Listado de descriptores libres y listado de palabras clave

Manela Juncà Campdepadrós

PID_00144350

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

Este módulo os introduce en el uso de los lenguajes libres. Comprende dos lenguajes: los listados de descriptores libres y los listados de palabras clave.

Itinerario de estudio

El módulo empieza con las características de los lenguajes libres, ya que los dos últimos lenguajes del curso comparten esta tipología. A continuación, se describen las listas de descriptores libres, su creación y el proceso de indización y recuperación. Finalmente, el curso acaba con las listas de palabras clave y la indización automática.

Tabla. Conceptos más importantes

Concepto	Ved
Lenguaje libre	1. Lenguajes documentales libres
Descriptor libre	2. Listado de descriptores libres
Palabra clave	3. Listado de palabras clave
Palabra vacía	3.1.2. Funcionamiento y evolución de los programas de indización automática
Cálculo de frecuencia	3.1.2. Funcionamiento y evolución de los programas de indización automática
Frecuencia inversa	3.1.2. Funcionamiento y evolución de los programas de indización automática
Discriminación	3.1.2. Funcionamiento y evolución de los programas de indización automática
Análisis morfológico	3.1.2. Funcionamiento y evolución de los programas de indización automática
Análisis sintáctico	3.1.2. Funcionamiento y evolución de los programas de indización automática
Análisis semántico	3.1.2. Funcionamiento y evolución de los programas de indización automática

Objetivos

Con el estudio de los materiales asociados a este módulo alcanzaréis los objetivos siguientes:

1. Conocer la aplicación de los lenguajes libres al análisis de contenido.

2. Valorar las ventajas de los lenguajes libres en la indización.

3. Valorar los inconvenientes de los lenguajes libres en la recuperación.

4. En cuanto a los **listados de descriptores libres**:
- Definir los listados de descriptores libres y sus tipologías.
- Conocer el proceso de creación del lenguaje.
- Adquirir una cierta habilidad en la indización de documentos.

5. Por lo que respecta a los **listados de palabras clave**:
- Introducirse en los mecanismos de la indización automática.
- Reconocer las aplicaciones estadísticas y lingüísticas aplicadas a la indización automática.
- Saber simular/entender el proceso de selección y elección de palabras clave.

1. Lenguajes documentales libres

Este último módulo está dedicado a los dos lenguajes documentales libres: los listados de descriptores libres y los listados de palabras clave.

Tipología según el nivel de control

Recordemos que la tipología según el nivel de control clasifica los lenguajes documentales en libres o controlados, en función de si los términos de indización corresponden a un lenguaje natural o a un lenguaje artificial construido para garantizar la indización y recuperación.

A diferencia de los lenguajes controlados, que neutralizaban las carencias de los lenguajes naturales y requerían conocimientos elevados para ser aplicados, los lenguajes libres no ofrecen tantas garantías en la recuperación, pero son amigables. Son lenguajes muy usados, ya que son:

- **Baratos**: los gastos de construcción son mínimos. Los elabora una persona o un programa informático.
- **Rápidos**: tanto en la construcción, que es inmediata, como en la actualización de su vocabulario, que se va incorporando al listado a medida que va ingresando en el fondo documental.
- **Terminológicamente ricos**: de resultas de la actualización inmediata de sus listas.
- **Fáciles de usar**: no necesitan conocimientos previos sobre lenguajes documentales, control de vocabulario, o precoordinación, entre otros.
- **Coherentes**: ofrecen coherencia máxima, de manera que dos servicios de información y documentación (SID[1]) con el mismo documento y software de indización automática llegarán a una indización idéntica (esta característica sólo es aplicable al listado de palabras clave).

[(1)]SID es la sigla de servicio de información y documentación.

Sin embargo, los lenguajes libres presentan inconvenientes en la recuperación, ya que al trabajar con lenguaje natural libre arrastran todos los problemas derivados de la ambigüedad (sinonimia, polisemia, homonimia), lo que provoca silencio y ruido documentales y presentan un número excesivo de resultados para cada petición. Es el proceso inverso de los lenguajes controlados, en los que la dificultad radicaba en el proceso de indización y aseguraba, en cambio, una recuperación unívoca.

Lenguajes controlados →	Dificultad en la indización →	Facilidad en la recuperación
Lenguajes libres →	Facilidad en la indización →	Dificultad en la recuperación

Actualmente muchas bases de datos combinan la utilización de los lenguajes controlados con los libres, de manera que podemos buscar y recuperar por diversas vías. Es habitual que los catálogos bibliográficos y otras bases de datos ofrezcan la posibilidad de buscar la materia de dos maneras:

a) Por el **campo *materia***: indizando con algún lenguaje controlado (tipo lista de encabezamientos de materia, tesauros o una lista de autoridades).

b) Por el **campo *palabra clave***: indizando con un programa de indización automática o lista de palabras clave.

2. Listado de descriptores libres

Un **listado de descriptores libres** es un vocabulario monolingüe de términos de indización ordenados alfabéticamente.

Estos términos son escogidos por el analista, sin verificar si existen ni cómo se introducen en una lista previamente establecida. Por lo tanto, no es un lenguaje controlado, sino libre, y por eso recibe este nombre.

2.1. Tipología

Las listas de descriptores libres son un lenguaje documental que tiene las **características** siguientes:

Ved también

Podéis ver la tipología de los lenguajes documentales en el apartado 5 del módulo "Análisis de contenido: resumen e indización" de esta asignatura.

1) Se trata de un tipo de **lenguaje analítico** (frente a los sintéticos). En el proceso de indización se identifican los conceptos que conforman el contenido del documento y se representan mediante una serie de descriptores. Por esta razón, su aplicación está ligada a la existencia de sistemas automatizados que permiten la recuperación de la información mediante la combinación de estos descriptores.

Ejemplo

Ponemos como ejemplo el resumen siguiente:

Dalmau Ribalta, A. (2002). *Los cátaros*. Editorial UOC ("Biblioteca Lectus Universitaria").

El catarismo es un movimiento cristiano disidente de la Baja Edad Media. Se extendió por diversos puntos de Europa, con una incidencia especial en el Languedoc. Esta obra da a conocer qué fue exactamente el catarismo y su trayectoria.

La obra *Los cátaros* procura distinguir la llamada *Iglesia de Dios* de toda una literatura esotérica y legendaria que la ha deformado de manera significativa y que, al mismo tiempo, se aleja de una concepción historiográfica tradicional, ya superada, que la consideraba como una simple prolongación del antiguo maniqueísmo.

A lo largo de esta obra se explica en qué consistía la liturgia, los ritos y la doctrina religiosa de esta Iglesia, considerada herética y perseguida por la Iglesia católica. Resumid su trágica historia con hechos tan relevantes como la llamada *cruzada albigense* y el nacimiento de la Inquisición.

El analista indiza más de un término[2]:

```
Baja Edad Media
cátaros
cristianismo
disidencia religiosa
Europa
liturgia
Languedoc
```

[(2)] En contraposición al LEMAC, que es un lenguaje sintético y sería un único término:

```
LEMAC: Càtars-Història
```

2) Se trata de un tipo de **lenguaje natural** porque la indización se basa en el uso de palabras o expresiones libres que pertenecen al lenguaje o discurso común y que se encuentran en los mismos documentos (como el título, el resumen o el texto) y a los conceptos de la consulta del usuario. Es decir, en ningún momento se utiliza un código (numérico, alfanumérico) que represente este contenido. El lenguaje natural presenta el problema de la falta de univocidad, es ambiguo y posibilita los fenómenos de la sinonimia, la polisemia y la homonimia.

Ejemplo

Listado de descriptores libres[3]:

```
Càtars
```

[3] En contraposición, indizado con un sistema de clasificación como la CDU, que es el único lenguaje documental codificado, sería:

```
CDU año 2004 27-789.67.
```

3) Se trata de un **lenguaje libre** porque está formado por términos de indización extraídos del lenguaje natural y que se utilizan o se sugieren en el mismo documento. El listado de descriptores resultante está constituido por una colección no ordenada de conceptos (sólo por orden alfabético). Estos conceptos están expresados por palabras o expresiones extraídas de los documentos, o propuestos por los mismos documentalistas, sin verificar si están en una lista establecida a priori (es también, por lo tanto, un lenguaje construido a posteriori).

4) Se trata de un tipo de **lenguaje postcoordinado**, ya que permite la coordinación de conceptos en el momento de la recuperación y la utilización de un gran número de términos de indización o puntos de acceso. La combinación precisa de estos diferentes descriptores, en la recuperación, se tiene que remitir al documento buscado.

5) Se trata de un tipo de **lenguaje de estructura combinatoria o asociativa**, en el que los conceptos o descriptores se organizan de una manera independiente y sólo se produce la combinación o intersección en las operaciones de indización y recuperación. Por lo tanto, se trata de un lenguaje en el que los términos que lo componen se organizan en una lista por orden alfabético sin respetar ningún otro tipo de estructura que sitúe cada descriptor en una cadena lógica, jerárquica o estructurada de conceptos.

