Evaluación de calidad de catálogos mediante el uso tests de suciedad

Mª del Pilar Ortego de Lorenzo-Cáceres
José Luis Bonal Zazo
Universidad de Extremadura. Facultad de Biblioteconomía y Documentación
C.e. portego@alcazaba.unex.es

Resumen

Se analizan los tests de suciedad como método de evaluación de catálogos automatizados, con los siguientes objetivos: exponer las características de este sistema de evaluación, proponer unas pautas de aplicación y analizar las diferencias entre los distintos tipos de test de suciedad, con el fin de establecer cuál es el sistema más adecuado para medir la calidad de las bases de datos. El trabajo se completa con un ejemplo de aplicación.

Palabras clave: *Evaluación de bases de datos bibliográficas/test de suciedad.*

Introducción

La tradicional división de los errores catalográficos en dos tipos[1]: errores de consistencia (errores en la aplicación de las normas) y errores de precisión (errores de caracteres), exige la utilización de métodos de análisis y evaluación diferentes para conocer el nivel de calidad de una base de datos. Uno de los procedimientos empleados para la evaluación de la presencia de errores de precisión es el uso de tests de suciedad. Los tests de suciedad se basan en la búsqueda deliberada de una secuencia de palabras erróneas representativas en una base de datos, de este modo, se buscará una término mal escrito, como por ejemplo «*Andaluía*» en lugar de «*Andalucía*», con el fin de conocer la frecuencia de aparición de la forma errónea, y a partir de

[1] CHAPMAN, Ann. «Up to standard? A study of the quality of records in a shared cataloguing database». *Journal of Librarianship and Information Science.* 1994, vol. 26, nº 4, pp. 201-210.

aquí intentar determinar el grado de «suciedad» de una base de datos. Paradójicamente, de este modo, los términos erróneos, se convierten en elementos de ayuda para la evaluación.

Aunque los primeros trabajos sobre el estudio de los errores mecanográficos y ortográficos en los catálogos se remontan a finales de los años 70[2], ha sido durante la década de los noventa cuando se ha producido el desarrollo de los tests de suciedad como método de evaluación de catálogos. En 1991 Jeffrey Beall planteó la posibilidad de utilizar una relación de 10 palabras mal escritas en diferentes bases de datos documentales con la finalidad de probar su calidad, Beall propuso la utilización de las palabras: *febuary, Guatamala, misssion, goverment, Fransisco, grammer, recieve, wensday, seperate* y *conditons.* Aunque su aplicación inicial sobre el catálogo de la Biblioteca del Congreso resultó fallido, su utilización posterior para evaluar las bases de datos de DIALOG demostró la utilidad del sistema[3].

El test de Beall fue perfeccionado por Jim Dwyer. Dwyer parte del mismo procedimiento de Beall, pero con una diferencia significativa: propone determinar además del número de palabras erróneas, el número de palabras correctas existentes por cada palabra errónea, con el objetivo de establecer en términos relativos el nº de palabras erróneas/nº de palabras escritas correctamente[4].

En 1994, a partir de las experiencias de Beall y de Dwyer, Pamela Cahn modificó el procedimiento de búsqueda, truncando los términos erróneos con la finalidad de aumentar la búsqueda, y de este modo detectar el mayor número posible de palabras mal escritas. El método exigía, sin embargo, un análisis exhaustivo de los términos erróneos en su contexto, con el objeto de determinar en qué medida el truncamiento

---

[2] BOURNE, Charles P. «Frequency and impact of spelling errors in bibliographic Data Bases». *Information processing & management.* 1977, 13, nº 1, pp. 1-12.

[3] Cit. por: CAHN, Pamela. «Testing database quality». *Database.* 1994, febrero, p. 24

[4] DWYER, Jim. «The cataloger's «invisible college» at work: the case of the dirty database test. *Cataloging & Classification Quarterly.* 1991, vol. 14, nº 1, p. 75-82.

había ayudado a recuperar el término mal escrito o si, por el contrario, eran términos correctos, dado que el uso del truncamiento podía ayudar a recuperar palabras correctas[5].

