Evaluación de bases de datos sobre el tema recuperación de la información

José Antonio Salvador Oliván
José María Angós Ullate
Universidad de Zaragoza

0.1. Resumen

El estudio compara la calidad de indización de las bases de datos online ERIC, LISA e ISA, e intenta resumir las líneas de investigación en *Recuperación de la Información*. Para ello se ha realizado una búsqueda bibliográfica sobre recuperación de la información en las bases de datos mencionadas anteriormente en el distribuidor Knight-Ridder, durante el período de tiempo de Enero de 1995-Junio de 1997. Se evalúan comparativamente las características de 26 referencias de documentos, que son los mismos para las tres bases de datos. Los resultados muestran que la indización en LISA es poco exhaustiva y específica, y en la que mayor número de errores se producen. ERIC e ISA, son las que mayor grado de afinidad tienen, con una indización más exhaustiva, específica y precisa.

Palabras clave: Recuperación de la información. Evaluación. Indización. Bases de datos online. LISA. ERIC. ISA.

0.2. Summary

The study compares the quality of indexing in the ERIC, LISA and ISA databases, and outlines the research subjetcs on *Information Retrieval*. A bibliographical search on information retrieval in the previoulsy mentioned databases of the Knight-Ridder host, during the period of time of January of 1995-June of 1997, requiring that the term only appeared in the title has been carried out. A comparison was conductes using tee same references of documents (26) in the three databases. The results shown that LISA databasc had a minor level of indexing exhaustivity and specifity, which more errors ocurrs. ERIC and ISA databascs are more similars, with a higher level of exhaustivity, specifity and accuracy.

Keywords: Information retrieval. Evaluation. Indexing. Online databases. LISA. ERIC. ISA.

1. Introducción

La evaluación de las bases de datos ha sido un tema ampliamente estudiado en la literatura mundial. Son muchos los factores que pueden afectar al rendimiento de los sistemas de recuperación, por lo que se hace necesario el estudio de estos factores para identificar donde pueden ocurrir los fallos en la recuperación, y que afectarán de forma negativa tanto a la llamada como a la precisión, dos de las medidas más frecuentemente utilizadas en la evaluación de dichos sistemas.

Por otra parte, el aumento en el número de bases de datos disponibles y accesibles tanto para el usuario final o para el intermediario, en el caso de que la búsqueda de información sea delegada, tiene como resultado que para un mismo tema puedan ser consultadas varias bases de datos, lo que produce algunas veces, en la persona que va a realizar la búsqueda, el sentimiento de inseguridad o de indecisión para seleccionar aquella(s) base(s) de datos que mejor van a satisfacer la necesidad de información del usuario. Por ello, todos aquellos estudios encaminados a la evaluación de las bases de datos, a un mejor conocimiento de sus características, ayudarán a tomar una decisión correcta a la hora de seleccionar la base de datos más idónea, así como a detectar los fallos del sistema y, como consecuencia, a intentar solucionarlos, lo que implicará un mayor y mejor rendimiento del sistema y una mayor eficacia en la búsqueda.

2. Objetivos

Uno de los principales factores que afectan a la recuperación eficaz y eficiente de la información es sin lugar a dudas el relativo a la indización de los documentos. Nuestro objetivo es estudiar algunos aspectos relacionados con la indización en las bases de datos más representativas en el ámbito de las Ciencias de la Documentación; dichos aspectos incluyen la calidad de la indización, la exhaustividad y la especificidad.

Por otra parte, uno de los principales temas de investigación en nuestro campo es todo lo relacionado con la "recuperación de información", de manera que se ha elegido este tema para evaluar la indización en las bases de datos, y a la vez, para conocer qué descriptores o términos se utilizan con más frecuencia sobre este tema, lo que nos indicará aquellos aspectos de la recuperación de información que más centran el interés de los investigadores.

3. Metodología

Para crear el conjunto de registros sobre los que se realiza este estudio, se han llevado a cabo los siguientes pasos:

- En primer lugar, se planteó una búsqueda de información sobre "*recuperación de la información*" en el distribuidor norteamericano Knight-Ridder

en las bases de datos LISA (Library and Information Science Abstracts), ISA (Information Science Abstracts) y ERIC (Educational Resources Information Center). Las razones que nos han llevado a seleccionar estas bases de datos han sido el considerar a LISA e ISA como las más representativas e importantes en el ámbito de la Ciencia de la Información y Documentación, y ERIC por constituir la principal fuente de información en educación.

