Un algoritmo informétrico para la evaluación de un vocabulario de búsqueda

Julián Colina (SEDIC)
Apdo. 47-126 28080 Madrid

0.1. Resumen

Evaluación informétrica de un vocabulario de búsqueda: Primero, se construye un vocabulario de términos, después se le añaden los sinónimos, lo cual crear un fichero único de los principales descriptores y los mismos descriptores con sus sinónimos. Entonces, para simular un modelo, usa el algoritmo NAT=[LOGe{ relaciones dentro y entre términos }], que es la cantidad de información, NAT, que es evaluada por el logaritmo natural: 3 es el óptimo, 5 ó 4 indican incertidumbre (ruido) ,2 ó 1 relaciones sin identificar (silencio); 3 también es la media aritmética del polinomio para 1 a 149 iteraciones, y asimismo *la capacidad potencial* de transferencia de información documental, por repeticiones de los grupos de sinonimias en los términos. Señalando pues la existencia de los términos (y lexemas) y sus relaciones dentro/entre. Aquí se han usado 10 estrategias, para 2 términos de búsqueda equiprobables o igualmente disponibles, para lo que podría haberse comunicado. Y facilita un análisis lexicográfico y semántico del vocabulario. Encuentra sinonimias alternativas y descriptores de términos, con signaturas válidas. También es válido el uso de este método para comprobar la consistencia de los operandos, es decir, si están completas y bien establecidas las relaciones. Mide pues las partículas elementales de información del thesaurus como una estructura organizada.

Palabras clave: Recuperación de la Información. Informetría. Modelos matemáticos. Evaluación.

0.2. Abstract

Informetric evaluation of a retrieval vocabulary in order ot find a methodology to alert the user of noise and silences. This is to be done before retrieving information, ensuring in this way better retrieval results. First, a vocabulary of terms is built. Thereafter, synonyms are added, merging in a unique file descriptors and equivalent terms. Then, an information retrieval model is achived using the logarithm NAT=[LOGe{ relations into and among terms }]. That is, the

quantity of infomation, NAT, evaluated by means of the natural algorithm: 3 means the optimal, 5 or 4 mean uncertainty (noise), 2 and 1 not identified relations (silence).

Keywords: Information retrieval. Informetrics. Mathematical models. Evaluation.

1. Introducción

Se examina como las propiedades informétricas de las bases de datos pueden ser usadas como herramientas para ayudar a un diseño más eficiente y más efectivo de los sistemas de recuperación de información. Se parte de un modelo general de un sistema de información, en el que se use un vocabulario de recuperación. Se trata de encontrar un método de evaluación de ese vocabulario de búsqueda, que nos advierta de los ruidos y silencios, antes de proceder al proceso de la recuperación de la información, haciendo así ésta más racional y más rica.

Para simular el modelo han de construirse los componentes que se consideren críticos o sensibles a los objetivos o propósitos finales y que intervienen en la medición y en la recuperación de la información y son: el vocabulario, unas reglas de simulación y finalmente un algoritmo que evalúe cuantitativamente las estrategias de búsqueda. Además hay que disponer de una base de datos sobre la cual ensayar todo ello. Finalmente, y para comprobar la adecuación de nuestras hipótesis de trabajo con los resultados esperados, hay que hacer ensayos reiterativos, lo cual se consigue simplemente simulando el comportamiento del modelo, —que ha de ser construido— y de su estructura. Se ha escrito para todo ello un programa informatizado, como herramienta de trabajo que consideramos la más adecuada y que incorpora esta novedad de una prevaloración de términos de búsqueda, que expresadas las opciones de búsqueda en lenguaje natural, aplique unas reglas expertas y valore —matemáticamente— la calidad de la búsqueda y remita a los descriptores, que el catalogador realmente empleó y con ellos encontrar la información solicitada. Se ensayó todo ello partiendo de relaciones jerárquicas de materias, siguiendo con sus sinonimias, continuando con sus equivalencias y asociaciones y también lexemas o significantes. Se utilizó un vocabulario de acorde con la base de datos disponible: Información y Documentación, con su Informática asociada, como puede verse en los ejemplos de evaluación de una búsqueda, que se acompañan como ensayos.