La misma autora, sin embargo, propuso una variante de su método: la búsqueda de todas las formas posibles de errores que podían aparecer asociados a una palabra mediante el operador «o» (por ejemplo, los errores asociados a la palabra «*management*» son «*managememement*», «*managemeht*»...). Cahn mejoró el análisis de los resultados, mediante la utilización de varias medidas de valoración complementarias: 1) el campo de localización del error, 2) el carácter del error, es decir, si era la única vez que aparecía esa palabra en el registro o si, por el contrario, la palabra aparecía varias veces, y alguna de ellas de forma correcta, de tal manera que la recuperación estuviera garantizada[6]. La autora propuso también el análisis del tamaño de la base de datos, con el fin de comprobar si existía algún tipo de relación entre el número de errores y el tamaño de la base de datos[7]. El modelo de Pamela Cahn ha sido utilizado posteriormente por autoras como Barbara Nichols Randall, quien ha aplicado el mismo método, adaptándolo al análisis comparativo de varios catálogos de bibliotecas universitarias[8].

Metodología

Siguiendo los modelos propuestos por Beall, Dwyer y Cahn el trabajo se ha desarrollado en las siguientes fases:

1. Selección de los términos erróneos necesarios para ser usados en el test. Siguiendo las pautas de Dwyer, se han seleccionado cuatro tipos de términos: nombres propios de lugares, nombres propios de personas, nombres de términos relacionados con la descripción bibliográfica y otros nombres comunes, todos ellos caracterizados por ser palabras con riesgo potencial de permutación y supresión de caracteres. Los térmi-

---

5 CAHN, Pamela. «Testing database quality»..., loc. cit., p. 23-30

6 Ibídem, p. 26-27

7 Ibídem

8 NICHOLS RANDALL, Barbara. «Spelling errors in the Database: shadow or substance». *Library resources & Technical Services*. 1999, vol. 43, nº 3, p. 161-169.

nos elegidos han sido *Barcelon, Ministeriode, Rodrigez, coleción, medician, dicionario, universida,* y *commemoración.* Para la selección, se ha optado por buscar los términos correctos en la base de datos de Bibliografía Española[9] y recuperar, a través del índice, todos aquellos términos érroneos relacionados, seleccionando, entre éstos, aquél que apareciera un mayor número de veces; por ejemplo, para seleccionar el término «*coleción*», se buscó el término correcto, «*colección*» y a partir de éste fueron buscados todos los términos erróneos asociados (*coleccción* (6), *coleccción* (1), *colecciión* (1), *colecciogn* (1), *colecciám* (2), *coleccioón* (1), *colecciuon* (1), *coleccom* (6), *colección* (2), *coleción* (27), *calección* (1), *colición* (1)). De las doce formas posibles recuperadas se seleccionó coleción, por ser aquélla que aparecía con más frecuencia.

2. Búsqueda del número de términos erróneos. Tras la selección de los términos se ha realizado su búsqueda en los catálogos de bibliotecas públicas de las comunidades Autónomas de Andalucía, Castilla y León y Aragón, así como en la propia Bibliografía Nacional Española, con la finalidad de obtener los valores reales y porcentuales apropiados. La búsqueda se ha completado con un análisis del contexto en que se encontraba cada término, dado que algunos de los términos seleccionados podían ser correctos, por ejemplo: «*dicionario*» (término aceptado en gallego y portugués) o «*Barcelon*» (término existente como apellido). Además del contexto de cada término se ha analizado también el área de la descripción bibliográfica en que se encontraba, con el objeto de intentar determinar las áreas con mayor riesgo de error.

3. Unificación de datos. A fin de comparar los datos resultantes del análisis de los distintos catálogos ha sido necesario reducirlos a unos valores relativos comparables entre sí, aunque también se proporcionan los valores absolutos. Siguiendo el modelo de Dwyer se ha obtenido la tasa: nº de palabras erróneas/nº de palabras similares bien escritas, por mil.

---

[9] BIBLIOTECA NACIONAL (España). *Bibliografía Nacional Española.* [Cd rom]. Madrid: Biblioteca Nacional, 1998. Disco 22, junio de 1998.

