DISPERSION LEXICA Y VOCABULARIO CONTROLADO. EVALUACION DE LOS SISTEMAS DE INDIZACION DOCUMENTAL DESDE LA PERSPECTIVA LEXICOMETRICA

Félix de Moya

Pedro Hípola

E. U. de Biblioteconomía y Documentación

Universidad de Granada

Resumen:

Las aportaciones de la lingüística computacional a los trabajos de la Documentación han producido a lo largo de los años una serie de mejoras en algunos de los sistemas de gestión de bases de datos documentales, como por ejemplo, la incorporación de funciones de ponderación en el proceso de indización automática de colecciones de documentos en texto íntegro.

En esta comunicación se analiza la posibilidad, las ventajas y limitaciones propias de aplicar algunas de esas técnicas a subconjuntos de términos de indización pertenecientes a léxicos de lenguaje documental con vocabulario controlado. Se aborda en especial el caso de una posible aplicación de funciones de ponderación a esos materiales.

Tema: Procesamiento de Lenguaje Natural y Gestión documental

DISPERSION LEXICA Y VOCABULARIO CONTROLADO. EVALUACION DE LOS SISTEMAS DE INDIZACION DOCUMENTAL DESDE LA PERSPECTIVA LEXICOMETRICA

Félix de Moya

Pedro Hípola

E. U. de Biblioteconomía y Documentación

Universidad de Granada

La mayor parte de los estudios y experiencias que, encuadrados dentro del entorno del Procesamiento del Lenguaje Natural, se ocupan de problemas de la **Documentación** suelen centrarse en una de estas dos áreas:

- cuestiones relacionadas con la posibilidad de utilizar aportaciones y técnicas de la **lingüística computacional** para los procesos de creación automática de índices documentales a partir de **textos íntegros** en lenguaje natural;
- intentos de uso de esas mismas técnicas para procesar las **búsquedas documentales** a partir de las consultas formuladas por el usuario final, que se expresa en términos de lenguaje natural;

En nuestro caso queremos referir ahora algunas experiencias asimilables a estas dos áreas de trabajo en cuanto a los procedimientos, pero diversas en lo que se refiere al tipo de materiales a los que se aplican las técnicas. Como es sabido, la mayoría de los investigadores y desarrolladores que aplican técnicas de PLN a la gestión documental han trabajado sobre todo a partir de materiales lingüísticos obtenidos en estado «bruto»: textos escritos completos en el primero de los casos, discurso hablado en el segundo.

El marco de aplicación en el que se sitúa este trabajo es el de las bases de datos documentales que contienen **información referencial bibliográfica**. Con independencia de que tales ficheros incluyan o no **textos íntegros** de los propios documentos originales, el modelo típico de registro referencial incluye uno o más campos dentro de los que se almacena una información condensada en una pequeña lista de palabras o expresiones que forman parte de lo que se suele denominar **lenguaje documental**[1]. En estas páginas se analizan específicamente algunos de los problemas que se plantean a la hora de llevar a cabo la gestión de información cuando el tipo de lenguaje documental que se utiliza es de los denominados **lenguajes de indización**[2].

Se parte de conjuntos de términos lingüísticos asignados tras un proceso de análisis documental: términos, sí, de lenguaje natural, pero que son propios de un lenguaje documental de indización, y además de un lenguaje de indización de **vocabulario**

---

1 Dentro del conjunto de procesos que constituyen la denominada «cadena documental», existe una operación de las más importantes y delicadas: transcribir a un lenguaje documental el contenido de los documentos del fondo que se procesa. Esta tarea, una de las que se ha llamado «operaciones bisagra de la Documentación», es decisiva para el acceso eficaz a la información incluida en las bases de datos documentales. El proceso implica seleccionar los puntos de acceso que resultarán más útiles para el momento de la recuperación de la información.

