes toman las decisiones acerca de las opciones en el terreno de la información.[3] Como el tamaño del sector privado en el mercado de la información de los países industrializados no deja de

1. TARTER, Blodwen, *Information liability: New interpretations for the electronic age,* Golden Gate University, 1991 (Tesis Univ.)
2. FERREIRO, Soledad, "Information management" En: *FID News Bulletin,* (1992), vol 42, 5, p. 125-127
3. YUSHKIAVITSHUS, Henrikas, "Information Policy" En: *FID News Bulletin,* (1992), vol 42, 5, p. 128-130

crecer, merced a la vigente conceptualización cultural de la información como mercancía, los bibliotecarios parecen quedar un tanto ignorados por los entusiastas de los ordenadores. Además, al ser en buena parte la tecnología información, y por adaptarse el usuario a la estructura de las herramientas que se ponen a su disposición,[4] parece que el vendedor dispone de importantes ventajas a partir del conocimiento que posee, a menudo excluyente, de las características de la tecnología (de la información) y de su control sobre la comunicación previa de las mismas.[5]

Aún a pesar de su baja visibilidad, en términos de estereotipo futurista, la comunidad dispone en sus bibliotecarios de una red de gente experimentada que siempre ha hecho honor a una ética de carácter fuerte en el servicio público. El hecho de que haya quien no perdone su viabilidad comercial a empresas que demuestran ser rentables desde el primer día sobre un terreno antes sólo frecuentado por entidades oficiales, puede que tenga que ver con que haciendo esto delimitan un margen escasamente operativo para que emerjan las posibilidades de un servicio de referencia, de información, propiamente público. La externalización de servicios de información hacia el sector privado con el pretexto de que los canales de comunicación de carácter tecnológico son allí viables, violenta una jerarquía intelectual entre datos e ideas, hechos y conocimiento de la que los hombres y mujeres que forman el personal bibliotecario configuran su sentido más saludable.[6] La larga experiencia manos a la obra a la caza de la información, presentada en la forma más adecuada, en el lugar más apropiado, garantiza no sólo que los ordenadores podrían generar más información para el público, sino que la propia información ocuparía un lugar apropiadamente subordinado en la cultura. Llama la atención, y no es lo único, que los brokers de información mantengan bases de datos para búsqueda de socios en consorcios de resultados limitados al 1% a la hora de comprobar los acuerdos realizados gracias al servicio de la base de datos, y que sea la base de datos "Eurocontact", en el marco del programa Esprit, la que consiga una metodología para asistir a las empresas a identificar socios con un nivel medio de eficacia del 15%.[7]

2. Bases

Nosotros, en un intento por dar cuenta de la gestión de la información en términos de calidad, vamos a incorporar un modelo de evaluación de la calidad de la indización en bases de datos, partiendo de la paradoja de la recuperación de la información: "la necesidad de describir aquello que se desconoce con el propósito de buscarlo".[8] Es decir, partimos de que formulaciones como 'no puedo describirlo, pero lo reconoceré cuando lo vea', y 'Déme más de esto!', ponen a prueba nuestras capacidades cognoscitivas y perceptuales: de que resulta de interés averiguar lo que está disponible en la base de datos antes de intentar establecer una descripción de lo que se nos pide.

4. HANCOCK-BEAULIEU, Micheline, *Subject searching behaviour at the library catalogue and at the shelves: Evaluating the impact of an online public access catalogue,* London, The city University, 1989 (Tesis Univ.)
5. VEGARA, Josep Mª, *Ensayos económicos sobre innovación tecnológica,* Madrid, Alianza, 1989
6. ROSZAK, Theodore, *The cult of information: The folklore of computers and the true art of thinking,* Cambridge, Lutterworth, 1986
7. CONROY, John, "Búsqueda de socios" En: *XIII Magazine,* (1992), nº 7, p. 10-13
8. HJERPPE, Roland, "Project HYPERCATalog: visions and preliminary conceptions of an extended and enhanced catalog" En: BROOKES, B.C. (ed.), *Intelligent Information Systems for the Information Society,* Elsevier Science Publishers B.V., 1986, p. 211-232