2. Construcción del modelo general

Un modelo general debería ser prioritario y éste necesariamente contendría las funciones clásicas de una biblioteca-hemeroteca, porque a este tipo de centros

de documentación nos referiremos, y que son en principio:

• Gestión de Compra de libros y Suscripciones y Recepción de ejemplares. (Adquisiciones).

• Gestión de Catalogación y Clasificación, Resúmenes y Vocabularios (Catalogación).

• Consultas por N' Registro, Títulos, Autores, Signaturas, Materias y Descriptores, y, eventualmente, Préstamos (Circulación).

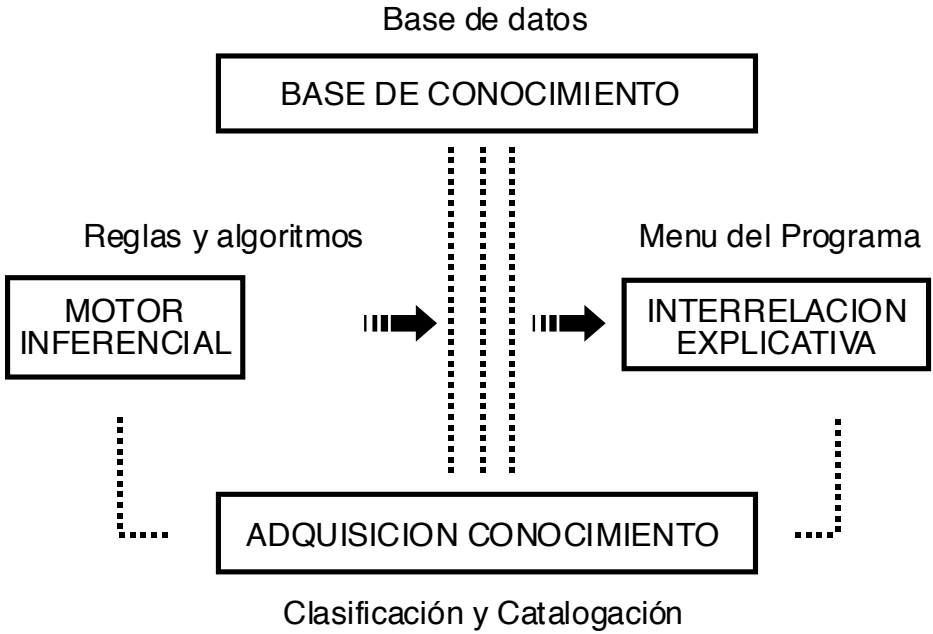

Fig. 1. Diagrama de un sistema experto y estructuras de relación

El diagrama abstracto de las operaciones de un sistema o un programa, como sistema experto, se concreta desarrollando la estructura con sus componentes precisos: fases, campos de datos y sus relaciones con la recuperación de la información. Ha de incluir, necesariamente, aquellos elementos que intervienen en su medición: vocabularios, reglas y algoritmos, que ahora son el objeto de nuestro concreto estudio:

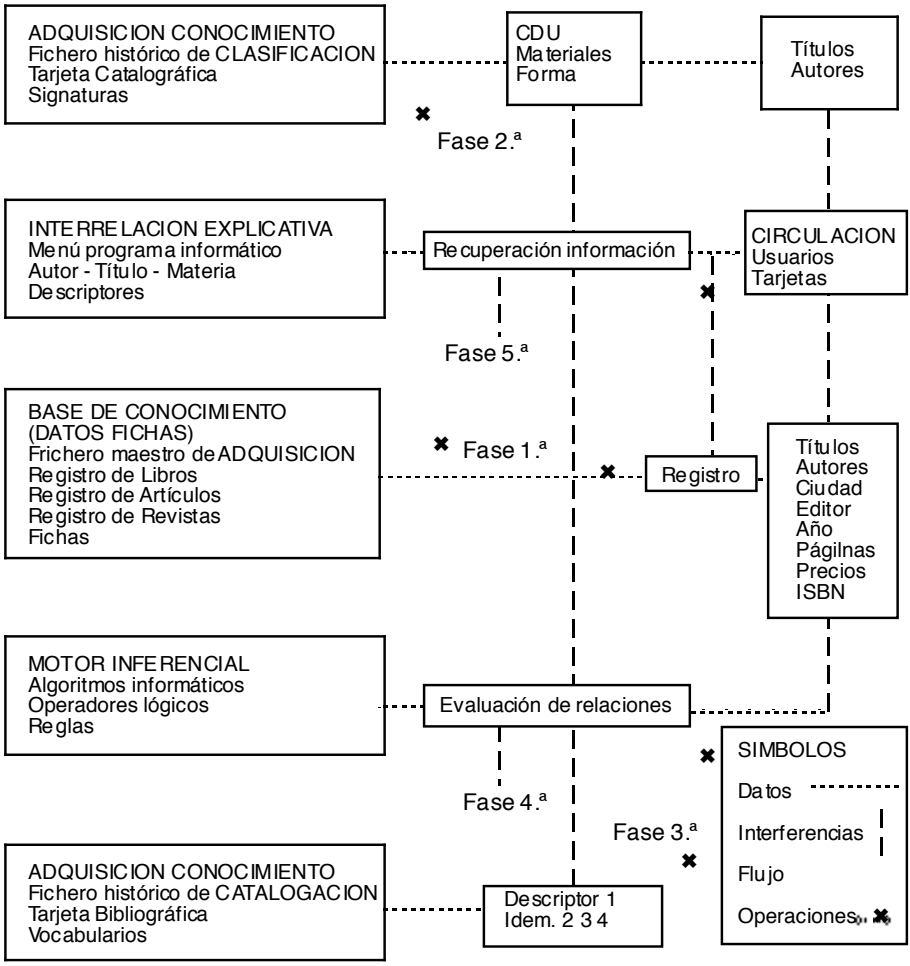

Fig. 2. Estructura del modelo como un sistema de información

3. Construcción de un vocabulario

Después de haber diseñado el Modelo General, para tener completos todos los requisitos, es necesario, en primer lugar, construir un vocabulario de trabajo para los temas específicos de nuestro interés. Podemos usar la misma CDU como punto de partida, añadiéndole después sus términos asociados y equivalentes. Al final habremos obtenido un fichero único de los principales descriptores (encabezamientos de materias) y de los mismos descriptores o palabras clave, con sus sinónimos.

Descriptor	**Términos equivalentes, relacionados, genéricos y específicos ("sinónimos") junto con el descriptor**
Bases	Bancos bases datos terminales Bancos datos bases datos información bibliotecas * bases Bases datos analista programador jefe senior junior Bases datos bancos ficheros gestión proceso relacional Bases datos bibliográficos gráficos documentales Bases datos dbase foxbase relacionales Bases datos distribuidos relacionales lenguaje consulta Dbase bases datos ficheros representación Diseño gestión bases datos Indización indexación ficheros bases Términos vocabularios bases ficheros terminología
Bibliotecas	Automatización informatizadas bibliotecas hemerotecas Bancos datos bases datos información bibliotecas * bibliotecas Bibliotecas bibliotecario documentalista libros Bibliotecas programas programación soporte lógico Biblioteconomía bibliometría bibliotecas documentación Hemeroteca revistas documentación bibliotecas Programas biblioteconómicos bibliotecas informatizadas
Información	Bancos datos bases datos información bibliotecas Centros información fuentes documentales fondos Documentación información informetría Formación información educación usuarios * información Informática básica teoría información científica historia Tecnología programación transferencia información Teoría información codificación pseudocodigos

Fig. 3. Estructura del fiechero de descriptores y términos "sinónimos"

Así es como hemos hecho una indización conceptual de nuestro vocabulario de trabajo. En los ejemplos que siguen hemos ido añadiendo a unas materias —*bases, bibliotecas* e *información*—, todos los otros "sinónimos" —en un sentido amplio—.

4. Reglas de simulación

Una vez que tenemos el vocabulario, vemos que hay términos que están en la misma línea del listado o relacionados con otros sinónimos o equivalente o simplemente asociados. Esto es así porque fue la forma como lo construimos.

Hay también parejas de términos que se encuentran al menos una vez en la misma línea, porque su sinonimia es total. Al primer suceso le llamamos -relaciones *dentro*- y al segundo suceso le denominamos —relaciones *entre*—. Con todas las posibilidades construimos la llamada Tabla de Verdad —la base de las Reglas—, donde hay un óptimo, posibilidades 1 y 8, de máxima ocurrencia.