- En segundo lugar, se interrogó en las citadas bases de datos por la frase *INFORMATION (W) RETRIEVAL*, realizándose la búsqueda en Julio de 1997. Se limitó a que aparecieran dichos términos solamente en el título, de manera que presumiblemente aquellos registros recuperados serían idóneos y altamente pertinentes para los propósitos de este estudio. El número total de referencias fue de 4157, por lo que se redujo el período de búsqueda a las publicaciones realizadas en los años 1995 hasta julio de 1997 e introducidas en las bases de datos, quedando un total de 257 referencias. La reducción a este período de tiempo se debió a considerar que en los dos últimos años y medio es un período de tiempo suficiente para analizar y descubrir las cuestiones principales, investigaciones recientes y avances que se producen en el campo de la recuperación de la información, y sobre todo, que equivalía a un coste económico razonable y asumible.
- Se creó una base de datos con estas 257 referencias, de las que 145 pertenecían a LISA, 80 a ISA, y 32 a ERIC. Para comparar la indización en las tres bases de datos, pensamos que lo ideal sería analizar únicamente aquellas referencias que estuvieran presentes en las tres bases de datos, evitando de esta manera cualquier posible fuente de sesgo que se produce cuando se comparan diferentes casos, en este caso, documentos. De esta manera, los resultados de los análisis reflejan con más precisión las diferencias que pueda haber entre las bases de datos. Por consiguiente, sólo se tuvieron en cuenta para evaluar la indización aquellas referencias que coincidían en las tres bases de datos, y que suponían un total de 26. Para el otro objetivo de este trabajo, el conocer qué aspectos de la recuperación de información se investigan con más frecuencia, se analizaron las 256 referencias, que suponían un total de 157 referencias únicas, eliminando las repetidas entre dos o las tres bases de datos.
- De cada una de las referencias bibliográficas se introdujo en la base de datos la información pertinente. Por una parte, se seleccionaron variables que podemos clasificar como identificativas o descriptivas de cada referencia, como nombre de la revista, autores, fecha de publicación, y base de datos en la que aparecía. Y por otra parte, se seleccionaron las variables relacionadas con la parte temática y contenido del documento: título, des-

criptores que aparecen, identificadores en su caso, y resumen.

- Posteriormente, se eligió el método para estudiar la calidad de la indización. La *especificidad en la indización* la medimos examinando si el concepto utilizado para la búsqueda (information retrieval), y que pensamos que tiene entidad propia, aparece como descriptor, lo que seria la situación normal y deseable. La *profundidad de la indización* la medimos simplemente contando el número de términos asignados a cada documento. La *exhaustividad de la indización*, que refleja el grado con que se representan todos los conceptos, se ha medido solamente con el título del documento. Se ha realizado un análisis de todos los conceptos importantes que aparecen el título, y se ha contabilizado si aparecen en el campo de descriptores.
- Por último, se realizó el proceso de datos con el paquete estadístico SPSS al objeto de analizar los datos.

4. Resultados

4.1. Factores de indización

Antes de comenzar a presentar y analizar los resultados obtenidos, conviene mencionar que LISA utiliza como vocabulario controlado una lista de descriptores y el procedimiento de indización en cadena; en ISA se utiliza como vocabulario controlado un esquema de clasificación, y en ERIC se utiliza el Tesauro de Descriptores de ERIC, contando además esta base de datos con un campo de identificadores.

4.1.1. Especificidad en la asignación de términos

La especificidad es el factor más importante que influye en la precisión. La falta de especificidad puede causar fallos tanto en la llamada como en la precisión. En ERIC aparece siempre el término information retrieval como descriptor (100%), en LISA aparece en 13 referencias, lo que supone sólo un 50%, y en ISA aparece en 25 registros (96,1%). (Fig. 1). Esto indica que es en la base de datos ISA donde más errores se producen por parte del indizador, ya que un a un concepto tan importante que aparece en el título se le debe de asignar un término índice en el campo de descriptores.

