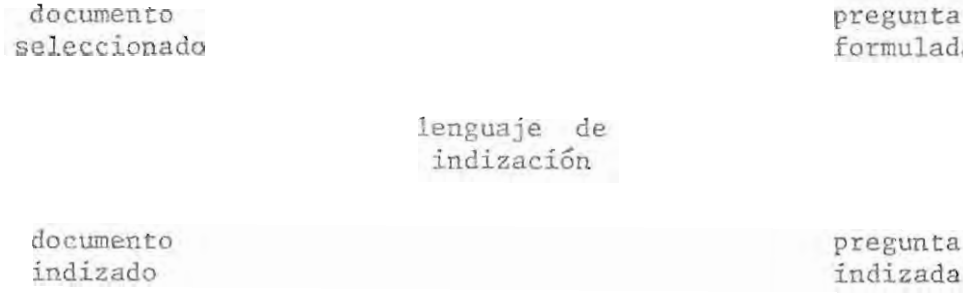
por las Profs. Ilda Idiart y Blanca Tamborenea

Todo proceso documentario involucra un problema de lenguaje. El contenido del documento se traduce al sistema en un lenguaje que llamaremos lenguaje de indización. El usuario, por su parte, expresa su necesidad de información en un lenguaje que el documentalista interpreta y traduce al lenguaje del sistema. O sea que el proceso documentario se realiza en dos etapas: a) se indiza el documento y se lo almacena o registra en el sistema (fichero tradicional, fichas muescadas, peek-a-boo, uniterm, sistemas fotoeléctricos, computadora...); b) ante una consulta se lo recupera de ese almacenamiento.

El punto a) constituye la entrada (input) al sistema y el b) la salida (output) del sistema. El primero es conocido como lenguaje de indización o lenguaje de entrada y el segundo como lenguaje de búsqueda o recuperación.

La mayoría de los sistemas utiliza un mismo lenguaje de indización y búsqueda. Otros utilizan un lenguaje de indización y otro de búsqueda. En ambos casos el lenguaje de indización es un puente entre el lenguaje de los documentos y el de los usuarios.



LENGUAJES

Los lenguajes se clasifican en:

a) Lenguaje natural: (no controlado, lenguaje libre).* Es el lenguaje utilizado por los índices permutados, KWIC, KWOC, concordancias, etc. Es un lenguaje de estructura y sintaxis complejas, rico por la cantidad de términos y matices. Simple como lenguaje de indización porque toma las palabras tal como se las encuentra en los documentos sin ningún esfuerzo de normalización. Dificultoso como lenguaje de búsqueda por los problemas de sinonimia y polisemia que acarrea.

b) Lenguaje controlado: Conjunto limitado de términos que se utilizan para la indización y recuperación en oposición al lenguaje natural que es limi

* Es el lenguaje utilizado por el autor para expresar su pensamiento.

tado. Aunque es nueva su denominación, son usados en bibliotecología desde antiguo. Son los utilizados por las clasificaciones, los encabezamientos de materia, los tesauros. Es un lenguaje de estructura y sintaxis simples, con menor riqueza que el lenguaje natural; más dificultoso como lenguaje de indización por la normalización que requiere y más simple como lenguaje de recuperación porque reduce la sinonimia y la polisemia.

En la actualidad, sólo se concibe al primero asociado a sistemas automatizados. Mientras que el lenguaje natural es esencialmente lineal, como el orden del discurso, el controlado, en oposición, es esencialmente vertical.

Según Chaumier (3), los lenguajes controlados se dividen de acuerdo con su organización interna en:

1) Lenguajes de estructura jerárquica: Son los que proceden por inclusión de una clase en otra.

C
B
A

Son las clasificaciones tradicionales tales como la de Dewey, CDU., LC., Colon Classification.

2) Lenguajes de estructura combinatoria: Nacen en la década del 50 como reacción a los anteriores y adquieren una gran expansión a partir de la década siguiente.

B

A

C

La indización se efectúa a nivel de concepto elemental. Los temas son representados por la intersección de clases más que por la inclusión de una clase en otra. Supone una etapa de análisis (extracción de las nociones elementales) y una de síntesis (reagruparlas por yuxtaposición) que se realizan también en dos etapas distintas del proceso documentario (indización y recuperación). Se los conoce como "indización coordinada".

Lancaster (7), establece que la sustancial diferencia entre los siste-

mas jerárquicos tradicionales y los coordinados reside en que si bien ambos tienen la habilidad de combinar clases, unos lo hacen en el momento de la indización y otros en el de la recuperación. Por eso llama a los primeros lenguajes precoordinados y a los segundos postcoordinados. Además los denomina respectivamente "no manipulativos" y "manipulativos". Los últimos permiten su recuperación por cualquiera de los elementos que lo componen mientras que los primeros solo pueden ser recuperados en el mismo orden en que fueron precoordinados.