Ejemplo

En este caso, el indizador hace:

```
Càtars AND Europa
```

2.2. Listado de descriptores libres en la indización

La lista de descriptores libres es un lenguaje documental que da libertad total al analista, por lo que lo convierte en muy amigable, rápido y fácil de usar.

Todos nosotros en algún momento de nuestra vida hemos ordenado fotos en álbumes, hemos clasificado los libros de la biblioteca o hemos ordenado artículos y fotocopias bajo algún tema o subtema. No éramos conscientes de ello, pero estábamos indizando con un listado de descriptores libres.

2.2.1. Creación de un listado de descriptores libres

El **proceso de creación** de una lista de descriptores libres es muy sencillo: el analista va leyendo los documentos que tiene que indizar y va tomando nota de los términos que son interesantes según su parecer. Los va anotando y los clasifica en orden alfabético. Les da un mínimo de forma y procura eliminar duplicidades de número y género (singular-plural, masculino-femenino), usa una única lengua y elimina algunos sinónimos.

Los descriptores libres tienen un gasto de construcción casi nulo. Son inmediatos, permiten un vocabulario actualizado, y se adaptan a la realidad del SID.

En algunos casos el listado de descriptores libres suele ser el primer paso en la tarea de crear un lenguaje documental. Los analistas empiezan por redactar un listado de descriptores libres con la intención de controlarlo y avanzar hacia un lenguaje documental realmente controlado. Si el analista quiere ejercer un poco de control sobre sus descriptores, sin llegar a trabajar con un vocabulario por completo controlado, como serían las listas de encabezamientos o los tesauros, tiene que tomar algunas medidas, como por ejemplo:

- Establecer una lengua o idioma de la lista.
- Controlar las formas singular/plural.
- Establecer los sintagmas nominales preferentes:
  - Sustantivo + adjetivo.
  - Sustantivo + preposición + sustantivo.
- Controlar los sinónimos.
- Utilizar formas usuales frente a otras muy técnicas.

Ved también

Las medidas de control de vocabulario detalladas en el apartado dedicado a las listas de encabezamientos de materia del módulo "Listas de encabezamientos de materia y listas de autoridades" son adecuadas también en este caso.

2.2.2. Indización

El analista lee el documento, determina la materia del documento y acto seguido propone un descriptor. Representándolo verbalmente, el analista diría "este documento trata de tal tema". No consulta ninguna lista, no comprueba si el término que piensa está aceptado o no, no tiene que seguir las reglas de combinación de ningún lenguaje; tiene libertad total a la hora de indizar.

El hecho de no necesitar una traducción del lenguaje natural a uno artificial controlado tiene ventajas y comporta inconvenientes. El primero, y el más evidente es la disparidad de indizaciones: ante el mismo documento, tres analistas pueden llegar a tres resultados muy diferentes.

Disparidad de indizaciones

Siguiendo con el ejemplo sobre el libro de los cátaros, vemos el resultado de la indización hecha por tres analistas diferentes:

Disparidad de indizaciones

Analista A	Analista B	Analista C
Baja Edad Media cátaros Francia movimientos cristianos disidentes	catarismo herejías órdenes monásticas religión	albigenses cismas historia medieval Languedoc

Como consecuencia de esto, la tasa de coherencia es la más baja de todos los lenguajes documentales.

Ejemplo

En el ejemplo, el uso de sinónimos y términos genéricos ha alterado tanto el resultado que no han coincidido en ningún descriptor (la tasa es del 0%).

Un segundo aspecto que hay que destacar es el grado de exhaustividad. Recordemos que el grado de exhaustividad puede ser alto, medio o bajo y que afecta a los lenguajes postcoordinados como éste, en el que el analista puede escoger el número de términos para representar un documento.

De esta manera, con una lista de descriptores libres podemos indizar el libro de los cátaros con tres grados diferentes:

Exhaustividad profunda o alta	Exhaustividad intermedia o media	Exhaustividad genérica o baja
Baja Edad Media catarismo cátaros cátaros, historia cristianismo cruzada albigense disidencia religiosa doctrina religiosa Iglesia de Dios Europa Inquisición liturgia Languedoc ritos religiosos	Baja Edad Media cátaros cristianismo disidencia religiosa Europa liturgia Languedoc	Baja Edad Media cátaros Languedoc

a) Las **ventajas** de este lenguaje en la indización giran en torno a la facilidad, la simplicidad y la libertad de uso:

- No necesita traducción de los conceptos del lenguaje natural de los documentos a un lenguaje artificial.

Ejemplo

Si en el texto de los cátaros que estamos usando como ejemplo se hablaba de *literatura esotérica y legendaria*, el analista puede indizar **leyendas** o **literatura esotérica** sin comprobar la forma en ninguna lista, si es aceptado o no, si se prefiere la forma singular/plural, si se puede adjetivar o no, si tiene relaciones semánticas que ayuden a la indización.

- Se trata de un tipo de lenguaje rápido y fácil de actualizar. Esta ventaja es muy apreciada en entornos tecnológicos en los que la terminología va más adelantada que su normalización en listas controladas. Recordemos que entidades como el Termcat son la fuente adecuada para consultar la forma correcta de escribir el término.

Ejemplo

Si el texto que hay que indizar habla del *Maniqueísmo* o de cualquier otro tema, no hay problema; se añade a la lista alfabética de descriptores **Maniqueísmo**.

- Se adapta perfectamente al nivel de usuarios y al tipo de SID, ya que es un lenguaje hecho a la medida de su centro.

Ejemplo

Si un SID de tipo genérico recibe este documento lo indizará con algún término del tipo **Religión**; en cambio, si es un SID especializado extraerá términos más concretos como, por ejemplo, **Cruzada albigense**.

- No hace falta una formación previa de los analistas. Precisamente la ausencia de reglas y principios hace innecesaria la formación.

b) Los **inconvenientes** giran en torno a la elección del término de indización entre todos los términos posibles. Un segundo problema derivado es prever de qué manera el usuario hará la búsqueda y así coincidir.

Ejemplo

En el ejemplo de los cátaros, el analista puede decidir indizar:

- Por el marco geográfico: Languedoc, Francia, Europa.
- Por el marco temporal: Baja Edad Media, Edad Media, historia medieval.
- Por el concepto *cátaros*: cátaros, catarismo, albigense.
- Por el concepto *disidencia*: disidencia, herejía, cisma.

Algunos analistas optan por indizar el término tal y como aparece en el documento (por ejemplo, Languedoc) y otros realizan la acción de escoger términos que creen que tienen más posibilidades de ser recuperados (por ejemplo, Francia).

2.2.3. Aplicaciones del listado de descriptores libres

Hemos comentado con anterioridad el uso generalizado de descriptores libres en el ámbito doméstico y que también suele ser un primer paso en el camino hacia la creación de un lenguaje controlado, pero también encuentra **aplicaciones** en los ámbitos siguientes:

1) En la **indización de información temporal**: cuando los ítems de información tienen una vida determinada ligada a un acontecimiento (político, deportivo o cultural) concreto que hace demasiado costoso invertir en un lenguaje controlado y su formación posterior.

2) En la **indización de documentos audiovisuales (imagen fija o en movimiento y audio) de la web**: los sistemas de indización automática de la web no pueden indizar, de momento, documentos que no contengan texto, como las fotografías o los vídeos de banda ancha, tan numerosos en la red. La solución pasa por indizarlos manualmente con la colaboración de los internautas.

3) En la **indización social** los internautas asignan etiquetas con descriptores a los recursos web. Mathes (2004) habla de *metadatos generados por el usuario*. Pueden indizar fotografías[4], webs[5], blogs[6] y compartir sus descriptores para colaborar en la indización de todo tipo de contenidos en el espacio web compartido y abierto. Cada recurso (por ejemplo, la web de la Wikipedia) es indizado por un grupo de usuarios que pueden coincidir o no con las etiquetas: uno puede indizar *wiki*, otro *enciclopedia*, otro *obra de referencia*, por ejemplo.

La asignación de las etiquetas se hace sin ánimo de lucro. Los internautas no buscan un beneficio económico, sino beneficiarse de búsquedas mejores. Los descriptores escogidos no se supervisan ni tienen ninguna estructura semántica.

Sinónimos

Otros nombres que recibe el fenómeno de la indización social son *tagging*, *etiquetado colaborativo*, *clasificación social*, *etiquetado social* o *folksonomies*. El origen del neologismo *folksonomía* lo debemos a Thomas van der Wal, que fusionó *folk* (gente, popular) y *taxonomía* (gestión de la clasificación), lo que dio como resultado una "indización gestionada popularmente".