Con respecto a la especificidad del lenguaje índice que se utiliza en las diferentes bases de datos, podemos mencionar lo siguiente:

- En ERIC se emplea el descriptor *information retrieval* como término con entidad propia, y además aparece siempre como descriptor importante (marcado con un asterisco).

- En LISA aparece 1 vez como *information retrieval*, 9 como *online information retrieval*, 2 veces como *computerized information retrieval*, y 1 vez como *computerized information storage and retrieval*. Observamos cómo existe una gran variedad de términos para representar un concepto, siendo además la única base de datos en la que existe un descriptor en el que no aparecen las palabras information y retrieval una a continuación de la otra, lo que debe de ser tenido en cuenta a la hora de utilizar operadores de proximidad.
- En ISA aparece como descriptor en forma de *information retrieval* (18 registros) o de information retrieval systems (7 registros).

Así, es importante conocer el lenguaje de indización utilizado en cada base de datos para una recuperación eficaz, ya que si la búsqueda la hubiéramos realizado exigiendo que *information retrieval* apareciera tal y como está (un término al lado del otro en ese mismo orden) en el campo de descriptores, en la base de datos LISA hubiéramos perdido 1 registro. Igualmente, es de gran importancia la precisión a la hora de asignar los términos índice por parte de los indizadores, ya que hubiéramos perdido en LISA al buscar sólo en el campo de descriptores el 50% de las referencias, cifra que parece exageradamente alta.

4.1.2. Profundidad de la indización

Se ha contado el número de términos asignados a cada una de las referencias. Esta medida nos va a permitir conocer con qué grado se representan los diferentes temas tratados en los documentos. En la figura 1 se muestra la estadística descriptiva del número de términos por registro. Como en ERIC existe un campo de identificadores que no existen en las otras bases de datos, se ha calculado también, a efectos de comparación, obviando este campo.

La figura muestra que es en ERIC donde se asigna un número significativamente mayor de términos, con un mínimo de 8 descriptores, y en LISA donde menor número de descriptores, con un mínimo de 2 descriptores. Para ser los

	Media ± D.E.	Mínimo	Máximo
Eric (Identificadores + Descriptores		8	17
Eric (Sólo descriptores)	9 ± 2	6	16
Lisa	4 ± 1,4	2	7
Isa	6,4 ± 1,9	3	9

Fig. 1. Número medio de términos asignados por documento

mismos documentos, parece una excesiva diferencia, 5 descriptores más en ERIC que en LISA, y 2,5 más que el ISA, teniendo presente además que en LISA hemos observado la repetición de algunos descriptores en el mismo documento (debido al proceso de indización en cadena), lo que produce una media sesgada en cuanto a descriptores únicos utilizados.

Estas diferencias pueden ser consecuencia bien de las políticas de indización de los productores de estas bases de datos, o bien fallos en los indizadores que no representan todos los conceptos importantes que deberían ser representados. Siendo que LISA e ISA son dos bases idénticas en su cobertura temática, la diferencia de 2,5 descriptores a favor de la primera, puede hacer pensar que se dan las dos circunstancias señaladas.

4.1.3. Exhaustividad y precisión en la indización de los conceptos del título

Se entiende por exhaustividad el grado en el se reconocen y representan todos los temas tratados en un documento, y por precisión la calidad de la indización por parte del indizador. La hemos medido observando si todos los conceptos del título estén reflejados en los descriptores, y además sean términos específicos. Para ello, se ha analizado conceptualmente el titulo de cada artículo al objeto de determinar en qué grado se expresan todos los conceptos del título en los descriptores. En la figura 2 se observa cómo en la base de datos ERIC en 20 ítems (77%), los descriptores reflejan todos los conceptos importantes del título, y en los 6 restantes falta un concepto (23%). En la base de datos ISA, son 21 ítems (80,8%) los que recogen todos los conceptos del título, por sólo 5 en los que falta un concepto. Sin embargo, en la base de datos LISA, solamente 5 registros

	Totalmente		Falta 1 concepto		Falta más de 1 concepto	
	Nº	%	Nº	%	Nº	%
Eric	20	77,0%	6	23,0%		
Lisa	5	19,2%	16	61,5%	5	19,2%
Isa	21	80,8%	5	19,2%		

Fig. 2. Reflejo de los conceptos del título en los descriptores

	Nº de conceptos que coinciden	Nº de términos índice que coinciden
Eric - Lisa	64	39
Eric - Isa	96	81
Lisa - Isa	52	36

Fig. 3. Coincidencia de conceptos y de términos índice en las bases de datos

(19,2%) reflejan todos los conceptos importantes del título, mientras que el 61,5% (16 items) pierde un concepto, y es la única base de datos en la que existen registros (5) con falta de 2 o más conceptos en los descriptores.