Los lenguajes postcoordinados no tienen más de treinta años (Batten, Taube, Moers). Ofrecen una solución a la recuperación: la de obtener económicamente acceso múltiple a temas complejos. Los sistemas precoordinados pueden expresar también cualquier grado de complejidad temática, pero sin acceso múltiple, económico y eficiente.

Se puede establecer otra distinción:

a) Lenguajes enumerativos: (LC., Dewey, CDU.) son verdaderos inventarios de términos de indización y no tienen la flexibilidad suficiente para crear nuevos términos.

b) Lenguajes sintéticos: (clasificaciones facetadas, tesauros) además de enlistar términos de indización permite combinarlos en todas sus posibilidades para obtener temas de cualquier grado de complejidad. El tesauro es una herramienta para el control del vocabulario. Se basa en el principio de la postcoordinación y la síntesis.

HISTORIA

El término "tesauro" fue usado por primera vez por Luhn en 1957 y el primer tesauro con características actuales fue elaborado para la empresa Dupont en 1959 para indizar documentos en ingeniería química.

Posteriormente, en 1960, aparece el Thesaurus of ASTIA Descriptors (actualmente Defence Documentation Center, DDC), primer antecedente del actual Thesaurus of Engineers and Scientific Terms (4). A partir de ese momento se han publicado gran cantidad cubriendo casi todos los campos del conocimiento, especialmente en ciencia y técnica, en casi todos los idiomas, muchos de ellos plurilingües.

FUENTES

Las informaciones sobre lenguajes de recuperación en lengua inglesa son relevadas por la Case Western Reserve University (Cleveland, Ohio, 44106, USA).

Las informaciones sobre lenguajes de recuperación en otras lenguas son relevadas por el Centranly Instytut Informacji Naukwo -- Technicznej i Ekonomicznej (Al. Niepodleglosce 188, Varsóvia, Polonia).

Ambas organizaciones publican listas de tesauros.

ESTRUCTURA

Un lenguaje de indización como cualquier otro lenguaje comprende: un vocabulario, una sintaxis y reglas para su uso. La disposición del vocabulario es similar a la de los encabezamientos. La sinonimia y cuasi sinonimia es controlada por medio de referencias que remiten de los términos no aceptados al descriptor preferido.

sueldos véase salarios

haberes véase salarios

La referencia recíproca es indicada por UF (use for).

salarios
UF sueldos
haberes

La polisemia es eliminada por el agregado de un modificador entre paréntesis

Depresión (Meteorología)
Depresión (Medicina)
Depresión (Geografía)
Depresión (Economía)

o el agregado de un calificativo

Estructura económica
Estructura linguística
Estructura social

A veces se le agrega una nota de alcance (scope note) que no forma parte del descriptor y se usa para:

- restringir el uso del descriptor
FRECUENCIA DE MICROONDA
(1 a 300 GHz)

- explicar abreviaturas y acrónimos
SYNTOL
(Syntagmatic Organization Lenguaje)

- excluir algún significado
SUELDOS
(excluye las DIETAS de los concejales)

- definir un término
MARGINALIDAD
(individuos o grupos excluídos de participar en las principales corrientes de la vida social, política o económica)

Como en los encabezamientos de materia el tesauro usa una red de referencias cruzadas que relacionar los descriptores

ej.: PLANEAMIENTO URBANO
UF PLANIFICACION URBANA
BT PLANEAMIENTO
NT ZONIFICACION
RT CIUDADES
DESARROLLO URBANO
URBANISMO

UF: es el sinónimo o forma alternativa no usada

BT - NT: indica la relación jerárquica

RT: indica relación no jerárquica, un tanto vaga e inespecífica. Equivale a: "usted puede estar interesado en..."

Ambos tipos de referencias son recíprocas, cada BT tiene su NT recíproco y viceversa.

Las relaciones son similares a las de los encabezamientos de materia, pero es diferente su disposición.

Tesauro	Encabezamientos
-use	see
-UF incluye	ref. 1a. ó x
-BT specific to	ref. 2a. ó xx

-NT
generic to
-RT

see also
véase además

Estos dos últimos tipos de relaciones (NT - RT) no son diferenciadas en los encabezamientos de materia. La diferencia entre el encabezamiento de materia y el descriptor reside más en su uso que en su forma misma:

- el encabezamiento se usa solo o precoordinado.
- el descriptor en conjunción con otros descriptores.

Muchos descriptores son iguales a encabezamientos de materia, sin embargo, varios de ellos no podrían ser incluídos en una lista de encabezamientos:

URBANO
RURAL
EFICIENCIA
PRUEBA
HORARIOS

descriptores casi vacíos de contenido pero muy especiales para ser usados en coordinación con otros. Por esta misma razón, el descriptor normalmente no tiene subdescriptores precoordinados, sino que se lo usa postcoordinados con otros.

El descriptor es un término autorizado, usado para representar en forma inequívoca los conceptos tratados en los documentos, con la menor cantidad de palabras - preferentemente una y no más de tres - aunque los hay más extensos.