La indización social participa de las características de las listas de descriptores libres en la filosofía de la indización, ya que cada participante indiza unos descriptores libres seleccionados por un proceso intelectual a partir del examen del recurso, sin verificar si los descriptores propuestos están o no en una lista establecida.

Los intentos por mejorar la recuperación van en la línea de aplicar algoritmos de ponderación y eliminación de etiquetas vacías que no aportan significado relevante para la comunidad. Estos mecanismos son totalmente invisibles para

(4)http://www.flickr.com/

(5)http://delicious.com/
http://www.mister-wong.es/users/500000111/

(6)http://technorati.com/

Lectura complementaria

Si queréis ampliar la información sobre los metadatos generados por el usuario, podéis consultar la página web de Adam Mathes:

- Folksonomies

el usuario. Algunos autores proponen alfabetizar al usuario dándole instrucciones para indizar, pero se duda de su eficacia, ya que una de las razones del éxito del *tagging* es la libertad que da al indizador internauta.

2.3. Listado de descriptores libres en la recuperación

a) Las **ventajas** que tienen los listados de descriptores libres en la recuperación son las siguientes:

- Como lenguaje analítico, permite la coordinación de diversos términos en la consulta, lo que permite búsquedas precisas.
- El lenguaje se actualiza continuamente.
- Facilidad de uso en la realización de consultas, ya que no hay que dominar previamente ningún lenguaje especializado, sofisticado ni artificial.

b) No obstante, también tienen **inconvenientes**. Si en la indización todo eran facilidades, en la recuperación encontramos todo tipo de problemas derivados del silencio y del ruido documentales.

El principal problema de este lenguaje documental se encuentra precisamente en este punto, la recuperación, ya que la abundancia de sinónimos y homónimos hace difícil predecir con qué término buscará al usuario y si coincidirá con el escogido por el indizador.

Las dificultades más notables se presentan a la hora de formular la petición:

- No elimina la polisemia del lenguaje natural, lo que provoca ruido documental.

- Presencia de sinónimos, lo que genera silencio documental e imposibilita recuperar los documentos que tratan sobre los conceptos de la consulta si éstos han sido indizados por los analistas con formas diferentes.

Ejemplo de sinónimos

Un ejemplo de búsqueda con todo tipo de sinónimos sería: CEE OR UE OR Unión Europea.

- Variantes ortográficas del mismo concepto (disco, disquete, diskete).

- Variantes idiomáticas (bafles y altavoces).

- Falta de una estructura semántica que ayude a la localización de términos. Esta falta provoca vacíos cuando pedimos un término genérico, con el cual no ha sido indizado el documento.

Ejemplo de falta de estructura semántica

Buscamos, por ejemplo, sobre lenguajes documentales y no recuperamos un documento sobre tesauro porque está indizado sólo como tesauro, no con un término conceptual-

mente más genérico como puede ser *lenguaje documental*. En un lenguaje controlado, esta vinculación se haría evidente (con siglas del tipo TG o TA) y ayudaría a la recuperación.

- Coordinación falsa. Entendemos por *coordinación falsa* recuperaciones inesperadas y erróneas que, a pesar de contener los elementos de la búsqueda, corresponden a temas diferentes.

Ejemplo de coordinación falsa

Buscamos, por ejemplo, sobre pintura catalana y pedimos Pintura AND Cataluña y recuperamos:

- Pintores catalanes (por ejemplo, M. Fortuny).
- Cataluña en la pintura (por ejemplo, la visión de J. Sorolla sobre el litoral catalán).
- Pintura en Cataluña (todos aquellos pintores que han pintado en Cataluña).
- Industriales de la pintura catalanes (pintores de paredes).

Las soluciones a estos inconvenientes pasan por recopilar toda la lista que seamos capaces de elaborar sobre el concepto: términos genéricos y específicos, polisémicos, sinónimos, siglas, variantes idiomáticas, cambios de género y número. Las estrategias de búsqueda suelen ser complejas.

3. Listado de palabras clave

El **listado de palabras clave** es un vocabulario ordenado alfabéticamente de los términos con carga significativa extraídos de un documento mediante un programa informático.

Listados de palabras clave

Consideramos *listados de palabras clave* los listados resultantes de la indización automática.

Una lista de palabras clave está constituida por una colección dispuesta en orden alfabético de las palabras significativas, llamadas también *no vacías* (es decir, todas las palabras que no son artículos, conjunciones, pronombres, preposiciones, numerales y ciertos verbos y adverbios), extraídas de una manera automática por el ordenador a partir del título, del resumen y cada vez más a menudo del texto completo de los documentos (van Slype). Acostumbran a ser listas monolingües.

Lectura complementaria

Si queréis ampliar vuestro conocimiento sobre la información dada por van Slype, podéis leer la obra siguiente.

G. van Slype (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales* (pág. 23). Madrid: Pirámide / Fundación Germán Sánchez Ruipérez (Biblioteca del Libro).

Los listados de palabras clave tienen la misma tipología que los listados de descriptores libres. Los dos lenguajes son:

- analíticos,
- naturales,
- libres,
- postcoordinados,
- de estructura combinatoria.

Las **diferencias** entre ellos, que son diferencias grandes, residen en que:

- Los listados de descriptores libres se basan en la indización humana y los listados de palabras clave, en la indización automática.
- Los listados de descriptores libres indizan conceptos y los listados de palabras clave indizan en su mayoría unitérminos.

3.1. La indización automática

La **indización automática** es el método por el que un ordenador aplica un algoritmo (o programa) al título, resumen o texto completo del documento para identificar los términos que puedan representar la materia y ser usados como términos de indización y recuperación en un índice o lista.

Hay dos **tipos** de indización automática:

a) Totalmente automatizada (en inglés, *automatic indexing*).

b) Semiautomática (en inglés, *machine-aided indexing*): el programa selecciona posibles descriptores procedentes de un tesauro o de una lista controlada y un documentalista acepta o rechaza la propuesta.

Los factores que hacen posible la indización automática son la elevada inversión de tiempo que conlleva la indización intelectual, el aumento de la documentación electrónica y a texto completo, y la extensión de sistemas de gestión documental en las instituciones y empresas, así como la evolución del procesamiento del lenguaje natural o PLN[7] (Méndez y Moreiro, 1999).

[7]PLN es la sigla de procesamiento del lenguaje natural. A partir de ahora, denotamos solamente PLN.

Lectura complementaria

Para ampliar la información sobre la indización automática, podéis leer la obra siguiente:

E. Méndez; J. A. Moreiro (1999). "Lenguaje natural e indización automatizada". *Ciencias de la Información* (pág. 11-24).

Aunque podemos encontrar aplicaciones dedicadas exclusivamente al resumen y a la indización automática, lo más habitual es que estos programas formen parte del mismo programa de gestión documental del centro.

3.1.1. Ciencias implicadas en la indización automática

En la indización automática intervienen diversas disciplinas científicas, la informática, la lingüística y la estadística:

1) La **informática** proporciona el algoritmo que representa, procesa, almacena y recupera las palabras clave.

2) La **lingüística** asigna etiquetas de tipo morfológico y lleva a cabo análisis sintácticos. Los programas de indización automática más evolucionados incorporan también análisis de tipo semántico:

- **Morfología**: trabaja en el ámbito de la palabra. Asigna la categoría gramatical de las palabras y la raíz común (o lema). Recordemos que las categorías gramaticales son sustantivo, verbo, adjetivo, adverbio, entre otras, y que morfológicamente pueden ser de diversos géneros, números, tiempo y modo. La lematización permite contar todas las veces que un concepto aparece en el texto aunque se exprese con formas diferentes.

Ejemplo de desambiguación

Por ejemplo, *columna* en un texto de arquitectura o *columna* en un texto de periodismo.

- **Sintaxis**: su campo de trabajo es la frase. Identifica los componentes de la frase, que son el sujeto, el verbo y los objetos, y su concordancia. Gracias a este análisis puede desambiguar gramaticalmente los términos no resueltos en la etapa morfológica.

- **Semántica**: se ocupa del texto. Desambigua términos polisémicos, ya que el texto permite optar por un significado u otro.

3) La **estadística** tiene diversas aplicaciones en documentación, pero en indización automática interviene especialmente en el cálculo de la frecuencia relativa de las palabras en un texto. Esta técnica calcula cuántas veces aparece una palabra en un texto y en función de unos baremos (ni las palabras más

repetidas, ni las menos) se seleccionan las palabras clave del documento. Los teóricos de este método son G. K. Zipf (1949), H. P. Luhn (1957), Sparck Jones (1972) y G. Salton (1989).

Trataremos con más detalle estos procesos en el próximo subapartado.