Vemos pues cómo existe una diferencia sustancial entre las bases de datos ERIC e ISA, con un porcentaje importante de registros que contienen todos los conceptos importantes del título, por lo que podríamos considerar que están indizados exhaustivamente y con precisión, y la base de datos LISA, muy pobre en este aspecto, con un 80,7% de registros en los que los descriptores reflejan de forma insuficiente el contenido de la publicación. Este aspecto consideramos que es muy importante, ya que si bien el título debe de reflejar con precisión el contenido de la publicación y es responsabilidad del autor o autores, son los productores de las bases de datos quienes tienen la responsabilidad de indizar fielmente y exhaustivamente el tema de que tratan dichas publicaciones, y el título debe de suponer el campo temático en el que deben de poner más atención. De esto se puede deducir que en la Base de datos LISA la exhaustividad de la indización es bastante pobre, y por consiguiente este factor disminuirá la precisión y la llamada en la recuperación de las búsquedas.

Se ha examinado también en qué medida coinciden las bases de datos en los diferentes conceptos expresados por los descriptores de las referencias. Para ello se han revisado todos los registros, comparando por parejas las bases de datos para ver el grado de concordancia. Los resultados se muestran en la Figura 3.

Se puede observar que las dos base de datos que más coinciden en los conceptos indizados son ERIC e ISA, con 96 conceptos, de los que 81 (84,3%) están representados por los mismos términos en las dos bases de datos. El número más bajo corresponde al emparejamiento de LISA e ISA, que coinciden en 52 conceptos, un poco más de la mitad entre ERIC e ISA, utilizando los mismos términos en el 69,2% de los conceptos, porcentaje sensiblemente inferior a la pareja ERIC-ISA.

Estas cifras dan lugar a pensar que ERIC e ISA utilizan un lenguaje índice y unas políticas de indización más parecidas que las que podrían esperarse para dos bases de datos, como LISA e ISA, que son específicas de la Ciencia de la Documentación y de la Información.

4.1.4. Otros aspectos relacionados con la calidad de la información

Otros aspectos relevantes relacionados con la información contenida en las referencias, y que pueden tener consecuencias para una recuperación eficaz en estas bases de datos, son los relacionados con los errores que pueden cometer al registrar los datos en los registros. Estos errores los hemos clasificado en tipográficos o errores por descuido de las personas que se encargan de introducir la información en las bases de datos. De cualquier manera, tienen un efecto negativo en la recuperación, ya que los sistemas de búsqueda buscan la coincidencia exacta de la cadena de caracteres. Comparando la información de los registros en las tres bases de datos, hemos detectado los siguientes errores:

- En la base de datos ISA, aparece el primer apellido de un autor con inicial, y el segundo apellido completo (error por descuido).
- En la base de datos ISA, aparece Cresanti en lugar de Crestani (error tipográfico).
- En la base de datos LISA, aparece un autor como Janes en lugar de James (error tipográfico).
- En la base de datos LISA, aparece un autor como Abdallah, N.N. en lugar de Abdallah, N.B. (error tipográfico).
- En la base de datos LISA aparece el autor como Rijsbergen, C.J.V.; mientras que en LISA y ERIC aparece como Van Rijsbergen, C.J. (error por descuido y/o desconocimiento). Este error aparece en dos registros.
- En la base de datos LISA, aparece el primer apellido de un autor como Cortex en lugar de cómo Cortez (error tipográfico)
- En la base de datos LISA, falta la mitad inicial del título del artículo (error por descuido).
- En la base de datos LISA, falta la mitad final del título del artículo (error por descuido).
- En la base de datos LISA aparecen repetidos dos descriptores en dos registros, pudiendo explicarse este error por el procedimiento de indización en cadena que se utiliza en esta base.
- Para hacer referencia al concepto "segunda forma normal", en LISA aparece como "NF2", en ERIC como "NF squared", y en ISA como "NF sup 2".