En la simulación de la búsqueda vamos a comprobar por el ordenador estos casos y por ello se expresan sólo dos palabras (una pareja) en lenguaje natural o dos términos, A y B, para hacer más simples las pruebas. Sus posibles relaciones con sus "sinónimos" —en un sentido lato—, *dentro*, y *entre* los dos términos A y B, son así :

Relaciones dentro del termino	**Relaciones entre los términos**
Relaciones si no si si no no si	si no no
Término A si si si no si no no	si si no
Término B si si no si no si no	si si no

Fig. 4. Tabla de verdad para la simulación y la evaluación del vocabulario

Posibilidades: 1 2 3 4 5 6 7 8 9 *

Las Reglas para todas las posibilidades son éstas :

1 Hay términos y relaciones dentro de los términos [usar algún sinónimo].

2569 No hay relaciones o no están aun en el vocabulario [establecerlas].

3456 Alguno de los términos no están aun en el vocabulario [establecerlos].

7 I m p o s i b l e

8 Hay términos y relaciones entre ellos [la búsqueda es completa].

9 Hay términos pero no relaciones entre los mismos [termino único].

* N u l a : no hay términos ni relaciones [usar otras palabras clave].

5. Construcción del algoritmo

Ahora vamos tras un óptimo de la "cantidad" de información. Una fórmula algebraica que actúe como una "ecuación de sensibilidad", que nos diga cual es el óptimo en nuestras evaluaciones. Que, al *simular* los parámetros para cada variable, veamos como se propaga el efecto sobre la trayectoria S —sucesos—, de un elemento del vector NAT —cantidad de decisión-información—. Ver en una representación gráfica situaciones posibles y situaciones más críticas, para validar su utilidad predictiva. Sobre las propiedades de la representación gráfica volveremos más adelante. La dinámica del modelo, que de por sí no tiene sensibilidad, depende de la precisión de las variables: qué parámetros tienen mayor influencia en el comportamiento del modelo. La estructura de la realimentación de un sistema produce su comportamiento dinámico, esta realimentación se refiere tanto a los descriptores que añadamos como al uso que de ellos hacemos; en este proceso de medida seguimos refiriéndonos a los descriptores como variables y no a la Base de Datos —Base del Conocimiento— y ni tan siquiera a la Adquisición del Conocimiento. En el Modelo estamos en la Fase del Motor Inferencial, con un objetivo bien diferenciado : responder a una pregunta objeto de una simulación o a un aprendizaje o a una búsqueda de información.

La fórmula sería una expresión algebraica que cuente ó las veces que algunos de los términos o sus lexemas están en algunas de las líneas, las veces que uno de los términos figure como palabra aislada y cuando ambos términos están juntos en una misma línea. A todos estos sucesos les hemos dado el mismo peso para medirlos, y se puede comprobar que el cuanto del índice es superior a contar el número de líneas, pues los términos se repiten en sus lexemas.

Hay amplio margen para que ello no produzca distorsión, como puede verse más adelante en la tabla de frecuencias del cuadro "Tabla de valores de NAT"; pero era necesario para aumentar la captación de la riqueza lingüística del vocabulario.

Siguiendo con la elaboración de la fórmula; buscamos resultados simples y precisamos una conversión de unidades, que exprese mejor el comportamiento esperado para los resultados numéricos, así sabemos por experiencia que no es útil usar muchos descriptores —ni en la catalogación ni en la búsqueda—, porque el ruido o ambigüedad aumenta muy rápidamente o exponencialmente; y buscamos una acotación, una formulación matemática que se comporte de igual manera, que no sea lineal, de cálculo fácil, disponible en calculadoras normales y también en ordenadores, éstos usan además logaritmos naturales, lo cual es una

pista. En una palabra, buscamos la transformación de la serie de números naturales en sus logaritmos, de base e. El rango o recorrido, como dispersión de la distribución esperada, queda así sólo entre valores 1 y 5, lo cual es una abstracción, muy útil por otra parte. Así vamos elaborando la fórmula, y ya tenemos que es una expresión algebraica, que es un polinomio o suma algebraica de varios monomios, que es una función logarítmica, que tiene una representación gráfica logarítmica o exponencial, no lineal.