4. Análisis de resultados. El análisis de los resultados se ha centrado en cuatro aspectos: 1) el estudio comparativo de las tasas de error de los catálogos de las tres Bibliotecas públicas analizadas; 2) el análisis comparado entre la media de errores existentes en los catálogos de las Bibliotecas públicas y la Bibliografía española; 3) el análisis de las áreas de la descripción bibliográfica en las que han sido detectados los errores; y 4) el estudio comparado entre el test de Dwyer (basado en un solo término erróneo) y el test de Pamela Cahn (basado en el estudio de todos los términos erróneos variantes de una palabra) sobre la misma base de datos, la de Bibliografía española.

Resultados

Por lo que respecta al porcentaje de errores por catálogos de bibliotecas públicas, de las tres comunidades autónomas analizadas es la comunidad aragonesa la que parece tener un porcentaje superior respecto al resto (53,05 por mil, frente al 49,2 por mil de Castilla y León y al 29,5 por mil de Andalucía). Sin embargo, hay que advertir que existe un elemento distorsionador, dado que entre los términos del test se encuentra una palabra con un índice de frecuencia muy bajo, pero con un alto índice de error (*commemoración*). La utilización de este término altera los resultados hasta tal punto que suprimiendo la palabra, los resultados generales se reducen notablemente (Andalucía: 4,52 por mil; Aragón: 3,05 por mil; Castilla y León: 2,31 por mil).

En cuanto a la relación entre el número de errores en los catálogos de las bibliotecas públicas y los errores de la Bibliografía Nacional Española, aunque existe una diferencia significativa entre ambos (43,95 por mil en las Bibliotecas públicas; y 49,8 por mil en la Bibliografía Nacional), ésta se encuentra directamente relacionada con el número de palabras correctas existentes en ambas bases de datos, mucho mayor en el caso de la Bibliografía Nacional Española (320.346 formas bien escritas de las 8 palabras analizadas en Bibliografía Española, frente a 193.836 palabras correctas en los catálogos de las bibliotecas públicas).

Términos erróneos	Catálogos de Bibliotecas públicas								B. Española	
	Andalucía		*Aragón*		*Castilla-León*		*Media*			
	Absoluto	X 1000	Absoluto	X1000	Absoluto	X 1000	Absoluta	Relativa	Absoluto	X 1000
Barcelon	18	0,225	32	0,54	66	0,22	38,6	0,3283	23	0,117
Ministeriode	5	1,47	0	0	1	0,075	2	0,515	15	1,28
Rodrígez	1	0,20	1	0,40	9	0,57	3,6	0,39	4	0,33
Colección	13	1,43	6	0,76	24	0,81	14,3	1	26	0,46
Medician	0	0	0	0	1	0,20	0,33	0,066	1	0,189
Dicionario	3	1,02	2	1,15	1	0,119	2	0,763	4	0,877
Universida	2	0,23	1	0,20	7	0,319	3,33	0,749	9	0,265
Commemoración	1	25	2	50	7	46,97	3,33	40,656	5	46,29
Total	43	**29,52**	44	**53,05**	116	**49,283**	64,79	**43,95**	87	**49,808**

Tabla 1. Datos absolutos y relativos

El análisis de la distribución de errores por las áreas de la descripción proporciona resultados significativos: las áreas con mayor número de errores son, de mayor a menor: el área de publicación (54,48% errores), el área de serie (22,41% errores), el área de título y mención de responsabilidad (15,86% errores) y el área de notas (4,82% errores). La existencia de un número tan elevado de errores en el área de publicación se debe a la ausencia de mecanismos de control de entrada de nombres geográficos y de editoriales de forma normalizada. El área de serie, en cambio, pese a ser uno de los puntos de acceso sometido a control de autoridades en el registro de encabezamientos secundarios, tanto en las Bibliotecas públicas como en la Biliografía Nacional Española, presenta un porcentaje excesivamente elevado de errores, debido a la falta de control en el cuerpo de la descripción bibliográfica, superior incluso a otras áreas con mayor riesgo de error, tales como la de título y mención de responsabilidad y la de notas.

Frente a las áreas de la descripción con mayor número de errores, las áreas y elementos con menor número de errores son las de edición, descripción física, encabezamiento principal y encabezamientos secundarios, las dos primeras por ser aquéllas que cuentan con menor número de palabras y los dos últimos, los puntos de acceso, por ser los elementos sujetos a mayor control normativo y técnico. De este modo, se garantiza el acceso a los registros por los puntos de acceso tradicionales.