Observamos cómo es en la base de datos LISA donde se producen casi todos los errores: 7 en total, 3 tipográficos y 4 por descuido, mientras que en ISA se produce 1 error tipográfico y 1 por descuido.

4.2. Análisis de contenido y temas de investigación

En la Figura 4 se muestran los descriptores utilizados más frecuentemente en las tres bases de datos. Sin tener en cuenta aquellos relaciones con *information retrieval*, está claro que los más utilizados en las tres bases de datos es el que hace referencia a las búsquedas (*search strategies*, *searching*). Otros descriptores utilizados con frecuencia hacen referencia a bases de datos, Internet, relevancia, indización, etc.

Exceptuando los temas de recuperación de información y de búsqueda, con un número de descriptores que sobresalen con respecto a los demás temas, la baja frecuencia con que ocurren los demás descriptores hace pensar la investigación sobre "recuperación de información" no se concentra en unas pocas áreas, sino que son bastantes las líneas de investigación en este sentido, como por ejemplo, búsqueda en bases de datos, indización, representación del conocimiento, métodos de evaluación, Internet y world wide web, inteligencia artificial y sistemas expertos, estudios sobre los usuarios de sistemas de información, interfaces hombre-máquina, redes neuronales y de información, rendimiento de sistemas de recuperación, análisis de citas, automatización de bibliotecas, lenguaje natural, procesos cognitivos, etc.

Eric	Lisa	Isa
Information retrieval (26)	Searching (13)	Information retrieval (18)
Search strategies (9)	Online information retrieval (9)	Searching (7)
Databases (6)	Hypertext (3)	Information retrieval systems (7)
Indexing (5)	Automatic text analysis (2)	Databases (3)
Information processing (5)	Bibliographic databases (2)	Hypertext (3)
Users (information) (5)	Online databases (2)	Indexing (3)
Relevance (5)	Choice of terms (2)	Text retrieval (3)
Semantics (5)	Machine learning (2)	Artificial Intelligence (2)
Computer system design (4)	Statistical techniques (2)	Boolean functions (2)
Knowledge representation (4)	Neural networks (2)	Computer networks (2)
Evaluation methods (4)	Research (2)	Design (2)
Models (4)	Relevance feedback (2)	Internet (2)

Fig. 4. Descriptores más utilizados en las bases de datos

Sí que nos ha sorprendido, por ejemplo, que mientras que en ERIC y en ISA aparecen varios documentos indizados bajo los términos *indexing* o *index terms*, en LISA no aparece ningún documento indizado de tal manera, lo que induce a pensar en la falta de especificidad del vocabulario empleado por esta base de datos, o bien en errores de los indizadores que no representan todos los conceptos del documento o eligen términos no apropiados.

El número de términos índice diferentes utilizados en ERIC es de 162, en LISA de 71 y en ISA, 112. Si dividimos estas cifras por el número total de descriptores que aparecen en todas las referencias en cada base de datos, nos da un resultado que puede indicar la proporción de descriptores nuevos que pueden aparecer en cada registro: en ERIC la proporción es de 0,56 (162/288), en LISA de 0,68 (71/194) y en ISA de 0,68 (112/165). Observamos cómo la proporción es igual en LISA e ISA, que son superiores a la encontrada en ERIC.

Esta medida esta en relación directa con la profundidad de indización, ya que a mayor número de descriptores utilizados en la indización de los documentos, menor será la probabilidad de que aparezcan descriptores nuevos. Así, estos resultados están en consonancia con los obtenidos en la figura 1, donde la media de descriptores era bastante mayor en ERIC que en las otras bases de datos.

Al objeto de comparar la similitud del lenguaje de indización utilizado por cada una de las bases de datos, se han examinado aquellos descriptores que aparecen exactamente igual (fig. 5).