La Guía de Unesco (11) incluye normas muy específicas relativas a la forma de los descriptores:

- deben suprimirse en lo posible artículos y preposiciones
  PSICOLOGIA LABORAL y no PSICOLOGIA DEL TRABAJO
- la forma preferida es la sustantivada
  DEMOCRACIA y no DEMOCRATICA
- en cuanto al número se sigue el uso normal en la lengua: singular para términos específicos; plural para términos genéricos

ACEITES VEGETALES
ACEITE DE LINO
singular para objetos: CONDENSACION
plural para propiedades: MONUMENTOS

- la grafía es la más corriente en la lengua correspondiente
  PSICOLOGIA y no SICOLOGIA
- las transliteraciones siguen las normas de ISO
- se usan abreviaturas sólo si su uso está generalizado
  RADAR
  LASER
- es preferible la entrada directa a la invertida
  ABSORCION ATMOSFERICA y no ATMOSFERICA, ABSORCION
  de todas maneras el índice de permutaciones permite una entrada por los demás términos.
- la puntuación se reduce a: paréntesis, para distinguir homógrafos; guión, para palabras compuestas; puntos y comas para las notas de alcance.
- los numerales que forman parte de descriptores se transcriben en arábigo.

SELECCION DE LOS DESCRIPTORES

Se los elige de la literatura corriente (diccionarios, enciclopedias, glosarios, listas de encabezamientos) y de consultas a especialistas.

Una de las características de los tesauros es la capacidad de asimilar inmediatamente los neologismos que proliferan en los distintos campos. Para decidir su eleccción se tiene en cuenta:

1.- frecuencia de aparición en los documentos.
2.- efectividad y propiedad en representar un concepto.
3.- su relación con descriptores anteriores.

Se justifica incluír un descriptor en un tesauro si es suficientemente usado en la literatura corriente e índica el nivel de especificidad buscada por los usuarios del sistema.

Es el término que es capaz de describir el contenido de los documentos y las consultas hechas al sistema. Hay por lo menos dos maneras de confeccionar un tesauro:

1) una es la que deriva el vocabulario de la indización de documentos, llamada también "estalagmítica" porque procede de abajo hacia arriba.

2) la otra procede a la inversa extrayendo los términos de diccionarios, glosarios y otras fuentes que luego son cuidadosamente discutidas por equipos de especialistas. Este método llamado "estalagtítico" porque procede de arriba hacia abajo, es quizás el más ordenado pero no el más apropiado porque:
   a) puede estar en desacuerdo con el contenido de la literatura corriente y los requerimientos de los usuarios.
   b) puede establecer matices y distinciones entre conceptos, ajenos a los propósitos de la recuperación.

Algunos llaman a los primeros tesauros "a posteriori" aludiendo a que se confeccionan a medida que se indizan los documentos y los otros "a priori", porque tienen su forma definitiva cuando se comienza a indizar.

RELACIONES SEMANTICAS

Las relaciones conceptuales entre descriptores son señaladas por referencias de manera similar a los encabezamientos de materia.

Relacion de sustitución: (equivalencia o preferencia)

Es la que remite de un término no usado a otro usado por el sistema. La referencia de "use" (úsese) remite de sinónimos, cuasi-sinónimos o formas equivalentes del descriptor preferido:

BASURA	use	RESIDUOS
SUBDESARROLLO	use	DESARROLLO

esta referencia tiene su forma recíproca

RESIDUOS	DESARROLLO
UF BASURA	UF SUBDESARROLLO

La referencia es usada también para enviar de términos específicos no usados en la indización a términos más genéricos en los que está incluído.

AVENA use CEREALES

La misma referencia es usada para indicar que determinado concepto fue indizado por la combinación de dos o más descriptores.

AMPLIFICADORES DE MICROONDAS use AMPLIFICADORES Y MICROONDAS

la forma recíproca

AMPLIFICADORES
UF AMPLIFICADORES DE MICROONDAS +
UFC AMPLIFICADORES DE MICROONDAS

UFC: use for combination

el signo + indica que AMPLIFICADORES es sólo uno de los descriptores que se usa para describir el concepto AMPLIFICADORES DE MICROONDAS.

Cualquiera que fuera el uso que se haga de este tipo de referencia, implica siempre una decisión de usar un término en vez de otro, de no llegar a un nivel mayor de especificidad o de combinar dos descriptores para representar mejor un tema complejo.

Algunos tesauros utilizan la referencia

USE para indicar uso obligatorio

SEE para indicar uso opcional

Relación jerárquica:

Expresa la relación de subordinación de los términos. Si AB es un término más específico que A. A es un término más genérico que AB. Esta relación es recíproca y se indica:

PLANEAMIENTO URBANO
NT ZONIFICACION

ZONIFICACION
BT PLANEAMIENTO URBANO

Esta relación puede ser de dos tipos:

a) relación género/especie:

IMPUESTOS MUNICIPALES
BT RENTAS MUNICIPALES

b) relación parte/todo:

PULMONES
BT APARATO RESPIRATORIO

Si bien no es recomendable establecer diferencia entre ambos tipos de relación, hay tesauros que lo distinguen y las presentan antecedidas de las siglas:

BTG (broader term generic)

BTP (Broader term partitive)

NTG (Narrower term generic)

NTP (Narrower term partitive)

Hay términos que pertenecen a varias jerarquías. En este caso hay dos posibles soluciones:

1) se elige una sola de las jerarquías, presentando solamente aquélla que representa la característica inherente (dejando de lado las ocasionales)

2) se incluyen todas las posibles jerarquías.