Área	Andalucía	Aragón	Castilla-León	B. Española	*Total*
Punto de acceso principal			3		3
Título y mención de responsabilidad	3	3	11	29	46
Edición	1				1
Publicación	26	33	68	31	158
Descripción física	1				1
Serie	11	5	25	24	65
Notas	1	2	8	3	14
ISBN					0
Encabezamientos secundarios		1			1
Fuente de publicación			1		1
Total	43	44	116	87	

Tabla 2. Distribución de errores por áreas

Pese a que el test de Dwyer es significativo, dado que proporciona datos uniformes sobre un término erróneo en las distintas bases de datos analizadas, la utilización del método B) de Pamela Cahn presenta algunas ventajas. En primer lugar, mejora notablemente la representatividad de los errores al disminuir la distorsión que puede provocar un número muy elevado de errores en una palabra determinada (tal como ocurre con la palabra *conmemoración*) (v. tabla 3). Además aumenta la representatividad de los posibles tipos de errores que puede experimentar una palabra (permutación, supresión de espacios, omisión, sustitución, repetición de letras, inserción de letras, e inserción de espacios en blanco): mientras que con el test de Dwyer simplemente se representa un tipo de error (omisión -*Barcelon* (23 casos), *Rodrigue* (4 casos), *colección* (27 casos), *dicionario* (23 casos), *universida* (17 casos)-, supresión de espacio -*Ministeriode* (15 casos)-, permutación -*medician* (1 caso)-, y sustitución -*commemoración* (5 casos)-), con el test B de Pamela Cahn aumenta la representación de los tipos de errores para las mismas palabras (permutación -28 casos-, omisión -191 casos-, sustitución -130 casos-, repetición de letras -23 casos-, inserción de letras -38 casos-, y supresión de espacios en blanco -15 casos-).

Palabra seleccionada	Nº total de palabras bien escritas	Valores sobre una palabra errónea (Test de Dwyer)			Valores sobre todas las palabras erróneas (Test B de Pamela Cahn)		
		Palabra errónea seleccionada -Frecuencia	Nº de errores por cada 1000 palabras correctas	Porcentaje sobre el nº total de errores por mil	Nº total de palabras erróneas	Nº total de palabras erróneas por cada mil palabras correctas	Porcentaje sobre el nº total de palabras erróneas por mil
Barcelona	196.517	*Barrelon* 23	0,117	**0,23%**	127	0,646	**0,75%**
Ministerio	11.647	*Ministerio de* 15	1,287	**2,53%**	34	2,919	**3,41%**
Rodríguez	11.999	*Rodríguez* 4	0,333	**0,65%**	71	5,917	**6,91%**
Colección	56.341	*Colerción* 26	0,461	**0,907%**	50	0,887	**1,36%**
Medicina	5280	Medician 1	1,189	**2,33%**	9	1,704	**1,99%**
Diccionario	4.557	Diccionario 4	0,877	**1,72%**	27	5,924	**6,921%**
Universidad	33.897	Universida 9	0,265	**0,52%**	94	2,773	**3,24%**
Conmemoración	108	Commemoración 5	46,29	**91,08%**	7	64,814	**75,73%**
Total			50,82			85,584	

Tabla 3. Datos comparativos de los tests Dwyer/Cahn aplicados a la misma base de datos: Bibliografía Española en CD-ROM

Conclusión

Aunque los tres métodos descritos son apropiados para la evaluación de la calidad de los catálogos, es el de Cahn el más completo, debido a que parte de la experiencia de los otros dos procedimientos, completándolos, y proporcionando datos más amplios, más correctos, y representativos de los diferentes tipos de errores posibles. No obstante, se trata de un procedimiento apropiado, exclusivamente, para el análisis y evaluación de los errores de precisión.

En cualquier caso, pese a las diferencias existentes entre los distintos métodos de test de suciedad, las fases de aplicación son comunes a todos ellos, independientemente de que se profundice más o menos en su estudio: selección de los términos erróneos, búsqueda de los mismos, unificación de datos y análisis de éstos.