3.1.2. Funcionamiento y evolución de los programas de indización automática

Los primeros estudios en indización automática datan de las décadas de 1950-1960. Se basaban en principios estadísticos y probabilísticos. En la década de 1970, G. Salton incorpora el modelo de valor de discriminación y de relevancia de los términos. En la década de 1980 aparecen los criterios lingüísticos y en la de 1990 se plantea la indización automática de información multimedia (imágenes y sonido). Desde la década de 1980 hasta la actualidad, los programas de indización automática combinan el cálculo estadístico con el etiquetado morfológico y el análisis sintáctico.

Evolución de los programas de indización automática

1950-1960	1970	1980	1980-actualidad
Cálculos estadísticos y probabilísticos	Relevancia de los términos	Análisis lingüístico	Modelo mixto: estadístico + lingüístico

Los programas informáticos de indización automática buscan emular el proceso mental humano. La cuestión central que nos interesa es "¿cómo pueden detectar las palabras con más carga semántica, con más significado del texto?". Con este objetivo, suelen seguir los siguiente pasos:

1) Lectura de los documentos

El primer paso es leer el texto. Para hacerlo el documento se debe encontrar en formato electrónico. Actualmente, muchos documentos ya se encuentran en soporte electrónico. También son fácilmente accesibles los títulos, resúmenes, sumarios o bien el texto completo en las bases de datos de publicaciones periódicas y catálogos bibliográficos. Si se trata de documentación en soporte papel, primero hay que escanearla y aplicar un programa de tipo OCR (reconocimiento óptico de caracteres) para transformar la imagen de la página escaneada en texto electrónico.

Programas OCR

Estos programas OCR están implementados en la mayoría de escáneres domésticos. Recomendamos hacer la prueba de escanear una página y convertirla en texto.

¿Cómo puede un programa detectar una palabra? El programa detecta la palabra porque la considera una cadena de caracteres entre espacio en blanco y espacio en blanco.

2) Diseño

El equipo de documentalistas y programadores debe decidir:

- Las partes del documento que serán indizables. Son obligatorios el título y el resumen y es deseable el texto entero, sobre todo en artículos.

- El tratamiento que darán a las cifras, los signos de puntuación, los guiones, las mayúsculas/minúsculas y los acentos. Habitualmente son caracteres que no aportan significado, pero en determinados contextos pueden ser determinantes.

Ejemplo

Hay veces que estos caracteres sí tienen importancia:
- Cifras: TV3, 1492.
- Puntos, guiones, signos: www.uoc.edu, Canal+, e-mail.
- Acentos (útiles para diferenciar diacríticos): deu/Déu en catalán, te/té en castellano.

Una vez tenemos el documento en formato electrónico y el programa ha leído las partes seleccionadas, el paso siguiente consiste en distinguir las palabras que tienen significado de las que no lo tienen y, entre las que lo tienen, seleccionar las de más relevancia en la indización y su recuperación posterior. Este paso se resuelve aplicando diversos métodos estadísticos y lingüísticos, ya sea en un orden secuencial o bien alternando los métodos.

3) Palabras con contenido y palabras vacías

En el lenguaje natural hay palabras que tienen mucho significado (como los sustantivos y los adjetivos) y otras que no tanto (como los artículos o las conjunciones). Estas últimas se conocen como *palabras vacías*.

Las **palabras vacías** son palabras sin significado, como los artículos, los pronombres, las preposiciones, las conjunciones o los adverbios, que son filtradas antes o después del procesamiento del texto.

Las palabras vacías son muy frecuentes pero aportan poco valor de contenido semántico al texto y poca ayuda a la recuperación, ya que el usuario no las utiliza en la búsqueda y tampoco son recogidas en el fichero inverso de la base de datos. Se conocen como *listas de detención* en castellano y como *stop word list* en inglés (Hans Peter Luhn fue el creador del término y del concepto *stop word*).

Se elabora una lista de estas palabras vacías y se introducen en el programa, que va leyendo el texto palabra por palabra y las contrasta con el fichero de palabras vacías. Si el programa las encuentra en la lista, es que son vacías y no las indiza. Si no aparecen en la lista, es que tienen significado.

Palabras vacías

Artículos	el, la, uno, una
Pronombres	mío, nuestro
Verbos	comer, ser, reír (en todas las formas verbales)
Adverbios	tranquilamente, eficazmente
Preposiciones	a, con, de, sin

Numerales	cuarto, octavo
Adjetivos	alto, bajo, grande
Conjunciones	y, por lo tanto, pero, porque

Hay sistemas en los que la lista de palabras vacías:

a) Viene **predeterminada**: el sistema dispone, ya desde el principio, de la lista de palabras vacías de su idioma o idiomas. De hecho, su elaboración es fácil, ya que sólo hay que añadir las categorías vacías de una base de datos de terminología al idioma que se quiera. Los artículos siempre son los mismos, las conjunciones también, incluso los verbos se pueden llegar a contabilizar y flexionar en todos los tiempos verbales.

b) Se **evita expresamente** para permitir al sistema la búsqueda por frases y sintagmas. Los sistemas que los evitan disponen de otras herramientas para reducir significativamente el número de palabras indizadas, como técnicas de *stemming* o lematización que veremos más adelante.

Ejemplo

La lista de palabras vacías se puede evitar, por ejemplo, para recuperar un concepto como el diario *El País*, en el que el artículo tiene un papel importante.

c) Está **contextualizada** (*stop word context-dependent*). Cada sistema elabora la lista de palabras vacías según su ámbito temático. Contextualizar el listado permite evitar dos inconvenientes graves:

- Palabras con significado que se convierten en palabras vacías.

Ejemplo

Si un usuario busca en el catálogo de una biblioteca especializada en astronomía, no buscará por el concepto "Astronomía", ya que toda la colección de la biblioteca hará referencia a este concepto. El concepto *Astronomía* deviene vacío en este contexto.

- palabras vacías que se convierten en importantes en la indización.

Ejemplo

En un texto de historia los números (como 1914, 1936), los numerales (Jaime I) y los adjetivos pueden tener mucha carga significativa (Alta y Baja Edad Media; sobrenombres adjetivados de personajes como Jaime I el Conquistador y acontecimientos como la guerra fría). En el ejemplo siguiente, las palabras subrayadas son palabras vacías:

Palabras castellanas que son vacías o que no dependen del ámbito temático

Palabra	En el texto	Economía	Derecho	Educación	Multidisciplinario
ESO	"Eso es lo importante..." "La reforma de la ESO"	ESO	ESO	ESO	ESO
Cabo	"Se precisa llevar a cabo..." "La composición rocosa del cabo..."	cabo	cabo	cabo	cabo (geografía)
Tuya	"Yo traeré la mía pero tú trae la tuya." "La tuya es un árbol perenne."	tuya	tuya	tuya	tuya (árbol)

Fuente: extracto de I. Gil Leiva (2008). *Manual de indización. Teoría y práctica* (pág. 333). Gijón: Trea.

En el caso de la contextualización, las palabras vacías pueden ser introducidas manualmente o bien ser el resultado de algún cómputo estadístico de frecuencia.

Ejemplo

Supongamos el texto "Osiris es el primero de todos los humanos y dioses que vive una segunda vez".

Descompongamos ahora el texto como lo haría un programa de indización automática. Partimos del supuesto de que el programa tiene una lista de palabras vacías predeterminada que incluye preposiciones, conjunciones, numerales, verbos y artículos.

Texto	Comprobación en la lista de palabras vacías		Resultado de la indización
Osiris	no está	es un nombre propio	Osiris
es	es una palabra vacía	es una forma verbal	-
el	es una palabra vacía	es un artículo	-
primero	es una palabra vacía	es un numeral	-
de	es una palabra vacía	es una preposición	-
todos	es una palabra vacía		-
los	es una palabra vacía	es un artículo	-
humanos	no está	es un sustantivo	humanos
y	es una palabra vacía	es una conjunción	-
dioses	no está	es un sustantivo	dioses
que	es una palabra vacía	es una conjunción	-
vive	es una palabra vacía	es una forma verbal	-
una	es una palabra vacía	es un artículo	-
segunda	es una palabra vacía	es un numeral	-
vez	no está	es un sustantivo	vez

El texto queda indizado con estas palabras clave:

```
Osiris
humanos
dioses
vez
```

¿Son muy numerosas las palabras vacías en un texto? La presencia de palabras vacías no es en absoluto desdeñable. Los estudios estadísticos calculan que el 50% de las palabras de un texto son vacías. Así, un artículo de 8.000 palabras tiene unas 4.000 palabras vacías y otras 4.000 con significado. En el ejemplo de Osiris, las quince palabras de la frase han quedado reducidas a cuatro palabras clave (en este caso, el ejemplo buscaba pretendidamente muchas palabras vacías para ilustrar este subapartado).

El procedimiento en PLN para distinguir las palabras vacías de las palabras con significado consiste en la aplicación de una lista previa de palabras vacías o de una combinación de métodos estadísticos (frecuencia) y lingüísticos (análisis

morfológico, sintáctico y semántico). Desde que H. P. Hans Meter Luhn en 1957 eliminó las palabras vacías, esta técnica ha sido una constante en los posteriores programas de indización automática.

