CONSEJOS PARA LA BÚSQUEDA DE BASES DE DATOS: PARTE 2

Por Marvin Hunn

Esta es la Parte 2 de una serie de tres partes. La Parte 1 introdujo algunas técnicas prácticas de búsqueda en una base de datos estructurada, basada en metadatos de vocabulario controlado, utilizando operadores explícitos y sintaxis de búsqueda. La Parte 2 es un poco más teórica porque explica brevemente algunos problemas fundamentales que limitan el rendimiento del motor de búsqueda.

Sistemas de descubrimiento

¿Cómo deben los estudiantes tratar con la multitud de bases de datos? Puede ser difícil recordar qué base de datos cubre qué tema. Puede ser difícil recordar las diferencias entre los motores de búsqueda. Puede ser tedioso y lento buscar en tres o cuatro bases de datos diferentes. ¿No sería mejor buscar muchas fuentes a la vez? Esa es la fuerza motriz que creó el clamor de "búsqueda única".

Los sistemas de búsqueda única permiten al investigador buscar recursos de muchas fuentes diferentes (bases de datos) con una sola declaración de búsqueda. Hay dos formas principales de hacer esto. Un enfoque utiliza un software intermedio para traducir la búsqueda en declaraciones compatibles con los diferentes motores de búsqueda, enviar comandos a todos los diferentes sistemas y recopilar los resultados. Este enfoque se denomina búsqueda **federada**, búsqueda de difusión, búsqueda distribuida o búsqueda cruzada de base de datos. Este enfoque de búsqueda única comenzó en la década de 1970. Alrededor de 2010 se reemplazó por otro enfoque que recopila los datos (registros o artículos) de todos los proveedores diferentes y los indexa en una base de datos combinada, de modo que solo hay un índice centralizado. Este enfoque de cosecha e índice produce un sistema de "**descubrimiento** unificado." De hecho, es mejor que el enfoque de búsqueda federada. Esto es lo que OCLC WorldCat aspira a ser.

En el mejor de los casos, la búsqueda única simplifica las complejidades de la búsqueda. Solo hay una interfaz que dominar, y parece que solo hay una base de datos para buscar. Sin embargo, las personas que usan la búsqueda única pueden usarla como su único medio de buscar en las bases de datos de la biblioteca. Y eso es malo.

Hay muchos problemas con una sola búsqueda. La búsqueda individual puede insistir en buscar bases de datos que no sean relevantes para sus necesidades actuales. Esto reduce la precisión. La búsqueda única normalmente no admite la navegación, referencias cruzadas o funciones de base de datos especializadas. Puede mostrar muchos duplicados (el mismo artículo de varias bases de datos diferentes). En el caso del enfoque de búsqueda federada, lo limita a operadores de búsqueda simples admitidos por todas las bases de datos, y es lento porque debe esperar para recuperar los resultados de muchas otras búsquedas intermedias antes de combinar resultados. Los sistemas de descubrimiento unificados, por otro lado, requieren que el proveedor normalice los datos de fuentes dispares en un esquema común utilizado por el sistema de descubrimiento. Esta homogeneización generalmente significa la pérdida de campos distintivos o especializados (como el campo de las Escrituras de ATLA, por ejemplo). Esta es una limitación significativa.

Quizás el mayor problema con los enfoques federados y de descubrimiento para la búsqueda única es que **enmascara la necesidad de personalizar los términos de búsqueda para que coincidan con el vocabulario de cada base de datos**. Recuerde que todavía estamos hablando de bases de datos de vocabulario controlado, no de bases de datos de lenguaje natural. Para comprender la necesidad de hacer coincidir el vocabulario de cada base de datos específica, comparemos la terminología del tema relacionada con la depresión (el trastorno del estado de ánimo) en dos bases de datos que cubren asesoramiento.[1]

Oct 2018 búsquedas de campos de materias (no de palabra clave)			
	declaración de búsqueda (configuración predeterminada)	EBSCO Medline	EBSCO PsycINFO
1	depresión	160.007	173,070
2	"depresión emoción"	1	**24,597**
3a	"depresión mental"	1	28
3b	"mental, depresión" (observe el orden de las palabras)	9	29
4a	"trastorno depresivo"	**96,148**	59,789
4b	" depresión mayor"	935	**113 060**
5a	Prozac	18	68
5b	fluoxetina	**8,956**	**5,196**
6a	'inhibidores de la captación de serotonina'	**18.444**	**7.727**
6b	'inhibidores selectivos de la recaptación de serotonina'	353	**5127**
6c	'inhibidores de la recaptación de serotonina' OR 'inhibidores de la captación de serotonina'	18.679	10,414