Otro problema es el de los distintos niveles de jerarquías. EL INIS Thesaurus y el SPINES señalan varios niveles jerárquicos

TRATAMIENTO DE RESIDUOS
BT 1 INGENIERIA SANITARIA
BT 2 INGENIERIA CIVIL

"Autoposting" es la adjudicación automática de descriptores desde un NT A UN BT. Una vez asignados los descriptores a un documento, se agregan automáticamente todos los BT de los mismos. Esta modalidad facilita las búsquedas genéricas.

Relación asociativa: (o de afinidad)

Es la más vaga e inespecífica. Reúne descriptores relacionados (no sustitutivamente ni jerárquicamente). No todos los tesauros señalan esta relación y no siempres son recíprocas.

Incluye las siguientes relaciones:

1) subordinación colateral: dos descriptores que son NT del mismo descriptor, son entre sí RT

LINGUISTICA	GRAMATICA	FONETICA
NT FONETICA	RT FONETICA	RT GRAMATICA
GRAMATICA		
LEXICOLOGIA		

2) medio/proceso

MAQUINA DE ENSEÑAR
RT ENSEÑANZA PROGRAMADA

3) proceso/producto

PINTURA
RT PINTURAS

4) causa/efecto

ENSEÑANZA
RT APRENDIZAJE

5) proceso/persona

INSPECCION GENERAL
RT INSPECTOR GENERAL

6) objeto/propiedad

LASER
RT COHERENCIA

7) objeto/proceso

BARCOS
RT NAVEGACION

RELACIONES SINTACTICAS

Cuando se utiliza el lenguaje de descriptores, la indización del documento y el modelo de búsqueda es un conjunto no ordenado de descriptores. Esta falta de sintaxis adecuada crea distintos tipos de fallas que pueden ser reducidas a:

1) falsa coordinación: se relaciona en la búsqueda términos que no están relacionados en el documento

ej.: un documento que trata de las técnicas de soldadura del aluminio y purificación del cobre (SOLDADURA/ALUMINIO/PURIFICACION/COBRE) puede ser recuperado en una búsqueda sobre soldadura de cobre, tema que en realidad no está tratado en el documento.

2) Relación incorrecta de términos: deriva de la falta de indicación del tipo de relación existente entre descriptores.

ej.: un documento que trata del uso de la computadora en el diseño de aviones (COMPUTADORA/DISEÑO/AVIONES) puede ser recuperado en una búsqueda por diseño de computadoras cuando en realidad no trata ese tema.

Ambos tipos de fallas se deben a falta de sintaxis y son infrecuentes en encabezamientos de materia o lenguajes precoordinados. Basta crear términos precoordinados del tipo

SOLDADURA DE ALUMINIO
PURIFICACION DEL COBRE
DISEÑO DE AVIONES

para evitar tales inconvenientes.

La falsa coordinación y la relación incorrecta de términos se eliminan por el uso de ciertos recursos auxiliares de los tesauros que comportan una verdadera sintaxis del lenguaje de indización.

Los enlaces: (links)

Son símbolos agregados a los descriptores en los modelos de búsqueda o en los códigos de ubicación que permiten agruparlos conceptualmente y evitar la falsa coordinación:

SOLDADURA	R 1
ALUMINIO	R 1
PURIFICACION	R 2
COBRE	R 2

El documento sólo se recupera cuando los descriptores de la búsqueda están ligados de la misma manera que en la indización del documento.

Roles o funciones:

Son símbolos especiales que agregados a un descriptor reducen su alcance, indicando la función lógica que el descriptor desempeña en un determinado contexto. El uso de roles evita la falsa recuperación por relación incorrecta entre términos. Si en el ejemplo dado creamos una pequeña tabla de roles

COMPUTADORA	2	Rol 1: objeto de la acción
DISEÑO		Rol 2: instrumento
AVIONES	1	

nunca se recuperaría por una búsqueda sobre diseño de computadoras

COMPUTADORA	1
DISEÑO	

Cada tesauro, o cada materia tienen sus roles específicos, la más conocida tabla de roles fue elaborada por el Engineers Joint Council. Una aplicación similar tienen los operadores de Farradane.

Enlaces y roles, si bien eliminan las recuperaciones falsas en un 10% ó 15%, aumentan considerablemente el tiempo y costo de la indización, haciendo más difícil la consistencia. En general se los reemplaza por una mayor precoordinación.