El análisis determina si una base de datos indiza una o varias áreas temáticas mejor que las demás bases de datos, al tiempo que indica la optimización de la indización alcanzada. Los resultados van a ser útiles a: 1) gestores de bibliotecas, profesionales de la búsqueda online o especialistas de la información a la hora de seleccionar bases de datos online para suscribirse a ellas o para hacer búsquedas, y 2) productores de bases de datos online para llamar su atención sobre áreas donde resulta necesaria una mejora en la indización.[9]

3. Análisis

Aquí se emplean racimos de documentos sobre la base de criterios de similitud de contenido. Para seleccionar los documentos con vistas a formarlos, la propuesta consiste en el empleo de expertos en la materia y de los autores para que interpreten la relevancia temática. Naturalmente los documentos no se seleccionarán de ninguna de las bases de datos haciendo uso de su propia terminología de indización, pues es ésta la que se trata de evaluar. Se estiman de 3 a 8 documentos por racimo.

Se van a evaluar tres dimensiones de la indización que representan facetas relevantes de la calidad de la indización que se trata de medir. Primera determinar el modo en que los vocabularios controlados de las bases de datos conectan documentos afines, identifica la forma en que una base de datos suele representar el contenido temático que tienen en común los documentos de cada racimo. Segunda, discriminar en términos generales entre subconjuntos conectados de documentos afines, entendiendo por tal que los términos de indización describan la mitad o más de los documentos de un racimo y se utilicen infrecuentemente en la base de datos. El valor de discriminación se calculará, obteniendo el número de veces en que se ha asignado a documentos cada término de indización a la hora de representar dos o más documentos en un racimo. Por último, se propone el cálculo del número medio de términos de indización asignados a un documento, para dar cuenta de lo accesible que está un documento y lo exhaustivo de la indización (progresando ambos conceptos con el aumento en el número de términos de indización asignados).

Al comienzo del estudio, se determinarán el número y alcance de los racimos a utilizar. Los problemas que pueden presentársenos al recuperar los documentos en estos racimos son:

- puede que en alguna de las bases de datos no haya control para los nombres de autor. Tal vez una búsqueda por títulos localice el item.

- la cobertura de los informes puede ser muy pobre.

- la cobertura retrospectiva de las bases de datos puede ser buena aunque lenta.

Una vez hecho esto, se obtiene por cada item la indización temática completa. Se ordenarán los términos de indización alfabéticamente, determinando los que son independientes, y para estos se calculará el número de veces en que han sido asignados a documentos. Las bases de datos observan sistemas de indización diferentes, se considerará pues, término de indización cualquiera que haya sido asignado para representar un concepto en un documento. Cabe optar por distinguir entre términos y no entre conceptos, por lo que vamos

9. CHU, Clara M. y AJIFERUKE, Isola, "Quality of indexing in library and information science databases" En: *Online Review*, (1989), vol. 13, nº 1, p. 11-35. (De este trabajo se ha hecho un empleo general).

a considerar a los términos específicos y sinónimos, que vienen dados por las relaciones que formulan los tesauros, términos independientes.

Atendiendo a la primera dimensión de la indización, vamos a identificar los términos que abarcan dos o más documentos, anotando la frecuencia de aparición de un término en el racimo. Se dirá que los términos de indización se ajustan a la medida de la consistencia, esto es que capturan la similitud del contenido temático de los documentos de cada racimo, cuando están asignados a la mitad o más de los documentos en un racimo.

Para obtener el valor de discriminación, y dar cuenta de la segunda dimensión de la indización, se especifica si el término de indización discrima muy por encima, por haber sido asignado a muy escasos documentos; bien, si fue asignado a un pequeño número de documentos; o de manera pobre, si se asignó a un gran número de documentos. Cabe emplear dos medidas:

Índice A:

Índice de Discriminación (Término A) = $1 / \log_{10}$ Nº veces en que el Término A ha sido asignado a documentos

Éste índice produce valores entre 0 y 1 con un valor umbral de discriminación de 0.25, indicando un valor superior una mejor discriminación. Es apropiado para evaluar bases de datos de tamaños comparables. Para comparar bases de datos de tamaños cualesquiera cabe hacer uso de:

Índice B:

Índice de Discriminación (Término A) = Nº de veces en que ha sido asignado a documentos el Término A / Tamaño de la base de datos

Este segundo índice también da valores entre 0 y 1, aunque las mejores discriminaciones vienen dadas por los valores más bajos, el valor umbral de discriminación es 0.05.