Finalmente, hemos sumado 1 a la fórmula, porque no hay ln de 0 (caso * en las Reglas), que ocasionaría un error en el programa; y contemplamos sólo la parte íntegra, despreciando decimales, porque es así más intuitivo y descriptivo el resultado final, que se mueve en el tramo de valores 1 a 5.

Este planteamiento aritmético es el mismo que mide la cantidad de decisión en la Teoría de la Información, y tomamos el concepto NAT para la formulación:

NAT = parte entera de [LOGe { 1 + relaciones dentro y entre términos }]

Este polinomio es la fórmula para cuantificar las veces que aparecen los términos juntos o relacionados, midiendo nuestras decisiones, y, por ende, la información que contiene, en dichas unidades NAT. Difiere del original de la teoría por algunas correcciones: a) para recibir las particularidades de los lenguajes, como la mentada de los lexemas; b) para su uso informático como sumar 1; c) para poder detectar todas las 10 posibilidades, necesario el caso de los resultados pobres o extremos, que deben ser explicitados, pues reflejan aquellos casos en los que no hay relaciones. Por lo anterior, el sumatorio también contiene veces que el término aparece aislado, único, y sin relaciones, casos 2 y 9, y por último cuando ambos términos están juntos, en la misma línea. Todo ello, lo que se mide, ya se ha definido más arriba, en este mismo capítulo.

Desarrollo de las Fórmulas para la construcción del algoritmo.

(Fuente : AENOR. UNE PNE 71-012)

Ho = log n Función logarítmica de la Cantidad de decisión.

o también

Ho = ln S Cantidad natural de decisión de un mensaje.

(para una base de logaritmo natural o neperiano)

para nuestro tema

NAT = d ln S Sumatorio del logaritmo natural de los sucesos (relaciones)

o también su función exponencial

$S = e^{Ho}$ ó $S = e^{NAT}$

siendo S el número de decisiones elementales distintas para elegir un suceso, y un suceso la presencia de un elemento equiprobable en un conjunto, que hemos enlazado por operadores aritméticos y lógicos.

Sucesos	Logaritmo
2	1
7	2
20	3
54	4
148	5
403	6

Fig. 5. Tabla de valores de NAT. Distribución de los valores de S en función de e

(Son computadorizados en una calculadora con la tecla de función EXP o la inversa de la función exponencial con la tecla de función Ln)

6. Ejemplo y discusión de evaluación de una búsqueda

Un ejemplo consiste en escribir dos palabras relacionadas con la Información y la Informática, puesto que el léxico usado es sobre estos temas o materias, por ejemplo *bases* y *bibliotecas*. El programa puede contar 19 sinonimias, que es una cantidad suficiente para tener un criterio e ir al paso siguiente: seleccionar algunas de ellas que se acerquen más a lo que estamos buscando. Así, por ejemplo, *bibliotecas información*, aquí las relaciones son 13, y la búsqueda es expresada en una valoración, según unas reglas y unas recomendaciones, y el algoritmo hace unas evaluaciones de la cantidad de información; así pues se expresaron el número absoluto de sinonimias y el relativo de relaciones, un pequeño texto que lo describe y un índice informétrico y finalmente el programa de ordenador recupera de sus bancos de datos los items o documentos adecuados. En el programa informático se ha añadido una última opción de seleccionar un término más de libre elección, entre los contenidos en el vocabulario, que oportunamente se

muestra. Por ejemplo el lexema "fox", haciendo referencia a este lenguaje, y dentro de la base de datos de 740 referencias, se obtenían 48 monografías, relativas al concepto de *Gestión de bibliotecas* o temas asociados, como Teoría de la *información.* Además, libros sobre el lenguaje relacional FoxBASE.

Nosotros mismos hemos probado en el CD-ROM de Biblioteconomía y Ciencias de la Información LISA/CRLIS Plus los términos *evaluation retrieval perfomance* al mismo tiempo, que las fichas los incluyan todos, con su comando SU o Subject Descriptor, recuperando 146 referencias. Con otra combinación : *evaluation retrieval informetrics*, sólo tres. Pero con la particularidad de que éstas no estaban entre las anteriores, ya que la lógica que empleamos precisa que todas están en la lista de palabras clave de cada ficha, alcanzando así precisión. Que es muy lógico dado que su base de datos es muy grande : 100.000 citas.