PRESENTACION DEL TESAURO

La presentación del tesauro puede adoptar diversas formas. Fundamentalmente se las podría clasificar en presentaciones alfabéticas, sistemáticas y gráficas. Por lo general incluyen una variedad de combinaciones que tienden a complementarse dentro de un mismo tesauro. Puede adoptarse un orden determinado para el total de términos que lo componen e incluir a continuación otras formas de presentación de esos términos. Cada una de ellas constituye una vía de acceso complementaria y favorece la correcta utilización del mismo.

La presentación alfabética consiste básicamente en una sola lista en la cual la totalidad de los términos que componen el tesauro están presentados en orden alfabético. Dentro de ese orden la lista puede estar confeccionada desde la forma más simple hasta alcanzar distintos niveles de complejidad.

La presentación sistemática puede adoptar varias formas: grandes grupos o campos temáticos y dentro de éstos los términos en orden alfabético; ordenamiento jerárquico; o presentación facetada en grupos genéricos o detallada.

La presentación gráfica es menos frecuente y generalmente se complementa con una presentación alfabética. Por medio de flechas relaciona los descriptores dentro de cada campo semático e inclusive indica las relaciones con descriptores de otros campos. Utilizando un sistema de coordenadas permite visualizar los términos más genéricos y sus relacionados. Normalmente las flechas unidireccionales indican las relaciones jerárquicas y las bidireccionales las relaciones asociativas. En ocasiones, el grosor de la línea de unión expresa el grado de dependencia entre uno y otro término. Ejemplo de estos tipos de presentación son el EURATOM Thesaurus y el de Documentación Económica en la Administración de Empresas del Bureau M. van Dijk y G. Szanto. El primer tesauro presentado en forma gráfica fue el llamado "tesauro circular" de la TECK (Technisch Documentation-en Informatie-Centrum voor de Krijgsmacht).

Esta forma de presentación tiene la ventaja de permitir a simple vista obtener un panorama de los términos que integran un campo y la forma en que se relacionan. Pero si el campo es muy extenso o las relaciones jerárquicas

incluyen muchos niveles corre el riesgo de volverse confusa.

Normalmente, un tesauro debe constar, además de la presentación del cuerpo principal, de una introducción, indicaciones para su manejo, el plan de clasificación usado e índices. En la introducción, luego del propósito y estructura del tesauro y el campo que cubre, puede encontrarse una descripción de los antecedentes, las etapas de elaboración del mismo, las fuentes y métodos utilizados, las entidades y personas que intervinieron en su confección y las modalidades de actualización previstas.

En las indicaciones para el manejo una descripción de cada una de las partes constitutivas del tesauro con sus respectivos objetivos, características e indicaciones para utilizarlas. Según sea el orden adoptado para la presentación del total de términos el plan de clasificación podrá ser más general o detallado.

El TEST presenta un plan de clasificación en 22 campos con las principales subdivisiones en cada uno y luego, un plan exhaustivo que contiene todos los descriptores incluídos en cada subdivisión.

Los índices también dependen del orden de presentación adoptado. Pero dado que hay términos compuestos por más de una palabra siempre es conveniente la presencia del índice permutado ya sea en su forma KWIC o KWOC.

Si bajo cada descriptor se indican sólo las relaciones jerárquicas inmediatas en el cuerpo del tesauro, es sumamente útil el índice jerárquico, que bajo cada uno de los descriptores más genéricos, presente la lista de todos aquéllos que le son respectivamente más específicos e indique los distintos niveles de especificidad (ej.: el índice jerárquico del TEST).

Según el tema del tesauro se pueden encontrar otros tipos de índices: geográfico , de términos incorporados y deshechados en la última edición, de fórmulas incluídas, de identificadores, etc. Algunos de éstos pueden aparecer en forma de apéndice.

Se recomienda que en todos los casos las indicaciones sobre cada una de las partes que constituyen el tesauro sean claras, precisas y en lo posible, estén ejemplificadas.

Estos son, a grandes rasgos, las modalidades más frecuentes adoptadas por los tesauros en distintos campos. Sin embargo, existen otros más elabora-

dos, complejos y diferenciados que requerirían un análisis particular. Tal el caso del Thesaurofacet, el Spines, etc.

EL LENGUAJE EN EL SISTEMA DE RECUPERACION

La calidad de un sistema depende de su habilidad para:

1) recuperar documentos relevantes de la base de datos en respuesta a las consultas (conocida como recall).

2) desechar los documentos no relevantes, (conocida como precisión).

"Recall" (acierto) es la relación entre los documentos relevantes recuperados y el número total de documentos relevantes existentes en la colección. El coeficiente de "recall" se mide por la fórmula

$$\frac{R \times 100}{C}$$

en que R es el número de documentos relevantes recuperados y C el número total de documentos relevantes contenidos en el sistema.

Si en una búsqueda recuperamos 8 documentos sobre un tema y en la colección existen 10, el coeficiente de "recall" será

$$\frac{8}{10} \times 100 = 80\% \text{ de "recall"}$$

El 100% se podría lograr revisando toda la colección y descubriendo que el tema es tratado además en obras más generales o relacionadas.