2 Los lenguajes de indización permiten representar el contenido de los documentos de forma analítica (en oposición a los lenguajes de clasificación, que se emplean normalmente para identificar los contenidos de forma sintética, a veces codificada). Se trata de vocabularios más o menos limitados, cuyos elementos léxicos están extraídos generalmente del conjunto de palabras del lenguaje natural dotadas de mayor y más distintiva carga semántica.

controlado[3]. En concreto, se han realizado pruebas con los **campos 650** de un subconjunto de registros IBERMARC de monografías[4], haciendo uso únicamente de los **encabezamientos de materias**, sin procesar los subencabezamientos. El conjunto elegido está extraído de la base de datos que alimenta la agencia bibliográfica nacional española, pues debemos suponer que este fichero, punto de referencia obligado para el colectivo profesional de las bibliotecas, destaca por la calidad de sus informaciones.

Se ha tratado de analizar de qué manera se produce la **dispersión de frecuencias** en el uso de los descriptores que se han integrado en esos campos de materias, con el fin de evaluar cuál puede ser la incidencia de esta dispersión en la eficacia del sistema de recuperación de información y en qué medida se pueden utilizar **técnicas lexicométricas** para optimizar esa recuperación.

Generalmente, al analizar una base de datos referencial bibliográfica en la que se ha incluido uno o varios campos con descriptores controlados, podría esperarse de antemano que se cumpla una tendencia señalada por quienes han analizado el problema: que en este tipo de bases de datos existe una relación inversamente proporcional entre el número de registros que conforman la base de datos documental y el número de términos de indización diferentes que son asignados por los indizadores a tales registros.

Es decir, en condiciones normales, mientras el aumento del número de registros de

---

3 Se distingue, según si existe o no un control previo del conjunto de términos del léxico, entre lenguajes documentales de **vocabulario libre** y lenguajes de **vocabulario controlado**. Los primeros son elaborados a partir de los textos en lenguaje natural que se contienen en la base de datos con la que se trabaja. Por el contrario, los términos de lenguaje natural que componen un lenguaje documental de vocabulario controlado son seleccionados con anterioridad al procesamiento de los textos que componen la base. Antes de que se comience el análisis, manual o automático, de los documentos, se ha dejado preestablecido un listado de todos los descriptores que se pueden utilizar en la indización. Con la técnica del control se pretende solucionar los problemas que plantean algunas de las propiedades, inherentes a las lenguas naturales, que pueden producir ruidos en el momento de la recuperación de la información: las ambigüedades, sinonimias, polisemias, la problemática de la flexión gramatical, etc. Un lenguaje cuyos términos han sido sometidos a un control previo trata de ser un conjunto de elementos lo más unívocos posibles. En el momento de la recuperación de la información el uso de los lenguajes libres asegura una mayor **exhaustividad**, si bien disminuye la **precisión** de la búsqueda. Con los lenguajes controlados sucede justo lo contrario.

4 Los registros MARC son registros que contienen materiales documentales de muy distinto tipo, ajustados formalmente a la norma ISO/2709, estándar que describe una estructura de registro de tipo directorio, de gran flexibilidad y potencia en cuanto a sus prestaciones. IBERMARC es la versión española del formato. El campo 650 del formato IBERMARC para monografías contiene información sobre el tema del documento utilizando términos del tipo encabezamientos de materias.

estas bases de datos progresa aritméticamente, con un incremento constante, el número de términos de indización -encabezamientos de materias en nuestro caso- debería de experimentar un incremento decreciente.

Por tanto, si hubiera que representar gráficamente la evolución del tamaño de un catálogo bibliográfico, lo más ajustado sería utilizar una regresión lineal, mientras que el crecimiento del número distinto de materias debería constituir una regresión logarítmica.

Gráfico 1: incrementos en la base de datos

Obsérvese que nos encontramos ante un tipo de materiales que obedecen al pie de la letra la formulación de la **ley de Zipf**.

Los valores de d_1, d_2 ... d_i dan cuenta de la existencia de esa relación inversamente proporcional. Se trata, en consecuencia, de un fenómeno que, además de ser comprobable empíricamente, podría permitir establecer una proyección hacia el futuro. Es decir, resulta

posible conocer el valor de *d* en cualquier momento. Y este es un dato que debe ser tenido en cuenta para optimizar la recuperación documental