4) Métodos estadísticos

Los métodos estadísticos han sido la primera aproximación a la indización automática y todavía hoy en día son una parte consustancial. La teoría de fondo es el cálculo del peso (ponderación) de las palabras: ni las palabras más repetidas (por vacías) ni las menos repetidas (por específicas) son adecuadas para ser seleccionadas.

Los métodos estadísticos aplicados en PLN son de tres tipos: frecuencia, frecuencia inversa y discriminación; y se pueden usar solos o en combinación:

a) Frecuencia de aparición (ley de Zipf): el psicólogo y psicolingüista estadounidense George Kingsley Zipf (1949) propuso que el número de veces que un término aparece en un texto (frecuencia) y la posición que ocupa en una lista de palabras más o menos frecuentes (rango) es constante. Para Zipf los escritores usan un abanico reducido de palabras, preferentemente cortas, y su longitud es inversamente proporcional a la frecuencia.

Ejemplo

En la web hay disponible una copia de las palabras francesas y su frecuencia. En esta lista vemos qué palabras de uso cotidiano tienen una frecuencia más alta y qué palabras menos cotidianas a la inversa.

Frecuencia de aparición de las palabras francesas

aliments	13,97		grand	674,68
aller	234,58		laisse	120,97
ami	95,16		les	16.011,00
avec	3.019,71		neutre	11,52
collège	24,42		permutable	0,03
comité	73,71		pour	5.332,48
être	1.986,23		translateur	0,06
goût	98,77		vitamines	3,84

Fuente: extraído de http://www.lexique.org/listes/liste_mots.txt

Hans Meter Luhn (1957) aplica la ley de Zipf en el campo de la indización automática. Luhn ve que la frecuencia con la que se repite una palabra puede ser un buen indicador para seleccionar las palabras clave con más significado. Las que tienen una frecuencia muy alta no acostumbran a tener carga significativa (en el ejemplo anterior serian *avec, les* y *pour*, pero recordemos que en

función del contexto también podría serlo *aliments*) y llevan a la recuperación de muchos documentos. Las que tienen una frecuencia baja suelen ser palabras muy específicas en el texto y, por lo tanto, con carga significativa, pero en la recuperación presentan pocos resultados. Para Luhn lo mejor para indizar eran los términos con una frecuencia media.

Luhn propone los pasos siguientes:

- Calcular la frecuencia de todas las palabras del texto o colección.
- Clasificarlas en orden decreciente.
- Eliminar las de frecuencia más alta.
- Eliminar las de frecuencia más baja.
- Indizar con el resto de palabras.

Luhn aplica esta técnica para eliminar las palabras vacías. A partir de él, todos los sistemas utilizan esta técnica para crear la lista de palabras vacías.

b) Frecuencia inversa: Sparck Jones (1972) puso de manifiesto la capacidad de discriminación de un término frente a otro. Esta discriminación tiene que ser vista en el conjunto de la colección, no en un solo documento. Hay que comparar las palabras clave entre los documentos del fondo para detectar cuáles son realmente discriminatorias.

Ejemplo

Imaginemos un fondo de cuatro documentos.

- Documento 1: "Visitad el Museo Picasso de Barcelona en el barrio del Borne".
- Documento 2: "En Figueres también podéis visitar el Museo del Juguete".
- Documento 3: "El Museo de Figueres y el Museo Picasso son los más famosos".
- Documento 4: "En Figueres no dejéis de visitar el Museo Dalí".

Lectura recomendada

Para la resolución del cálculo ved:

Francisco Javier Martínez Méndez. *Recuperación de información en la red*.

Eliminemos las palabras vacías *el*, *los*, *a*, *al*, *del*, *de*, *los*, *más*, *no* y todas las formas del verbo *visitar* (*visitad*, *visitar*). Nos queda:

Doc.	Museo	Picasso	Barcelona	Barrio	Borne	Figueras	Juguete	Famosos	Dalí
1	1	1	1	1	1	0	0	0	0
2	1	0	0	0	0	1	1	0	0
3	2	1	0	0	0	1	0	1	0
4	1	0	0	0	0	1	0	0	1

Siguiendo el ejemplo, la palabra *museo* que aparece en cada documento no tiene mucha utilidad para discriminar el documento 1 del resto, en cambio Picasso, Barcelona y todas las palabras que aparecen poco serían discriminatorias. Para medir este valor de discriminación se propone la *frecuencia inversa*. El peso de una palabra aumenta si aparece pocas veces en un documento y disminuye si aparece a menudo en el resto de documentos.

Pesos de cada palabra

Museo	0		Borne	0,602
Picasso	0,301		Juguete	0,602
Figueras	0,301		Famosos	0,602
Barcelona	0,602		Dalí	0,602
Barrio	0,602			

c) Valor de discriminación del término: G. Salton (1989), a partir de la idea de que los vocablos de un texto se clasifican según su capacidad para discriminar unos documentos de los demás en una colección, ideó un sistema de indización, conocido como el **modelo de valor de discriminación**, que atribuye el peso o valor más alto a aquellos términos que causan la máxima separación posible entre los documentos de una colección. Es decir, el valor de un término depende de cómo varía la separación media entre los documentos cuando a un término se fija una identificación de contenido. Por lo tanto, las mejores palabras son aquellas que consiguen la distancia mayor. El análisis del **valor de discriminación** consigna una función específica en el análisis de contenido a las palabras simples, a las yuxtapuestas, a las frases y a los grupos de palabras.

El valor de discriminación de un término se define como la medida de los cambios en la separación espacial, que se manifiesta cuando una palabra cualquiera es asignada a una colección como término de indización para representar mejor las diferencias que pueda haber entre los documentos. Precisamente la asignación reduce la densidad espacial y, por el contrario, un discriminador pobre incrementa la densidad espacial. De este modo, si primero se calculan las densidades espaciales y se atribuyen a cada término, es posible especificar los términos en orden decreciente según sus valores de discriminación (Gil y Rodríguez, 1996).

Lectura complementaria

Si queréis ampliar la información sobre el sistema de indización ideado por Salton, podéis leer la obra siguiente:

I. Gil Leiva; J. V, Rodríguez Muñoz (1996). "Tendencias en los sistemas de indización automática. Estudio evolutivo". *Revista Española de Documentación Científica* (vol. 19, n.º 3, pág. 273-291).

5) Métodos lingüísticos

Los primeros analizadores lingüísticos son de la década de 1960-1970. Su aportación al análisis del contenido es primordial, ya que permiten analizar el texto en tres niveles de profundidad: palabra, frase y texto.

Cada uno de estos niveles es analizado por módulos del programa basados en diferentes disciplinas:

- Palabra: morfología.
- Palabra dentro de la frase: sintaxis.
- Palabra dentro del texto: semántica.

a) Análisis morfológico: es la rama de la gramática tradicional que estudia la forma de las palabras, independientemente de sus relaciones o funciones dentro de la frase. La forma de las palabras se distribuye en diversas categorías llamadas *partes de la oración*, por ejemplo nombres, adjetivos y verbos. También estudia las variaciones que pueden experimentar por motivo de los accidentes gramaticales como el género, el número, la declinación o la conjugación.

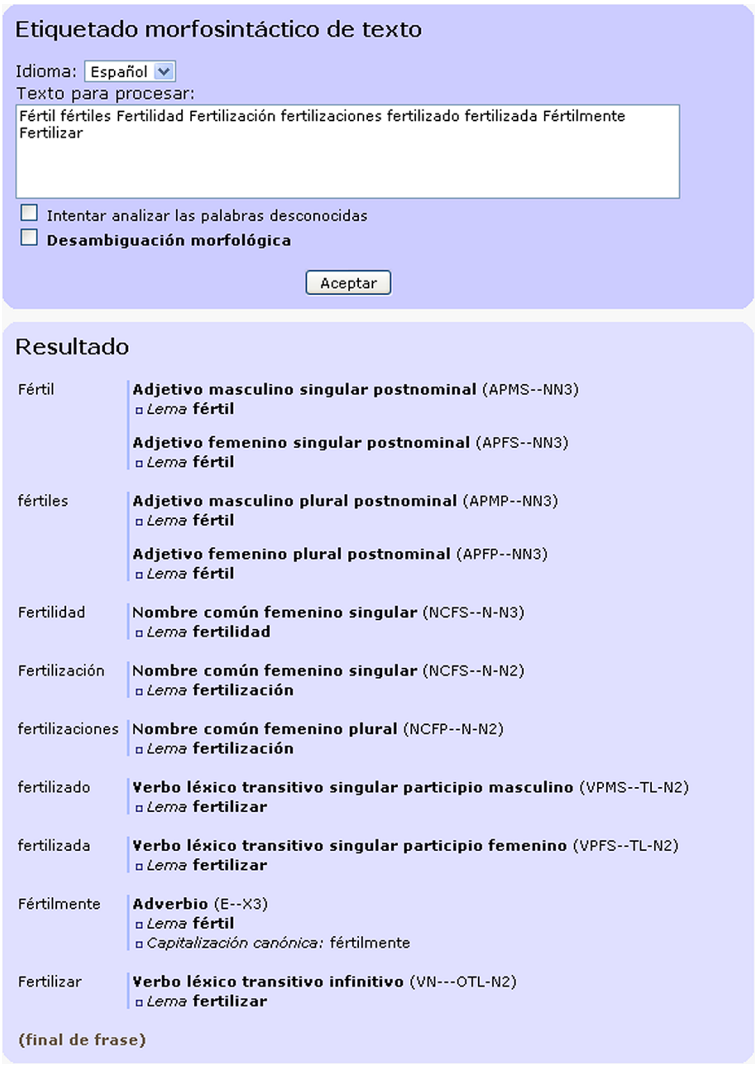
El análisis morfológico en indización automática se basa en reglas gramaticales y diccionarios de sufijos y afijos.