Esta medida por sí sola no es muy importante si no está acompañada de otra que expresa la habilidad del sistema para separar lo que no interesa: coeficiente de "precisión", que se mide por la fórmula

$$\frac{R \times 100}{L}$$

en que "R" es el número total de documentos relevantes recuperados y "L" el tal de documentos recuperados.

Si en una búsqueda recuperamos 40 documentos, de los cuales sólo 8 son relevante y 32 no

$$\frac{8}{40} \times 100 = 20\% \text{ de "precisión"}$$

quiere decir que para nuestro caso el sistema opera con un 80% de tasa de "recall" y un 20% de "precisión". Esto indica una gran capacidad de recuperación y baja precisión.

La tasa de precisión mide el esfuerzo (tiempo) del usuario para separar los documentos irrelevantes del total de documentos recuperados. Evidentemente, es más difícil separar 8 relevantes de 32 que no lo son (20%) que separar 8 relevantes de 8 que no lo son (50%).

La tasa de precisión es una medida del esfuerzo del usuario para lograr un determinado "recall".

"Recall" y precisión son dos magnitudes inversamente proporcionales. Si aumentamos el "recall" ampliando las búsquedas, disminuímos la precisión. Si afinamos la precisión tendemos a reducir el "recall". Según Lancaster (7): el sistema de recuperación es un verdadero filtro capaz de dejar pasar lo que queremos mientras retiene lo que no deseamos.

Los lenguajes tienen recursos que permiten aumentar el "recall", es decir, asegurar la recuperación del mayor número posible de documentos relevantes y otros que tienden a impedir la recuperación de documetos no relevantes.

Recursos que permiten aumentar el "recall": en general tienden a reducir el vocabulario

a) control de sinónimos: se reduce el vocabulario, aumenta la recuperación, pero se pierde la capacidad de hacer diferencias sutiles (precisión).

b) control de la forma de los descriptores: evita la dispersión y aumenta la recuperación. Se puede llevar el control de forma al máximo y reducir los descriptores a raíces.

DROGA *

se pierde la diferencia (precisión) entre DROGA y DROGADICCION.

c) relaciones jerárquicas cuidadosamente presentadas: guían en búsquedas cada vez más genéricas que tienden a aumentar el "recall".

d) estructura de un vocabulario de entrada con remisión de términos específicos no usados a términos más genéricos que los incluyen.

Recursos que permiten aumentar la precisión: en general tienden a aumentar el vocabulario:

a) especificidad: abunda en detalles, aumentando el vocabulario y la precisión.

b) coordinación: tanto la precoordinación como la postcoordinación aseguran una identificación minuciosa.

c) uso de roles y enlaces.

d) ponderación: símbolos que agregados a los descriptores, indican el grado de importancia relativa con que son tratados los temas.

TIPOS DE TERMINOS

Descriptores:

Son los términos asignados a un documento para describir su contenido.

Especificadores:

Son términos que describen o especifican una clase. No siempre son descriptores sino que identifican algunos temas por combinación.

CAMARAS DE REFRIGERACION use CAMARAS y REFRIGERACION

Términos de entrada:

Son términos que permiten la entrada al sistema. Si se decide no llegar a gran especificidad se hace referencia de términos específicos a los más genéricos que los incluyen

AMPLIFICADORES DE MICROONDAS use AMPLIFICADORES

de esta manera el lenguaje disminuye la precisión pero no pierde "recall".

Esta distinción entre tipos de términos permite comparar vocabularios y determinar la capacidad de precisión de cada uno de ellos. La precisión de un vocabulario no está dada por la cantidad de descriptores, sino por la cantidad de temas que puede identificar separadamente, o sea por la cantidad de especificadores.

Un lenguaje bien estructurado tiene más términos de entrada que especificadores y más especificadores que descriptores.

Vocabulario de entrada:

Es el que conduce al usuario del término específico al término elegido por el sistema. Está constituído por términos de entrada y especificadores.

Una cantidad reducida de descriptores, con un buen vocabulario de entrada puede llegar a tener una eficiente capacidad de recuperación. Cuando el lenguaje de descriptores es genérico debe aumentar el volumen del vocabulario de entrada.

Un buen vocabulario de entrada facilita el trabajo intelectual de los indizadores y usuarios.

ESPECIFICIDAD DEL LENGUAJE

Es la habilidad de un vocabulario para expresar un tema con profundidad y detalle.

Los vocabularios muy específicos:

1) tienen gran cantidad de términos.

2) la indización es más dificultosa. Requiere personal más experto.

3) son más costosos.

A la inversa, los vocabularios más genéricos:

1) tienen menor cantidad de términos.

2) es más fácil obtener indización consistente.

3) son económicos.