Primero examine las líneas 1-4. Medline usa "*depresión*" para el trastorno del estado de ánimo leve / temporal, y "*trastorno depresivo*" para afecciones graves / crónicas. PsycINFO usa "*depresión (emoción)*" y "*depresión mayor*" para hacer la misma distinción entre depresión menor y mayor, pero el "*trastorno depresivo*" es común en los títulos. Ahora considera las líneas 5-6. *Prozac* es el nombre de marca de un medicamento utilizado para tratar la depresión, pero rara vez aparece en los campos temáticos. El nombre genérico (y técnico) de *fluoxetina* es un término de búsqueda mejor. Este medicamento es un miembro de la clase de inhibidores de la absorción de serotonina. Tenga en cuenta la importancia de usar "*inhibidores de la captación de serotonina*" en Medline, pero también usar "*inhibidores de la recaptación de serotonina*" en PsycINFO.

Diferentes bases de datos utilizan diferente terminología. A veces usan los mismos términos de maneras diferentes, incluso contradictorias. Una declaración de búsqueda que funciona bien en una base de datos puede funcionar mal en otra base de datos. La traducción automática entre títulos de materias en diferentes bases de datos podría, en teoría, mitigar el problema. A menudo, no existe una equivalencia de

[1] Medline cubre literatura médica técnica, incluyendo psiquiatría. PsycINFO cubre la psicología académica en general, incluida la psicología clínica. Estas bases de datos utilizan terminología estandarizada, pero no siguen los mismos estándares.

uno a uno predecible entre términos en diferentes bases de datos, por lo que la traducción automática es difícil, y esa puede ser la razón por la que rara vez se ofrece en la actualidad. La variación en la terminología entre bases de datos es un problema muy serio. Es un problema lingüístico fundamental (como lo son la sinonimia y la polisemia en general).

Otro problema importante con los sistemas de descubrimiento es que tienden a contener muchos registros dispersos. ¿Por qué? Los sistemas de descubrimiento están bajo presión para ser integrales. Pero no pueden obtener todos los metadatos que necesitan. Algunos creadores de A&I (compañías de extracción e indexación que crean los metadatos, como PsycINFO o ATLA) no cederán sus metadatos a los proveedores de descubrimiento. Estas compañías de A&I quieren que los usuarios utilicen el motor de búsqueda y la interfaz de A&I nativos para que los usuarios reconozcan de dónde provienen los metadatos. El reconocimiento de la marca ayuda a justificar un alto precio por el producto. Así que los sistemas de descubrimiento están recurriendo a otras fuentes para obtener metadatos. Por ejemplo, como no pueden obtener metadatos de PsycINFO, obtienen metadatos dispersos de los editores de revistas para muchas de las mismas revistas que están indexadas en PsycINFO. A veces, los propietarios de A&I licenciarán registros abreviados. Crean registros de metadatos completos (incluidos encabezados de materias y resúmenes) para su uso en su propio sistema, pero arriendan versiones dispersas / abreviadas de esos metadatos (que carecen de títulos de materias o resúmenes) a los proveedores de descubrimiento. La escasez de registros es un defecto común y significativo en los sistemas de descubrimiento.

Por lo tanto, los sistemas de descubrimiento usualmente tienen una combinación de registros completos con temas de vocabulario controlado y registros dispersos con títulos en lenguaje natural pero sin encabezados de materia. La búsqueda de un sistema de descubrimiento con tales registros mixtos puede ser difícil. Nuestra introducción al uso de WorldCat trata este tema en detalle (http://library.dts.edu/wc-intro).

A pesar de estas limitaciones, el descubrimiento tiene su lugar. Comprenda sus limitaciones, benefíciese de ello y sepa cuándo no confiar en él. Piensa en cómo buscas. Los productos de descubrimiento a menudo producen búsquedas "suficientemente buenas" de una manera rápida y conveniente. Pero no se deje engañar en la investigación perezosa. No permita que un atajo cortocircuite su pensamiento y su educación. La confianza exclusiva en la conveniencia del descubrimiento es un síntoma de la investigación perezosa. Para los indisciplinados, apoya el mito de la investigación fácil, refuerza los hábitos perezosos y facilita el uso no crítico de las fuentes. Resista estas tentaciones. Use sistemas de descubrimiento pero no haga mal uso; utilice pero no sea engañado en hacer trabajo descuidado.

Conceptos, no palabras

Generalmente, buscamos conceptos, o ciertos significados semánticos. Pero los motores de búsqueda buscan cadenas de caracteres (palabras, palabras parciales, frases, etc.) No entienden lo que significan las cadenas / palabras. La diferencia entre un concepto / significado y una palabra / cadena es muy importante. Considera lo siguiente.