Su intervención en la indización automática es muy relevante en la formación de la lista de palabras vacías y una primera agrupación de las palabras, ya que:

- Asigna la categoría gramatical de cada una de las palabras del texto (como nombre común, verbo, adjetivo, adverbio o artículo), según cuál sea el género, el número, el tiempo y el modo. Con ello se reduce el abanico de palabras susceptibles de contener información, ya que detecta las palabras vacías y las flexiones derivadas de los accidentes gramaticales (como el género y el número). En la red hay aplicaciones gratuitas que analizan morfológicamente un texto; los ejemplos más comunes son Daedalus en castellano o Thera en castellano y catalán.

Ejemplo

Analizaremos un texto sobre fertilidad con Daedalus:



- Asigna la raíz común de diversas palabras. El mismo análisis morfológico procede a la lematización (*stemming*, en inglés), que consiste en localizar la raíz común de un grupo de palabras (fértil, fertilidad, fertilizar, fertilización). La lematización permite contar todas las veces que un concepto aparece en el texto, aunque se exprese con formas diferentes.

Ejemplo

En la recuperación, cuando el usuario busque por *fértil* recuperará todos los demás conceptos, ya que se encuentran indizados conjuntamente.

Sin lematizador	Con lematizador
fértil: un resultado fértiles: un resultado fértilmente: un resultado	fértil: tres resultados

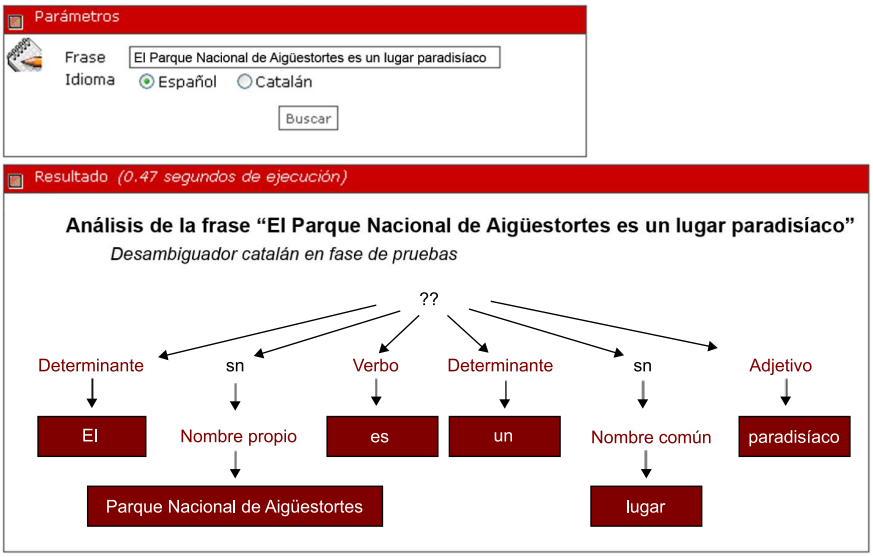
Gracias a este análisis, los cálculos estadísticos se pueden elaborar sobre un conjunto ya seleccionado de términos.

b) Análisis sintáctico: estudia la función o las asociaciones de las palabras dentro de la frase, lo que se llama *estructura sintáctica*. Identifica los componentes de la frase: sujeto, predicado, complementos u objetos y su concordancia. El programa de análisis sintáctico toma la frase e intenta reconocer la estructura con la que se ha construido dentro de un conjunto de análisis posibles conocidos como *gramática*.

Gracias a este análisis, puede desambiguar gramaticalmente los términos no resueltos en la etapa morfológica. Los analizadores sintácticos (*parsing*) se basan en analizadores morfológicos de la etapa anterior, gramáticas y diccionarios.

Ejemplo

Análisis sintáctico de la frase "El Parque Nacional d'Aigüestortes es un lugar paradisiaco" con Thera:



El análisis sintáctico es útil para los polisémicos o alguna palabra que por su similitud podría ser considerada vacía por el programa.

Ejemplo

Analizamos la frase: "Esta nota para mañana"[8], en la que morfológicamente cada palabra tiene diversas opciones, pero que el análisis sintáctico puede aclarar. Por ejemplo, *nota* puede ser un nombre común femenino (la nota escrita) o una forma verbal (3.ª persona del singular del presente de indicativo o 2.ª persona del singular del imperativo del verbo *notar*), pero dentro de la frase, por su posición, sólo puede ser considerada como nombre común y, en consecuencia, sería una palabra que habría que indizar, no una palabra vacía. Sin embargo, cabe decir que actualmente no todos los analizadores sintácticos pueden llegar a estas prestaciones y cometen errores.

[8]El ejemplo está extraído de I. Gil (2008). *Manual de indización. Teoría y práctica.* Gijón: Trea.

c) Análisis semántico. La semántica es la rama de la lingüística que estudia el significado de las palabras. Su utilidad reside en desambiguar sinónimos, polisémicos y vincular anáforas con el término al que hacen referencia. Sin embargo, esta parte del análisis es la más compleja para el software actual. Los analizadores semánticos se basan en diccionarios léxicos, redes semánticas, tesauros y ontologías.

Las redes semánticas

Estas redes son bases de datos que representan el conocimiento de una lengua o ámbito de una manera estructurada.

Para la indización automática son muy interesantes las bases de datos del tipo WordNet, que dan información léxica y semántica de los términos. Un algoritmo vinculado a la WordNet puede resolver casos de sinonimia o hiperonimia, ya que la base de datos proporciona relaciones básicas de significado.

WordNet y Euro WordNet son bases de datos multilingües en diversos idiomas (Euro WordNet comprende el neerlandés, el italiano, el castellano, el alemán, el francés, el checo y el estonio). Cada idioma diseña su WordNet pero mantiene la misma estructura que la inicial de Princeton en cuanto a los términos sinónimos con relaciones semánticas de relación de equivalencia entre ellos. Las diferentes WordNet están conectadas entre ellas y proporcionan acceso a una ontología compartida. En acceso gratuito sólo se encuentra la Wordnet de Princeton.

Los tesauros participan en el análisis semántico en que aportan las relaciones semánticas de equivalencia, jerarquía y asociación de cada palabra clave seleccionada. Las ontologías aportan, además, el detalle en cada tipo de relación. Un ejemplo es todo tipo de relaciones de asociación entre dos descriptores.

3.2. El listado de palabras clave en la indización

El programa de indización automática reconoce palabras, para ser más exactos, cadenas de caracteres entre un espacio y el siguiente. Por lo tanto, reconoce palabras, no expresiones formadas por más de una palabra. Decimos, pues, que el listado de palabras clave indiza unitérminos.

Unitérmino

Recordemos que los unitérminos son el elemento con significado más pequeño de un término de indización. *Lluvia* es un unitérmino. *Lluvia de estrellas* podría ser un descriptor de un tesauro formado por dos unitérminos: *lluvia* y *estrellas*.

Las **ventajas** que tienen las listas de palabras clave en la indización son las siguientes:

- Indización inmediata, en cuestión de segundos.
- Coherencia de indización exacta entre diversos SID. Dos SID con el mismo programa indizador llegarán al mismo resultado.
- Indización exhaustiva y tan específica como lo sea el contenido del documento indizado.
- Actualización rápida de la base de datos.
- Riqueza terminológica extraordinaria: evolución automática de la terminología, paralela a la misma evolución del conocimiento y de la ciencia.
- Costes mínimos de construcción y mantenimiento.