En los sistemas "A", "B" y "C" las especificidad es distinta:

A	B	C
MAQUINAS AGRICOLAS	MAQUINAS AGRICOLAS	MAQUINAS AGRICOLAS
	COSECHADORAS	COSECHADORAS
		COSECHADORAS DE TRIGO

A pesar de que el sistema "A" no es tan específico como el "C", los tres tienen la misma posibilidad de recuperación si se hacen las respectivas referencias:

Sistema A: COSECHADORAS DE TRIGO use MAQUINAS AGRICOLAS

Sistema B: COSECHADORAS DE TRIGO use COSECHADORAS

La desventaja es que siendo distinto el nivel de precisión, en el sis-

tema "A", las COSECHADORAS DE TRIGO se recuperarán junto con todos los documentos sobre MAQUINAS AGRICOLAS y en el sistema "B", con todos los tipos de COSECHADORAS. O sea que trabajando con lenguajes poco específicos, pero con un vocabulario de entrada bien estructurado, si bien se ve perjudicada la precisión no se ve disminuída la capacidad de recuperación del sistema. Puede asimismo adoptarse un alto nivel de especificidad para temas centrales y disminuírlo para temas marginales.

PRECOORDINACION Y POSTCOORDINACION

Los tesauros son lenguajes eminentemente sintéticos. Los descriptores pueden ser combinados para representar temas complejos. Estos temas pueden recibir distinto tratamiento:

a) Precoordinados: estos términos tienen un alto nivel de precisión. Evitan las falsas recuperaciones. Un documento que trata de depósitos de basura y transporte de combustible, no será objeto de falsas recuperaciones si se lo indiza por descriptores precoordinados:

DEPOSITOS DE BASURA
TRANSPORTE DE COMBUSTIBLE

Tiene la desventaja de que aumentan notablemente el vocabulario y se corre el riego de que no se los busque precoordinados de la misma forma en que se los indizó.

ESTADISTICAS DE MORTALIDAD INFANTIL

puede ser precoordinado

ESTADISTICAS + MORTALIDAD INFANTIL o
ESTADISTICAS DE MORTALIDAD + NIÑOS

b) Poscoordinados por síntesis de sus factores linguísticos: los temas son representados por la combinación de sus palabras constitutivas

ESTADISTICA DE POBLACION use ESTADISTICAS y POBLACION
ABASTECIMIENTO DE AGUA use ABASTECIMIENTO y AGUA

Poscoordinados por síntesis de factores semánticos: los temas son representados por la combinación de conceptos más simples de acuerdo con su significado intrínseco.

TERMOMETRO use INSTRUMENTO y MEDIDA y TEMPERATURA

Las formas postcoordinadas son menos específicas que las precoordinadas, pero alcanzan un nivel mayor de "recall".

Se elige la forma precoordinada cuando su uso es frecuente o cuando su significado difiere al usarlo en forma postcoordinada

AMA DE LAVES
AMA DE CASA

Se elige la forma postcoordinada cuando no es frecuente el uso del descriptor por si mismo o cuando la forma precoordinada no es la habitual

JURISDICCION
HORARIO
CONSTRUCCION

por si mismos casi no tienen significado salvo cuando se los combina con otros descriptores. A su vez la forma precoordinada no es frecuente

HORARIO DE MERCADOS
JURISDICCION EN ZONA PORTUARIA

Cuando se plantean dudas acerca de la modalidad a seguir, es aconsejable precoordinar antes que postcoordinar.

EL ALGEBRA BOOLEANA

Cuando se trabaja en servicios automatizados, las búsquedas documentarias se deben traducir en una ecuación de búsqueda, una verdadera fórmula que es la clave lógica del acceso a la información almacenada en la computadora. Se usa el lenguaje de descriptores como vocabulario y el álgebra booleana como sintáxis.

Se aplican los principios esenciales de los conjuntos. De las operaciones lógicas sobre conjuntos, la recuperación de la información toma principalmente las tres que le interesan

unión A B

intersección A B

diferencia A B

/

Las formulaciones de búsqueda son presentadas en forma de sumas lógicas, productos lógicos y diferencias lógicas.

En la indización los descriptores no están relacionados entre si, funcionan independientemente. En cambio las búsquedas son formuladas en forma de operaciones lógicas. Si se busca información sobre: microformas, especialmente microficha y microfilm, la solicitud de búsqueda se formulará como sumas lógicas:

A	MICROFORMAS	
B	MICROFILM	A + B + C
C	MICROFICHA	

Serán considerados relevantes aquellos documentos que contengan cualquiera de los tres descriptores.

Si en cambio lo que se busca es: clasificación de la microficha y el microfilm, la ecuación de búsqueda será:

A	MICROFICHA	
B	MICROFILM	(A + B) C
C	CLASIFICACION	

Serán considerados relevantes únicamente los documentos que contengan los tres descriptores o por lo menos A y C o B y C .