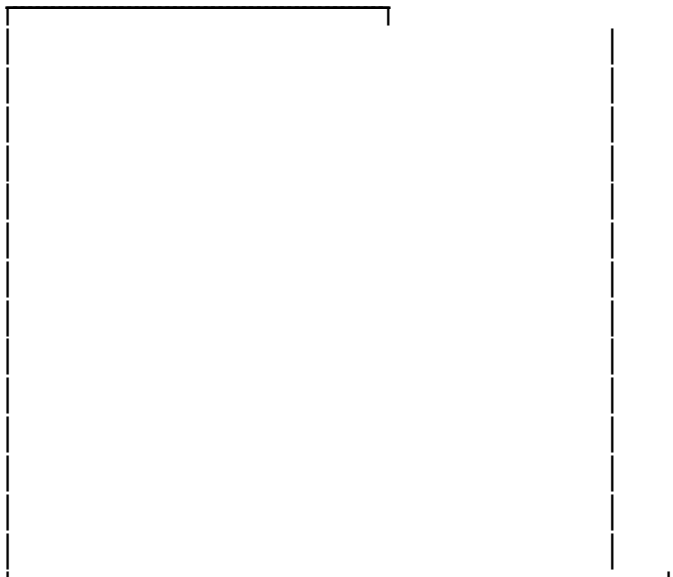

Gráfico 2: proyección de los datos anteriores

La observación de este sencillo fenómeno puede ponerse en relación con los que, hace unos treinta y cinco años, fueron los primeros intentos de preparar software para llevar a cabo procesos de **indización automática**, programas que desde entonces se emplean para analizar automática o semiautomáticamente los documentos, y también las consultas, de las bases de datos documentales, con el fin de poder eliminar buena parte del trabajo manual que tradicionalmente realizan los indizadores humanos.

En efecto, los primeros procedimientos automáticos para la creación de índices que se inventaron fueron los que realizaban **cálculos estadísticos** a partir de documentos

completos (generalmente abstracts de artículos científicos). Trataban de extraer del texto que se analizaba cierto número de términos característicos, retenidos en función de la **frecuencia absoluta** con que aparecían en el texto. Estos desarrollos se comenzaron a preparar a partir de hipótesis como la que formula Hans Peter Luhn en 1959: es probable que, cuantas más veces se utilice una palabra determinada dentro de un documento, más se pueda decir que esa palabra es un indicador del tema que dicho documento aborda. Si se comprobaba que tal hipótesis es cierta, se podría entonces poner a punto un programa informático que permitiera realizar de modo automático la indización del contenido de los materiales, extrayendo elementos de índices por medio de estudios estadísticos[5].

El rudimentario procedimiento se fue completando con ciertas mejoras, como las que se pudieron implementar gracias a la existencia de diversas herramientas que aportaba la **lexicometría**: por ejemplo la comparación de los listados de **frecuencia relativa** extraídos de un documento con las tablas de frecuencia media preparadas a partir de corpus de lenguaje estándar, que se utilizaron para retener como entradas de índices sólo aquellos términos que presentaran una frecuencia «anormalmente elevada»[6].

---

5 Lo primero que debería realizar el programa es un análisis de los textos que se le presenten, para extraer el listado de cada una de las palabras que los componen, con información sobre el número de apariciones de cada una. A continuación habría de reducir todas las flexiones a sus formas canónicas y clasificarlas por orden de frecuencia. Al hacer esto se observa que el resultado, en el caso de un artículo científico, por ejemplo, es una lista con gran cantidad de términos que aparecen una sola vez, otro gran conjunto, menor, integrado por los que aparecen dos veces, un grupo algo menor de palabras que se usan en tres ocasiones... En definitiva, que la distribución de palabras, agrupadas según el total de frecuencias de cada una, constituye una especie de pirámide, en cuya base -muy amplia- se sitúa el gran conjunto de frecuencia uno, dos, tres... y en la parte alta de la pirámide -muy estrecha- se encuentran una serie de términos, pocos en número, pero de muy frecuente aparición.