- Una cadena de caracteres puede representar múltiples significados diferentes (conceptos). El desajuste entre concepto y cadena es evidente en la homografía (por ejemplo, una palabra deletreada "banco" que se refiere a una institución financiera y una **diferente** palabra deletreada "banco" que se refiere a un asiento que puede ser usado por varias personas a la vez) y la polisemia (por ejemplo, la palabra "estrellado" 'significa' lleno de estrellas 'y la **misma** palabra' estrellado 'significa' sufrir un choque violento 'como en' El huevo terminó estrellado contra el suelo. ').[2]

- Dos cadenas diferentes pueden tener el mismo significado (por ejemplo, "enojo" e "ira"). Esta es una sinonimia.

- Un concepto / significado se puede expresar de muchas maneras diferentes, usando muchas palabras y frases diferentes. El concepto / significado puede incluso expresarse de una manera original o novedosa. El lenguaje figurado hace que esto sea aún más complejo. Piensa en las parábolas que Jesús usó.

Por lo tanto, a menudo no hay una correspondencia de uno a uno entre una cadena y un concepto. Los motores de búsqueda utilizan tres enfoques principales para hacer frente a este problema.

1. El enfoque clásico es buscar cuerdas. Así es como funcionan los motores de búsqueda EBSCO y WorldCat. El enfoque clásico también puede considerar otras pistas textuales (como párrafos, mayúsculas, encabezados, etc.)[3] El enfoque clásico también puede usar vocabulario controlado y referencias cruzadas de un término a otro (término más amplio, término más estrecho, término relacionado). En los sistemas modernos, las referencias cruzadas se pueden mejorar con una base de datos de "datos vinculados" que especifica relaciones entre palabras como "esto es un miembro de eso" o "esto es un sinónimo de eso". Las relaciones apoyan inferencias para construir un "gráfico de conocimiento".

2. El enfoque lingüístico primero analiza un documento basado en un modelo de cómo funciona el lenguaje. Por ejemplo, para determinar qué podría modificar qué, el software identifica partes del habla, frases, etc. Tiene un diccionario organizado por "conceptos" (no realmente). Intenta discernir los conceptos deseados y buscarlos. El enfoque lingüístico también utiliza todos los elementos del enfoque clásico. Nuestros modelos de cómo funciona el lenguaje son primitivos, y

[2] Técnicamente, los homógrafos son dos palabras no relacionadas que se escriben de la misma manera. Pueden haber sido incorporados al inglés de dos idiomas diferentes, por ejemplo. Un polisemia, por otro lado, es una palabra con múltiples sentidos o significados diferentes. A veces hay un único significado original con el que los otros significados se relacionan lógicamente y del que históricamente se derivan. Verá muchos homógrafos en su vocabulario hebreo, y la distinción entre homógrafo y polisemia será importante si tiene que rastrear una raíz hebrea a través de lenguajes afines. Pero la distinción entre homógrafo y polisemia no importa para nuestros propósitos. Ambos son 'una cuerda, muchos significados'.

[3] BRS y Dialog hicieron eso en la década de 1980.

este enfoque no funciona (todavía) bien. Es poco probable que encuentre un motor de búsqueda de este tipo.

3. El enfoque de clasificación estadística utiliza el enfoque clásico para generar una lista de elementos coincidentes, y luego aplica ponderaciones calculadas para ordenar los resultados. Recientemente, el enfoque estadístico ha sido emparejado con métodos de aprendizaje de inteligencia artificial. Una forma de entrenar el motor de búsqueda es alimentarlo con calificaciones (por personas) para miles (incluso millones) de búsquedas. Las calificaciones indican precisión y recuperación. El software aprende a mejorar la clasificación de los documentos de esta manera. En algunas variaciones, el software también puede aprender a expandir la búsqueda al identificar otras formas de expresar la declaración de búsqueda. Así que el enfoque estadístico comienza con el enfoque clásico, mejora la clasificación y también puede evolucionar desde allí. Es el aspecto "evolutivo", la capacidad de aprender y mejorar, lo que se llama inteligencia artificial. Pero el software no "entiende" el lenguaje; solo usa prueba y error para encontrar nuevas formas (reglas) para acercarse más a la coincidencia con los datos de capacitación que proporcionan las personas.

El punto principal aquí es que los motores de búsqueda no han superado los problemas lingüísticos básicos. Las técnicas de inteligencia artificial producen resultados cada vez mejores, pero no confíen ingenuamente en los resultados. No busques de forma pasiva. En su lugar, revise activamente los resultados y piense en cómo mejorar los resultados hasta que sean satisfactorios. Usa tus conocimientos para mejorar tus búsquedas.