Si lo que se busca es: "transporte de combustibles, principalmente, en países extranjeros", la ecuación será:

A	TRANSPORTE	
B	COMBUSTIBLE	(A.B) - C
C	ARGENTINA	

Será considerado documento relevante todo aquel que contenga A y B como descriptores pero será rechazado todo el que contenga C . La diferencia lógica debe ser usada con prudencia (si el documento tratara el transporte de combustible en Argentina, Estados Unidos y Canadá sería rechazado, aún cuando en realidad es relevante), porque la operación de diferencia es eliminatoria.

La unión también es simbolizada por O , la intersección por Y y la diferencia por SALVO y la formulación entonces tiene esta presentación:

(MICROFILM O MICROFICHA) Y CLASIFICACION

Se puede variar la estrategia de búsqueda variando la fórmula: disminuyendo el número de grupos unidos por la relación de intersección y aumentando los ligados por la operación de unión se obtiene mayor "recall". Procediendo a la inversa se obtiene mayor precisión.

TESAUROS Y COMPUTADORAS

La computadora cumple distintas funciones en relación con los tesauros:

Indización automática:

Se realiza por aplicación de técnicas estadísticas y basada en el principio de que las palabras estadísticamente relacionadas, también lo están semánticamente. La indización por asignación es la asignación de términos extraídos de un vocabulario controlado, basándose en al frecuencia de las palabras en los documentos.

Confección de tesauros:

Realiza distintas tareas como:

- generar automáticamente las entradas recíprocas.
- controlar consistencia en el formato de descriptores cuando aparecen como BT, NT o RT.
- agregar y eliminar descriptores.
- almacenar un tesauro de lectura automática para control de las indizaciones y búsquedas.
- llevar un registro de la historia de los descriptores y frecuencia de asignación de los mismos.
- impresión definitiva
- por la aplicación de técnicas estadísticas confecciona vocabularios controlados que se comportan como tesauros.

Sistemas "on-line":

En los sistemas que operan "on-line" el tesauro se utiliza en la indización para:

- traducir el lenguaje del indizador al del sistema.
- incorporar nuevos descriptores.
- sugerir descriptores.
- detectar errores y falta de consistencia en la indización.

en la búsqueda para:

- traducir el lenguaje natural del usuario.
- ayudar en la formulación de las estrategias de búsqueda, permitiendo la ampliación o especificación de las mismas a los efectos de obtener mayor "recall" o mayor precisión.

FUTURO DE LOS TESAUROS:

Queda un gran problema por resolver: la compatibilización de los vocabularios controlados. Las bases de datos operan con vocabularios distintos. Para facilitar la utilización de todos los recursos sería imprescindible la creación de lenguajes intermedios.

Lancaster anuncia un creciente uso del lenguaje natural en los sistemas de información.

El lenguaje natural tiene una gran ventaja respecto de los controlados, es totalmente específico.

El tesauro del futuro será totalmente distinto: en lenguaje natural, débilmente estructurado y concebido como instrumento de búsqueda más que como control del vocabulario de indización.

BIBLIOGRAFIA

1- AITCHISON, J. The thesaurofacet: a multipurpose retrieval language tool. (Journal of documentation, v. 26, n°3, 1970, p.187-203).

2- AITCHISON, J. y GILCHRIST, A. Thesaurus construction; a practical manual. London, Aslib, 1972.

3- CHAUMIER, J. Les techniques documentaires. Paris, PUF, 1971.

4- ENGINEERS JOINT COUNCIL. Thesaurus of engineering and scientific terms. New York, 1967.

5- LANCASTER, F. W. Information retrieval systems. New York, Wiley, 1968.

6- LANCASTER, F. W. Vocabulary control for information retrieval. Washington, Information Resources Press, 1972.

7- LANCASTER, F. W. Vocabulary control in information retrieval systems. (Advances in librarianship, v.7, 1977, p. 1-40).

8- MIJAILOV, A. I.- GUILIAREVSKII, R.S. y CHIORNII, A. I. Fundamentos de la informática. Moscú, La Habana; Nauka, Academia de Ciencias de Cuba, 1973.

9- NACIONES UNIDAS. Consejo Latinoamericano de Documentación Económica y Social. Uso de descriptores y tesauros. Santiago de Chile, 1971.

10- SOERGEL, D. Indexing languages and thesauri; construction and maintenance Los Angeles, Melville, 1974.

11- UNESCO. Guidelines for the establishment and development of monolingual thesauri. Paris, INISIST, 1973.

12- UNESCO. Guidelines for the establishment and development of multilingual thesauri. Paris, UNISIST, 1976.

13- VAN DIJK, M y VAN SLYPE, G. El servicio de documentación frente a la explosión de la información. Buenos Aires, CONICET, 1972.

14- VAN SLYPE, G. - VAN DIJK, M. y GUILLOT, M. Systemes documentaires et ordinateur. Paris, Les editions d'organisation, 1973.

15- VICKERY, B. C. Classification and indexing in science. 3.ed. London, Butterworthe, 1975.

16- VICKERY, B. C. Thesurus; a new word in documentation. (Journal of documentation, v.16, 1960, p. 181-189).