Los términos que más veces aparecen en los textos suelen ser de diversos tipos: por una parte, las **palabras más gramaticales**: artículos, preposiciones, pronombres...; luego, otras palabras de bajo contenido significativo - las palabras **atemáticas**-; otras muchas...; y, por fin, los términos que, por el tema específico del documento en cuestión, necesitan ser frecuentemente utilizadas. Estos serían, de acuerdo con el método de indización por cálculo de frecuencia, los términos seleccionados por el sistema como puntos de acceso a los documentos. Para conseguir localizarlos automáticamente, el programa, a continuación del recuento, ha de excluir las palabras contenidas en un **antidiccionario**: un listado de las **palabras vacías** de significado. Al procesar los textos y encontrar palabras contenidas en el antidiccionario, el programa no acumula las frecuencias de esas palabras, sino que directamente las ignora. Después de haber desechado tales términos, queda el listado de las **palabras significativas**, del que es posible entresacar las más frecuentes. Se podría entonces considerar como términos de índice las que aparecen por encima de una frecuencia preestablecida.

6 Después de elaborar el listado de términos extraídos de un documento y haberlos ordenado por orden de frecuencia absoluta, se calcula la frecuencia relativa de cada entrada del índice, en relación con el total de palabras que componen el documento. Las entradas, acompañadas entonces de sus respectivas frecuencias relativas, se ponen en comparación con otros listados de frecuencias relativas. Esos listados pueden haber sido elaborados a partir de corpus generales -extraídos de textos de la lengua estándar-, o bien a partir de documentos restringidos a un área temática específica.

Realizada la comparación, se observa que una serie de términos presentan un alto índice de frecuencia en el texto que se indiza, pero que esos mismos términos también cuentan con una frecuencia relativa similar en el listado que se ha elegido para hacer la comparación. Son las palabras omnibus, o atemáticas, presentes de forma relativamente generalizada en todo tipo de textos. Sin embargo, otras palabras, que presentan una notable desviación con respecto a las frecuencias relativas de los corpus, pueden ser consideradas probablemente indicadores del contenido de los documentos.

Estos ejemplos de **indización automática selectiva en lenguaje libre** sirven para poner de manifiesto la íntima relación de la cuestión con algunas de las disciplinas que por aquel entonces adquirían una importancia creciente dentro de lo que iba a ser la **lingüística computacional**: la **estadística lingüística**. Y también tendremos ocasión de referirnos a las leyes de la **bibliometría**.

La evolución de estos trabajos proporcionó con el tiempo la posibilidad de utilizar sistemas estadísticos para llevar a cabo procesos de **ponderación**[7] automática a partir del análisis de frecuencias de las palabras cuando son procesados textos completos. Nadie duda de la utilidad de tales procedimientos en ese contexto. La pregunta que nos hemos hecho nosotros es si puede tener sentido aplicar métodos similares en el caso de los materiales que ahora nos ocupan: grandes cantidades de descriptores –elementos léxicos de un vocabulario controlado– utilizados en los campos específicos de una base de datos bibliográfica.

El motivo de la pregunta es claro. Se puede tender a considerar que todas las entradas extraídas de un listado de encabezamientos de materias tienen el mismo nivel de jerarquía semántica. Sin embargo, la experiencia demuestra que esto no es así en ningún ámbito lingüístico: el uso (y el no-uso) pone de manifiesto la existencia de una jerarquía –¡la del uso!– con la que los elementos léxicos son capaces de marcar una mayor o menor especificidad.

Esta realidad destaca aún más en el caso de la base de datos que hemos analizado, pues ¡con el 10% de las materias que se han asignado, se puede recuperar casi las tres

7 El sistema de la ponderación es empleado en sistemas documentales tanto con procedimientos manuales como por procesos automatizados. Así algunos productores de bases de datos asignan descriptores de dos tipos: los **descriptores principales**, que describen el tema principal del documento que se analiza, y los **descriptores secundarios**, que se refieren a contenidos sobre los que el documento incluye información en menor medida. Otros asignan a cada descriptor un peso distinto, por ejemplo de 1 a 4, con el fin de evaluar la importancia de cada concepto indizado.

cuartas partes de los registros!

El dato es llamativo, y pone en entredicho la eficacia del sistemapero de todas formas hay que señalar que, al margen de las características específicas de la base de datos que nos ocupa, existe mucho interés en buscar nuevas soluciones para optimizar la indización y recuperación documental. Es por la necesidad que, desde hace años, tiene la Documentación de implementar nuevos sistemas para gestionar la información de cara a mejorar los sistemas de consultas a sus ficheros. La búsqueda en bases de datos documentales se articula, de forma casi exclusiva y generalizada, en torno a la combinatoria booleana, que, además de cargar con otros defectos, parte de una apreciación binaria sobre la pertinencia de cada registro con respecto a la consulta: un documento o es pertinente o es no pertinente (según si se produce o no un 'exact matching'). Sería deseable poder matizar esta apreciación, definiendo para cada documento un grado de pertinencia óptimo ('partial match') con respecto a una consulta[8] (Belkin, 1978).