Observamos que los lenguajes de indización con mayor número de términos iguales son los utilizados por las bases de datos ERIC y por ISA. Es este un aspecto importante, sobre todo para el intermediario o el usuario que vaya a realizar la búsqueda. De sobras es conocido que cuando se interroga en varias bases de datos a la vez, la diferencia del vocabulario empleado por cada una de ellas puede ocasionar, si no se tienen en cuanta las diferentes formas en que pueden aparecer los términos, una pérdida de información relevante. En la Figura 6 podemos ver qué términos existen idénticos en los lenguajes de estas bases de datos, lo que nos puede ayudar a la hora de diseñar la estrategia de búsqueda y seleccionar los términos adecuados.

5. Conclusiones

Como conclusiones, hemos encontrado algunas diferencias importantes en las tres bases de datos, destacando lo siguiente:

- ERIC es la base de datos que indiza de forma más exhaustiva, lo que favorecerá una alta llamada en la recuperación de información. Utiliza un vocabulario controlado específico, y la indización es precisa. No hemos detectado errores tipográficos o por descuido de los indizadores.

ERIC-LISA	ERIC-ISA	LISA-ISA
Bibliographic databases	Abstracts	Citations
Citations	Artificial intelligence	Databases
Databases	Automation	Eastern Europe
Evaluation	Biomedicine	End users
Feedback	Citations	Evaluation
Hypermedia	Cluster analysis	Feedback
Information retrieval	Computer networks	Generic algorithms
Information science	Cybernetics	Hypertext
Information theory	Databases	Information retrieval
Internet	Evaluation	Information theory
Library of Congress	Feedback	Internet
Mathematical models	Indexes	Library of Congress
Natural language processing	Indexing	Machine learning
Neural networks	Information infrastructure	Networks
Online catalogs	Information networks	Neural networks
Research	Information retrieval	Online catalogs
Search strategies	Information theory	Relevance feedback
Situated action	Internet	Research
Syracuse university	Library automation	Searching
	Library of Congress	Syntatic analysis
	Logic	
	Models	
	Natural language	
	Navigation	
	Neural networks	
	Online catalogs	
	Performance	
	Query processing	
	Relevance	
	Research	
	Semantics	
	User studies	
	World wide web	

Fig. 5. Descriptores idénticos en las bases de datos

- LISA es la base de datos que perores resultados ha ofrecido en comparación con las otras dos. Es la que menos términos índice asigna por documento, lo que puede indicar que no se representan todos los conceptos importantes del documento, lo que implicará una repercusión negativa en la llamada. El procedimiento de indización en cadena que utiliza tiene como consecuencia la aparición de descriptores repetidos. Posee un vocabulario controlado poco específico, y además se han observado fallos tanto en la indización como de tipo tipográfico y por descuido, lo que influirá negativamente en la llamada y en la precisión.
- ISA es una base de datos que dispone de un vocabulario controlado, y la indización es específica y consistente, lo que favorece la precisión y exhaustividad tanto en la búsqueda como en la recuperación. Sólo se han observado dos errores en la introducción de información en la base de datos.

Con respecto a las líneas de investigación sobre *Recuperación de Información*, aunque la muestra de registros es pequeña, se intuye que se está trabajando en una gran variedad de temas y aspectos, destacando sobre todos el que hace referencia a la búsqueda de información.

En resumen, se desprende que ERIC e ISA son las dos bases de datos más afines, tanto por el lenguaje que utilizan como por la política de indización y seriedad de los indizadores. Sin embargo, LISA da una imagen bastante peor, y debería de cambiar su política de indización, tanto en lo que se refiere a especificidad como exhaustividad. Sería deseable que las dos bases de datos más importantes que cubren literatura mundial en el ámbito de las Ciencias de la Documentación e Información, (y de las pocas existen), normalizaran y se parecieran un poco más en sus políticas de indización y en el vocabulario utilizado.

6. Bibliografía

Ellis, D. (1996). Progress and problems in information retrieval. 2ªed. London : Library Association Publishing, 1996.

Hood, W. ; Wilson, C.S. (1994). Indexing terms in LISA database on CD-ROM. Information Processing and Management, 30 : 3 (1994) 327-342.

Lancaster, F.W. (1986). Vocabulary control for information retrieval. 2ª ed. Washington, D.C. : Information Resources Press, 1986.

Lancaster, F.W. ; Warner, A.J. (1993). Information retrieval today. Washington, D.C. : Information Resources Press, 1993.