Uno de los **inconvenientes** que tienen las listas de palabras clave en la indización es que las palabras con significado se indizarán tal como salgan en el texto. Por lo tanto, puede indizar:

- El mismo término en diferentes idiomas, por ejemplo: *fuentes de información*, *fonts d'informació*, *information resources*, en función de la lengua del texto. Y no tiene relaciones de equivalencia entre los términos, como tienen los tesauros multilingües.
- Los errores ortográficos (*tesauris*, *thesaurus*).
- Singular, plural, masculino, femenino, entre otros (encabezamiento/encabezamientos, lista/listado).
- Siglas y conceptos sin distinción (SID/servicio de información y documentación; UE/Unión Europea).
- Sinónimos (lista de palabras vacías/*stop word list*).
- Homónimos (*Hierro*, y no sabremos si hace referencia a la isla o al elemento químico).

Como no hay control sobre el vocabulario, no controla la ambigüedad del lenguaje natural.

3.3. El listado de palabras clave en la recuperación

La indización automática sustenta los buscadores; en consecuencia, los usuarios de Google y Yahoo conocen de una manera implícita las ventajas y los inconvenientes de la recuperación de documentos con este lenguaje. Tal como dice Isidoro Gil (2008, pág. 108), el uso de Internet está convirtiendo a cada usuario en un *paradocumentalista* en potencia. Los usuarios son conscientes de que hay que escoger con cuidado las palabras clave, que deben ser cuanto más específicas mejor, poner el texto entre comillas, restringir la búsqueda con operadores booleanos, especificar fechas o ir añadiendo términos, entre otros, si queremos evitar el ruido documental.

Las **ventajas** que tienen las listas de palabras clave en la recuperación son las siguientes:

- Las búsquedas pueden ser muy precisas (todos los términos significativos están presentes).

- Son muy fáciles de usar para el usuario.

- Es el lenguaje más actualizado, evoluciona automáticamente al mismo tiempo que la terminología del documento. Podemos buscar *iPhone*, *teléfonos 3G*, *iPod*, *spam*, *spyware*, entre otros, términos que no constarían en ningún otro lenguaje documental, dado que el proceso de actualización es lento.

- La inmediatez: el documento llega y, en cuestión de segundos, está indizado y disponible para la búsqueda.

En cuanto a los **inconvenientes**, la indización automática se caracteriza por la posibilidad de indizar, si se desea, el texto entero del documento, lo que genera muchos puntos de acceso por materia, muchas palabras clave. Cuando realizamos una búsqueda, el buscador nos puede devolver miles de registros, lo que es un problema molesto (ruido documental). Sin embargo, esta característica también minimiza problemas potencialmente peligrosos (el silencio documental), como la pérdida de conceptos y la sinonimia.

Los principales inconvenientes son los siguientes:

- Pérdida de conceptos. Como el programa indiza unitérminos, los conceptos expresados de una manera compuesta se separan en dos unidades y pierden el sentido.

Ejemplo de pérdida de conceptos

Universidad de Barcelona, por ejemplo, se separa en universidad y Barcelona. Separar conceptos puede dar lugar a coordinación falsa en la recuperación como en este caso, en el que recuperamos todas las universidades de la ciudad de Barcelona: UB (Universidad de Barcelona), UPF (Universidad Pompeu Fabra), URL (Universidad Ramon Llull), entre otras.

- No reconoce la sinonimia. El programa no puede reconocer los sinónimos del mismo concepto. Si en el texto aparecen todos estos sinónimos, el programa no detectará ninguna semejanza entre ellos y los indizará todos: *cautivar*, *encantar*, *hipnotizar*, *sugestionar*.

- No reconoce la anáfora y la elipsis.

- No permite la univocidad.

- Aparece ruido y silencio documentales producidos por la ambigüedad propia del lenguaje natural (fenómenos de sinonimia, polisemia y homonimia).

- No aparece ningún tipo de control sobre variantes ortográficas, sobre plurales, adjetivos, verbos, siglas, entre otros, ni se ejerce ningún control sobre los errores ortográficos o de impresión que pueden aparecer en los documentos.

- Hay complejidad en el momento de plantear la estrategia de búsqueda; se necesita una agrupación de sinónimos completa, el uso de truncamientos y de operadores de proximidad.

Ejemplo de estrategia de búsqueda

Para buscar documentos sobre la Unión Europea, por ejemplo, habría que pensar en todos sus sinónimos y siglas como *UE* / *Unión Europea* / *European Union* / *CEE* / *Comunidad Económica Europea*.

Precisamente para minimizar los inconvenientes en la recuperación, los SID acostumbran a complementarse con otros lenguajes documentales preferentemente controlados, como un tesauro o una lista de encabezamientos de materia.

Actualmente la mayoría de catálogos en línea de bibliotecas universitarias nos permiten realizar búsquedas por diversos campos, de los cuales destacamos especialmente el de *palabra clave de materia*. Esta opción descompone el texto

(incluso el encabezamiento construido con una lista de encabezamiento de materia, es decir, con un lenguaje controlado y humano) en una sucesión de palabras clave, es decir, en un lenguaje libre y automático.

Actividades

Listado de descriptores libres

1. Indizad con un listado de descriptores libres el resumen siguiente con los tres grados de exhaustividad.

López Alonso, C.; Séré, A. (eds.) (2003). *Nuevos géneros discursivos: los textos electrónicos*. Madrid: Biblioteca Nueva.

Esta obra describe la transformación de determinados discursos sociales ante la revolución de las nuevas tecnologías de la información y de la comunicación (TIC). En efecto, Internet forma parte de nuestra vida diaria y ya no es fácil prescindir de este medio tecnológico. Esta obra, que consta de tres partes, analiza estos nuevos géneros discursivos:
En la primera parte, se describen los distintos textos electrónicos –el correo, los chats, los foros– y otros escritos en la red, como prensa, newsletters, E-zinc, etc.
En la segunda, se introduce al lector en la génesis y evolución de los lenguajes informáticos y sus diferentes tipos, especialmente los lenguajes de marcado y los modelos hipermedia.
La tercera, finalmente, se centra en los nuevos productos hipertextuales en el campo del discurso de transmisión de conocimientos con muy variadas utilizaciones: la enseñanza a distancia o textos de consulta como los diccionarios.

2. Indizad el resumen anterior para dos bases de datos diferentes: una genérica y la otra especializada en documentación.

3. Analizad detenidamente el ejemplo de los tres analistas que indizan el documento sobre los cátaros y decid por qué motivos son tan diversos los resultados.

Analista A	Analista B	Analista C
Baja Edad Media	catarismo	albigense
cátaros	herejías	cismas
Francia	órdenes monásticas	historia medieval
movimientos cristianos disidentes	religión	Languedoc

4. Elaborad una lista con las operaciones que hay que llevar a cabo sobre el léxico para transformar una lista de descriptores libres en un lenguaje controlado no codificado.

Listado de palabras clave

1. Indizad el texto siguiente simulando el funcionamiento de una lista de palabras clave:

Piero de Benedetto dei Franceschi (Sansepolcro, Toscana, 1416-1492) fue un pintor del *quattrocento* italiano conocido por su alias Piero della Francesca. También fue geómetra y matemático.

Es uno de los personajes principales y fundamentales del Renacimiento. Piero della Francesca es un personaje itinerante: es una figura que encontraremos en Ferrara entre 1447 y 1448 trabajando con Lionello d'Este; en Rimini para Sigismundo Malatesta en torno al 1450; en Roma para el papa Pío II entre 1458 y 1459, y en Urbino para Federico de Montefeltro en diversas ocasiones. Su actitud itinerante se suele comparar con la de Leon Batista Alberti. El autor nace en Borgo San Sepolcro (al norte de Florencia, en la zona de la Umbría) en 1416. El autor proviene de una familia de mercaderes y, por eso, sabía aritmética, cálculo, álgebra, geometría y contar con el ábaco. Después de la formación que tuvo para llevar el negocio familiar se formó con un maestro local que se llamaba Antonio di Anghiari para ser pintor. En aquellos momentos, había muchos estilos: todavía estaba vigente el linealismo y el lirismo de Fra Angelizo, Panozzo Gonzolo o Filippo Lipi y, por otra parte, estaba el realismo geométrico de Paolo Uccelo.

Tiene un estilo pictórico muy particular y, por lo tanto, es fácil de identificar. Utiliza una luz muy diurna y unos colores muy brillantes, que le provienen de su contacto con Domenico de Venezziano. En su obra siempre está presente el peso de la geometría, que implica, por una parte, utilizar la perspectiva lineal y, por la otra parte, reducir las figuras a la esencia. Las figuras de Piero della Francesca son muy estáticas, poco nerviosas, lo que va al revés que en el resto de pinturas renacentistas de Florencia, que a medida que avanzamos en el tiempo cada vez son más dinámicas. También vemos que sus figuras son poco expresivas y monolíticas. Longhi, cuando habla de Piero, dice que sus figuras son *columnes*.

"Piero della Francesca", *Wikipedia*

a) Diseñad las opciones del programa que simularéis (partes del texto, signos y símbolos).
b) ¿Cuántas palabras vacías habéis localizado?
c) ¿Qué inconvenientes planteará en la recuperación?