Por estos motivos, si se comprueba la existencia de la ratio a la que nos referíamos anteriormente -relación proporcional inversa entre número de registros de la base y número de materias distintas utilizadas- ésta podría ser utilizada para asignar un factor de ponderación a cada elemento de indización diferente incluido en la base. En efecto, cuanto menos frecuente es un término dentro del conjunto de la base de datos, cabe asumir que mayor es su poder **discriminante**, y su **peso** debe ser más elevado dentro del conjunto de términos de la base. Dicho de otra manera, la capacidad discriminativa de un elemento de indización es inversamente proporcional a su frecuencia de uso en el total de

8 Entre los procedimientos «innovadores», cabe citar los métodos probabilísticos, los de espacio vectorial y los que recurren a la lógica difusa (Bookstein 1985).

los registros analizados.

Aplicar este tipo de procedimientos parece algo urgente. Si con un reducido número de descriptores es posible recuperar el 75% de las referencias de la base, habría que decir que tales descriptores son palabras omnibus, casi términos vacíos, lo cual cuestiona el sentido mismo del uso de las materias.

Se ha señalado (Larson, 1991) que el problema fundamental de la recuperación por materias es la cantidad de búsquedas que son abandonadas por los usuarios porque el sistema, como contestación a una consulta, suministra un conjunto de referencias de tamaño excesivo. Este fenómeno de 'information overload' delata que los analistas han sacrificado la precisión de los descriptores a costa de una abundante exhaustividad

El problema es determinar qué sistema de ponderación automática podría ser utilizado en este contexto. Las funciones más valoradas hoy son quizá las de Bookstein (1985), Robertson (1982), Salton (1983) y Van Rijsbergen (1979), y todos los años aparecen fórmulas nuevas. Hay que señalar, por otra parte, que excepto en muy pocos casos, como en SPIRIT y STAIRS, los sistemas de ponderación no son aún explotados en productos comerciales.

En un trabajo de Noreault, McGill y Koll (1981) se incluye un listado de 37 funciones distintas para calcular sistemas de ponderación de términos en los documentos y en las consultas. Aparte de una serie de funciones excesivamente simples, no resulta posible aplicar la mayor parte de ellas, pues están concebidas generalmente para ser aplicadas a colecciones de documentos completos.

La función que, por exclusión, parece más adecuada a nuestros trabajos es la de

Salton, Wong y Yang:

$$\log (N/d_i)$$

donde

N= Número de documentos (en nuestro caso referencias) dentro de la base de datos

d_i= Número de apariciones del término *i*

La utilización de esta función de ponderación nos permite establecer unas conclusiones similares a las expresadas por Van Rijsbergen: un término que tiene una frecuencia de aparición elevada no es muy útil para el momento de la recuperación, sea cual sea su distribución en la base. Los términos de frecuencia media son los más útiles (...).

Hasta el momento no es normal aplicar técnicas de ponderación partiendo de la distribución de los propios descriptores contenidos en las mismas bases de datos referenciales bibliográficas, quizá por el hecho de que se ha considerado que, al usar un lenguaje de vocabulario controlado, ya se producía una especie de proceso de ponderación en el momento en el que el indizador asignaba los descriptores. Sin embargo, al analizar el comportamiento de los usuarios ante las grandes bases de datos catalográficas de hoy, se revela que empiezan a aparecer problemas similares a los que fueron planteados en los días en los que comenzaba a ser posible gestionar bases de datos documentales de cierto tamaño con materiales en lenguaje libre. Parece que ha llegado el momento en que funciones de ponderación estadística sean incorporadas a los sistemas documentales de acuerdo con las especificaciones a las que nos hemos referido. Más adelante habrá que

estudiar la posibilidad de incorporar el **análisis de clusters** ligados a estas funciones de ponderación

REFERENCIAS BIBLIOGRAFICAS