Hasta aquí, y para saber qué términos se portan bien en las dos dimensiones de indización evaluadas, esto es, una vez más, a la hora de conectar documentos afines y de discriminar entre estos documentos por toda la base de datos, se calcularán para cada índice dos tipos de términos de alcance/discriminación. Los términos A/D del tipo 1 son términos que abarcan la mitad o más de los documentos en el racimo y tienen un valor por encima del umbral 0.25 para el índice A y por debajo de 0.05 para el índice B. Los términos A/D del tipo 2 son los que abarcan la mitad o más de los documento en el racimo y tienen un valor entre 0.25 y 0.75 para el índice A y por debajo de 0.05 para el índice B. El tipo 1 sólo descarta los términos que discriminan de forma pobre. El tipo 2 es una medida más fina pues establece límites superior e inferior y prescinde de aquellos términos que discriminan en exceso o pobremente.

Por último, y para atender a la tercera dimensión de la indización, se calcula la medida de exhaustividad, el número medio de términos empleados para describir un documento en un racimo para cada base de datos. Parece que entre 8 y 12 términos por racimo es una medida apropiada de exhaustividad de la indización.

4. Resultados

Como resultado cabe decir, teniendo en cuenta el índice B que tiene presente el tamaño de la base de datos, y los términos A/D de tipo 2, que probablemente esté

10. KIANG, Yihwa, *Exploration of qualitative reasoning and its applications to management,* The University of Texas at Austin, 1991 (Tesis Univ.)

bien indizado un racimo de documentos afines que reciban cada uno una media de entre 8 a 12 términos asignados, donde dos o tres términos abarquen todos los documentos y entre tres a seis términos se estimen A/D de tipo 2. He aquí, entonces, un criterio para establecer la mejor calidad en la indización de bases de datos.

5. Conclusiones

Es notable, ya lo vemos, el interés que tiene incorporar información adicional cuantitativa y capacidad matemática para dotarse de una metodología de razonamiento causal y cualitativa a modo de herramienta que apoya la toma de decisiones con valor añadido en el dominio de la información.[10] Este es un resultado que un bibliotecario está llamado a conseguir, como así lo apunta la investigación reciente.[11] Eludirlo conduce, posiblemente, hacia ofuscados panoramas (como el actual alrededor del MARC para holdings?) donde los consorcios industriales cortan por lo sano (en términos de externalidades!?).[12]

11. THOMPSON, Christiane Epps, *Hard science or soft science: A bibliometric analysis of selected library science-information science journals,* Texas Woman's University, 1989 (Tesis Univ.)
12. WEISS, Martin y CARGILL, Carl, "Consortia in the Standards Development Process" En: *Journal of the American Society for Information Science,* (1992), vol. 43, nº 8, p. 559-565

Grupo Distribuidor Editorial

❑ CATEDRA ❑ PIRAMIDE ❑
❑FUNDACION GERMAN SANCHEZ RUIPEREZ ❑
❑TECNOS ❑ EUDEMA ❑ ED. EL ARQUERO ❑
❑ ANAYA MULTIMEDIA ❑ VERSAL ❑
❑ BARCANOVA ❑ EDICIONS XERAIS DE GALICIA ❑
❑ ANAYA ❑ ALIANZA EDITORIAL ❑
❑ ANAYA & MARIO MUCHNIK ❑ BIBLIOGRAF ❑ AURA ❑

GRUPO DISTRIBUIDOR EDITORIAL, S.A.

Ferrer del Rio, 35
28028 MADRID
Tel. (91) 361 08 09 (6 líneas)
Fax (91) 356 57 02

Polígono Pisa
C./ Horizonte, parcela 16
41927 - MAIRENA DEL
ALJARAFE (Sevilla)

Polígono Industrial
(LA UNIDAD-ASEGRA)
C/. Málaga, s/n Parcela 8 D
PELIGROS (Granada)
Tel. (958) 40 52 49
Fax (958) 40 26 30