3.3- Recuperación de la información. Estrategias de búsqueda.

El proceso de recuperación de información consiste esencialmente en extraer de una colección de documentos aquellos que se ajustan a las especificaciones de un petición determinada. Se trata pues de una comparación sistemática entre los documentos o sus representaciones y la petición o demanda de información. Podemos descomponer el proceso en tres fases:

* Traducción del documento en un lenguaje de indexación. La expresión que resulta de este proceso se llama **modelo de búsqueda** del documento. Se trata de representar el documento por una serie de descriptores que lo determinen lo más directamente posible para su posterior búsqueda.
* Expresión de la petición de información en el mismo lenguaje del modelo de búsqueda. Se obtiene el denominado **perfil de búsqueda**.
* Comparación sistemática de los modelos de búsqueda de los documentos con el perfil de búsqueda, a fin de seleccionar los que se ajusten a este.

La operación resultaría relativamente sencilla si los lenguajes de indexación y de búsqueda coincidieran exactamente. Ello, sin embargo, no es así en la mayoría de los casos, ya que los productores de bases de datos y repertorios bibliográficos suelen indexar en lenguaje libre o semilibre, lo que resulta mucho más fácil y económico para sus fines. Por tanto, al seleccionar **los descriptores para la búsqueda**, habrá que tener en cuenta *todas las posibles formas de expresión de un concepto en la indexación de los documentos* (sinónimos, conceptos más generales y más específicos, etc.) ya que de otro modo podrían perderse cantidades importantes de información. Por otra parte, si la selección es demasiado amplia, se obtendrán documentos carentes de interés. De ahí la importancia de preparar adecuadamente el perfil de búsqueda, operación que resulta así la mas importante en el proceso de recuperación de información.

Preparación de perfiles de búsqueda

El perfil de búsqueda parte de una petición de información en lenguaje natural, y consta esencialmente de tres elementos:

* Identificación de los conceptos **(1)**
* Desarrollo y expansión de los conceptos, mediante una colección de términos **(2)**
* Expresión de las relaciones entre los términos, mediante operadores lógicos (AND, OR, NOT) **(3)**

Vamos a referirnos a la elaboración de un perfil para una búsqueda temática)también es posible realizar búsquedas por autores, revistas, idiomas, etc.). Tomemos un ejemplo sencillo:

** Petición de información sobre “pinturas anticorrosivas de alto contenido en cinc”**

(1) En esta petición identificamos como conceptos:

- pintura
- corrosión y su prevención
- cinc

(2) La segunda fase, o expansión de conceptos, es imprescindible debido a que, para efectuar la búsqueda, el ordenador compara los distintos términos del perfil con los que contienen los modelos de búsqueda de los documentos carácter a carácter y sólo da como aciertos los que coinciden exactamente. Así, un documento sobre "recubrimientos anticorrosivos a base de cinc" no se registraría como acierto. Es, pues, necesario desarrollar cada concepto teniendo en cuenta los sinónimos, palabras más genéricas, más específicas, etc., para lo cual será apreciable la ayuda de un *thesaurus* (conjunto de términos de una bases de datos). En nuestro ejemplo, y sin ser muy exhaustivos, el desarrollo de los conceptos conduciría al siguiente cuadro:

A	B	C
Pintura	Corrosión	Cinc
Recubrimiento	Anticorrosivos	Zinc
Barniz	Degradación	Zn
Laca	Desgaste	
Esmalte	Incrustación	

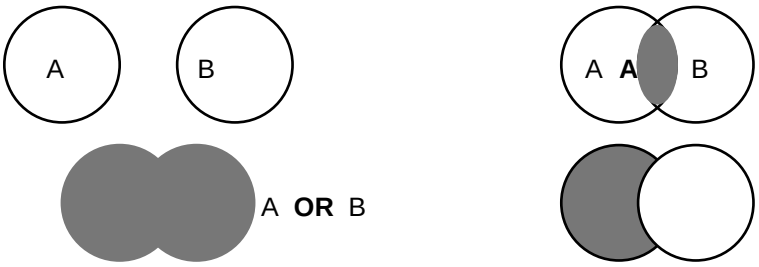
Como se puede ver, no sólo se tienen en cuenta palabras relacionadas sino también las diversas formas ortográficas. En este sentido, ocurre con frecuencia que, para agotar todas las posibilidades, es preciso considerar palabras derivadas de la que se ha utilizado como término. Por ejemplo, si se ha elegido como término "polimerización" será frecuente que interesen también términos como "polímero", "polimerizado", "copolimero", etc. Para tenerlos en cuenta se hace uso del **truncado**, artificio mediante el cual el ordenador considerará aciertos a todos los términos que contengan determinados fragmentos de palabras, sean cualesquiera las letras que se hallen antes o después de los mismos.

Existen tres tipos de *truncado*:

- De **sufijo**: polimer*, cubriría polimer-o, polimer-os, polimer-izado, polimer-ización,…
- De **prefijo**: *polimero, cubriría polímero, co-polimero, homo-polimero,…
- De **infijo**: *polimer*, cubriría todas las posibilidades apuntadas.

De todas formas no se debe abusar del truncado porque puede conducir a la recuperación de muchos documentos irrelevantes.

(3) La tercera fase, relaciona los términos y conceptos mediante los operadores lógicos.



* El operador **OR** permite obtener un nuevo conjunto formado por los documentos que contienen indistintamente el término A o el B (*operación de unión*)

* El *operador de intersección*, **AND**, permite obtener un conjunto formado por los documentos que contienen simultáneamente los términos A y B.

* El operador de *exclusión*, **NOT**, permite formar conjuntos de documentos que contienen el término A, pero no el B.

En general, se relacionarán con la lógica OR los términos correspondientes a un mismo concepto; con la lógica AND los conceptos que deben estar presentes simultáneamente y con la lógica NOT, aquellos que se desee excluir. En nuestro ejemplo, el perfil resultante sería:

(pintura OR recubrimiento OR barniz OR laca OR esmalte)
AND
(corrosión OR anticorrosivos OR degradación OR desgaste OR incrustación)
AND
(cinc OR zinc OR Zn)

Aquí no hemos utilizado el operador NOT. Si se desea, por ejemplo, recuperar información sobre “pinturas anticorrosivas, excepto las que contengan cinc”, bastaría sustituir, en el perfil anterior, el último operador AND por el operador NOT.

Una vez preparado el perfil de búsqueda, se efectuará la comparación con los modelos de búsqueda de los documentos, para obtener los que se ajustan al perfil. Finalmente habrá que comprobar si los documentos obtenidos satisfacen los requisitos del peticionario, es decir, la **relevancia** de dichos documentos para la búsqueda solicitada. Muy a menudo, la información que en verdad se necesita no coincide exactamente con lo que se pide. Ello se debe, normalmente, a que el usuario no conoce con precisión sus necesidades, o no es capaz de expresarlas de forma adecuada.