Glosario

análisis morfológico *m* Rama de la gramática tradicional que estudia la forma de las palabras, independientemente de sus relaciones o funciones dentro de la frase. La forma de las palabras se distribuye en diversas categorías llamadas *partes de la oración*, que son, entre otros, nombres, adjetivos y verbos. También estudia las variaciones que pueden experimentar por motivos de los accidentes gramaticales como el género, el número, la declinación o la conjugación.

análisis semántico *m* Rama de la lingüística que estudia el significado de las palabras. Su utilidad reside en desambiguar sinónimos, polisémicos y vincular anáforas con el término al que hacen referencia.

análisis sintáctico *m* Estudio de la función o de las asociaciones de las palabras dentro de la frase, lo que se llama *estructura sintáctica*. Identifica los componentes de la frase: sujeto, predicado, complementos u objetos y su concordancia. El programa de análisis sintáctico toma la frase e intenta hacer evidente la estructura con la que se ha construido dentro de un conjunto de análisis posibles conocidos como *gramática*.

automatic indexing *f* Ved **indización automática**.

clasificación social *f* Ved **indización social**.

coordinación falsa *f* Recuperación inesperada (y errónea) que a pesar de contener los elementos de la búsqueda corresponden a temas diferentes.

descriptor libre *m* Término de indización propio del lenguaje documental. Lista de descriptores libres.

etiquetaje colaborativo *m* Ved **indización social**.

etiquetaje social *m* Ved **indización social**.

folksonomía *f* Neologismo (Thomas van der Wal) resultado de la fusión de *folk* (gente, popular) y *taxonomía* (gestión de la clasificación), lo que da como resultado una "indización gestionada popularmente".
sin. **indización social**.

frecuencia de aparición *f* Ley de Zipf. El psicólogo y psicolingüista estadounidense George Kingsley Zipf (1949) propuso que el número de veces que un término aparece en un texto (frecuencia) y la posición que ocupa en una lista de palabras más o menos frecuentes (rango) es constante. Para Zipf los escritores usan un abanico reducido de palabras, preferentemente cortas y su longitud es inversamente proporcional a la frecuencia.

frecuencia inversa *f* Sparck Jones (1972) puso de manifiesto la capacidad de discriminación de un término frente a otro. Esta discriminación tiene que ser vista en el conjunto de la colección, no en un solo documento. Hay que comparar las palabras clave entre los documentos del fondo para detectar cuáles son realmente discriminatorias.

indización automática *f* Método por el que un ordenador aplica un algoritmo (o programa) al título, resumen o texto completo del documento con el fin de identificar los términos que puedan representar la materia y usar como términos de indización y recuperación en un índice o lista.
en *automatic indexing*.

indización social *f* Tipo de indización en red en el que los internautas asignan etiquetas con descriptores a los recursos web. La asignación de las etiquetas se hace sin ánimo de lucro, no se busca un beneficio económico sino beneficiarse de búsquedas mejores. Los descriptores escogidos no se supervisan ni tienen ninguna estructura semántica.
sin. **clasificación social, etiquetado colaborativo, etiquetado social, folksonomía**
en *tagging*.

lematización *f* Localización de la raíz común de un grupo de palabras.
en *stemming*.

listado de descriptores libres *f* Lenguaje documental. Vocabulario monolingüe de términos de indización ordenados alfabéticamente. Estos términos son escogidos por el analista, sin verificar si existen o cómo se introducen en una lista previamente establecida.

listado de palabras clave *f* Lenguaje documental. Vocabulario ordenado alfabéticamente de los términos con carga significativa extraídos de un documento mediante un programa informático.

machine-aided indexing *m* Ved **método de indización semiautomática**.

método de indización semiautomática *m* El programa selecciona posibles descriptores procedentes de un tesauro o de una lista controlada y un documentalista acepta o deniega la propuesta.
en *machine-aided indexing*.

método estadístico *m* Primera aproximación a la indización automática. Calcula el peso (ponderación) de las palabras: ni las palabras más repetidas (por vacías) ni las menos repetidas (por específicas) son adecuadas para ser seleccionadas. Los métodos estadísticos aplicados son de tres tipos: frecuencia, frecuencia inversa y discriminación. Se pueden usar solos o en combinación.

método lingüístico *m* Método que permite analizar el texto en tres niveles de profundidad: palabra (análisis morfológico), frase (análisis sintáctico) y texto (análisis semántico).

OCR *m* Ved **reconocimiento óptico de caracteres**.

palabra vacía *f* Palabra sin significado como, por ejemplo, artículos, pronombres, preposiciones o conjunciones, que es filtrada antes o después del procesamiento del texto. Son palabras muy frecuentes pero que aportan poco valor de contenido semántico al texto y también poca ayuda en la recuperación, ya que el usuario no las utiliza en la búsqueda y tampoco son recogidas en el fichero inverso de la base de datos.
sin. **lista de detención**
en *stop word list*.

palabra clave *f* Término de indización propio del lenguaje documental. Lista de palabras clave.

reconocimiento óptico de caracteres *m* Programa que transforma la imagen de la página escaneada en texto electrónico (de la expresión inglesa *optical character recognition*, OCR).

stemming *m* Ved **lematización**.

stop word list *f* Ved **palabra vacía**.

tagging *m* Ved **indización social**.

valor de discriminación del término *m* Método estadístico aplicado a la indización automática, desarrollado por Salton (1985). La medida de los cambios en la separación espacial que se manifiesta cuando una palabra cualquiera es asignada en una colección como término de indización para representar mejor las diferencias que pueda haber entre los documentos. La asignación reduce la densidad espacial y, por el contrario, un discriminador pobre incrementa la densidad espacial. De este modo, si primero se calculan las densidades espaciales y se atribuyen a cada término, es posible especificar los términos en orden decreciente según sus valores de discriminación.

Bibliografía

Alonso Berrocal, J. L.; Figuerola, C. G.; Zazo Rodríguez, A. (2009). *Recuperación avanzada de la información*. Salamanca. [Fecha de consulta: 31 de julio de 2009].

Cid, P.; Cuadrado, M.; Aguiriano, C. (1999). *Fonaments de llenguatges documentals* [documento electrónico]. Barcelona: UOC.

Gil Leiva, I. (2008). *Manual de indización. Teoría y práctica*. Gijón: Trea.

Gil Leiva, I.; Rodríguez Muñoz, J. V. (1996). "Tendencias en los sistemas de indización automática. Estudio evolutivo". *Revista Española de Documentación Científica* (vol. 19, n.º 3, pág. 273-291).

Hassan Montero, Y. (2006). "Indización social y recuperación de información". *No Solo Usabilidad* (n.º 5). ISSN 1886-8592.

Méndez, E.; Moreiro, J. A. (1999). "Lenguaje natural e indización automatizada". *Ciencias de la Información* (pág. 11-24).

Quintarelli, E. (2005). "Folksonomies: power to the people". ISKO Italy-UniMIB meeting: Milán. [Fecha de consulta: 31 de julio de 2009].

Salton, G. (1988). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading: Addison-Wesley.

Slype, G. van (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide / Fundación Germán Sánchez Ruipérez (Biblioteca del Libro).

Páginas web interesantes

Cibermetría

Cybermetrics. Recursos presentados por el departamento de CINDOC dentro del Consejo Superior de Investigaciones Científicas; artículos, seminarios, etc.

Blog spot. Acceso a algunos artículos sobre cibermetría.

Cybermetrics. Software disponible relativo a la cibermetría.

Webometrics and organizations. Enlaces hacia publicaciones y grupos de trabajo sobre cibermetría.

Applications of informetrics to information retrieval research. Artículo que relaciona la cibermetría con la recuperación de información.

The diameter of the world wide web. Artículo sobre la medición de la web realizada por los autores.

Inteligencia artificial

World Scientific: Connecting great minds. Vínculos a varias publicaciones periódicas sobre campos especializados como inteligencia artificial o sistemas neuronales.

AI international. Vínculos a entidades investigadoras sobre el tema como universidades y laboratorios, así como vínculos a congresos que versen sobre el mismo como la AAAI o las PRICAI.

Otras páginas web de interés

The Open Archives Initiative.

GOOGLE-WATCH. Página crítica sobre el sistema de recuperación de Google (el aparente uso comercial del orden de la recuperación).

RDF: modelo de metadatos flexible. Página sobre metadatos, con recursos web relacionados.

DUBLINCORE: Metadata Initiative (DCMI). Iniciativa de metadatos para páginas web, conversión de formatos... Proyectos, software, grupos de trabajo.

Pew Internet & american life. Web dependiente del Pew Research Center, con acceso a publicaciones de actualidad sobre internet.

z39.50 International Standard Maintenance Agency. Página sobre el mantenimiento de este protocolo de búsqueda e intercambio de información.