ALGORITMO DE FILTRADO MULTI-TÉRMINO PARA LA OBTENCIÓN DE RELACIONES JERÁRQUICAS EN LA CONSTRUCCIÓN AUTOMÁTICA DE UN TESAURO

Velasco, I. Díaz, J. Lloréns, A. de Amescua
Departamento de Informática. Universidad Carlos III de Madrid. Correo electrónico: llorens@inf.uc3m.es
V. Martínez
Departamento de Inteligencia Artificial, Facultad de Informática. Universidad Politécnica de Madrid.

etc.) que utiliza como repositorio una estructura basada en los tesauros documentales. Una de las fases indicadas en esta metodología para la generación del tesauro es la adquisición de conceptos, que utiliza técnicas de filtrado de información estadísticas. En este trabajo se presentan modificaciones a estas técnicas de filtrado para proporcionar términos compuestos.

Palabras Clave: análisis de dominios, filtrado, relaciones jerárquicas, indización, ciencias de la información, palabra compuesta, tesauro.

Abstract: The new techniques of domain analysis (thematic, etc.) supply methods to create repositories or sets of structured information. A specific methodology to automatically generate domains (thematic, etc.) uses as repository a structure based on documental tbesauri. One of the stages described in this metbodology to generate a tbesaurus is the acquisition of concepts using statistical techniques to filter information. This paper presents modifications to these filtering techniques to obtain composite terms.

Key words: domain analysis, filtering, hierarchical relations, indexing, information science, composite Word, thesaurus.

1 Introducción

El problema de la construcción automática de tesauros (1, 2, 3) ha traído en jaque a varias generaciones de investigadores. Su construcción se realiza en la actualidad de forma manual; sólo con unas pocas aproximaciones que intentan la automatización del proceso (4). Esta supondría el ahorro tanto de gran parte del personal dedicado a la construcción manual como del tiempo necesario para ello, que en el caso de tesauros construidos manualmente, y dependiendo del área de conocimiento que abarque el tesauro, oscila en torno a un año (5).

Por ello, la riqueza semántica que proporcionan estas estructuras de conocimiento es desaprovechada en otras áreas de investigación, ya que, por ejemplo, en las nuevas técnicas de análisis de dominios (6,7, 8) no se propone la utilización de tesauros como repositorio inteligente para gestionar la información de un dominio o área de conocimiento, salvo en (5), donde se utiliza como repositorio el Tesauro de Software, definido en (9, 10). En este caso, el análisis de dominios está principalmente enfocado a la reutilización de software pero puede extrapolarse su uso a cualquier tipo de información, no necesariamente software. Existen multitud de conexiones entre el análisis de dominios y las ciencias documentales, tal como se comenta en (4, 11, 12). En (5) se presenta una metodología para la construcción automática de dominios utilizando como repositorio el Tesauro de Software; es decir, que en el fondo se define una metodología para la construcción automática de tesauros. Esta metodología divide el proceso global en seis fases, las cuales son similares a las que se realizarían en un correcto proceso de construcción manual de un tesauro:
-Identificación y definición del dominio.
- Obtención del corpus.
- Identificación y adquisición de componentes. Se intentan identificar características comunes (entre conceptos, operaciones, eventos, relaciones 0 estructuras complejas compuestas de alguna de las anteriores), variaciones que ayuden a encapsular y parametrizar, combinaciones que sugieran patrones o comportamientos y trade-offs que posibiliten descomposiciones de módulos o arquitecturas para satisfacer conjuntos incompatibles de requisitos encontrados en el dominio.
- Indización o referenciación de la información.
- Creación de relaciones entre componentes.
- Contraste o validación del dominio.

La adquisición de información se realiza a través de la documentación que forma el corpus representativo del área de conocimiento mediante técnicas bilbliométricas definidas en (4).

No toda la información que aparece en el corpus es válida. En el caso de no disponer ni siquiera de un diccionario es complejo poder determinar de forma automática los descriptores representativos del dominio. Se utilizan procesos de filtrado de información para obtener rápidamente descriptores, principalmente sustantivos, sin necesidad de conocimiento semántico.

En este artículo se va a prestar principal atención a esta fase del proceso, la utilización de las técnicas de filtrado para la adquisición de Pos descriptores que formaran el tesauro y la posterior y automática deducción de las primeras relaciones entre los descriptores a partir de los descriptores compuestos obtenidos en el proceso de filtrado. Los descriptores compuestos se obtienen mediante la modificación de conocidas técnicas de filtrado de información.

2 Aproximación estadística a las técnicas de filtrado

Como se ha comentado en el apartado anterior, es muy interesante realizar filtrados en la indización, pudiendo realizarse estos procesos antes de, o durante el proceso de indización. El tiempo de indización se reduce considerablemente. Posteriormente, a la hora de buscar relaciones entre los descriptores para formar la jerarquía del tesauro, es necesario que el número de éstos sea reducido ya que las técnicas estadísticas y de redes neuronales que proporcionan estas relaciones trabajan con un conjunto limitado de elementos.

Algunos autores realizan un filtrado manual pero esto impide la idea de construcción automática de tesauros ya que, en este caso,

la construcción sería manual. Las distintas técnicas que se han analizado son capaces de discriminar entre los términos que consideran representativos de un texto y los que consideran sin importancia. En (13, 5) se han desarrollado dos algoritmos diferentes, el primero IDF (14), basado en frecuencias estadísticas de términos y sus apariciones en los distintos documentos que forman el corpus, y el segundo, n-grams (15), que observa las frecuencias estadísticas de cadenas de caracteres; trabajando con una longitud fija para las cadenas. Este último tipo de filtrado presenta la posibilidad también de obtener descriptores compuestos.

En este artículo se presentan las modificaciones y resultados relativos al estudio del primer tipo de filtrado, el filtrado IDF y la ley de Zipf (16), asociada a este tipo de filtrado.

3 Método IDF

IDF hace referencia a las siglas de indización estadística de términos por frecuencias (14, 17). Esta técnica de filtrado está basada en la ley de Zipf (16), que establece que las palabras con mayor frecuencia absoluta son las palabras vacías, mientras que las más infrecuentes son aquéllas que reflejan el estilo y riqueza del vocabulario del autor. Aquéllas que aparecen en la zona media de la función de distribución de frecuencias son las que representan al documento. El punto, referente a la frecuencia, en torno al cual se encuentran estos términos significativos se llama punto de transición de Goffman (18, 17).

La técnica IDF establece un sistema de pesos en función de la frecuencia relativa de cada término en cada documento. En los casos en los que un término tenga una frecuencia en un documento mayor que la media en el resto de documentos se tomará como descriptor. En el momento en que se tome como descriptor para un documento será considerado como tal en el resto de documentos.

En (5) se han integrado ambas técnicas, aplicando primero la ley de Zipf para el cálculo de la zona de transición y después el método IDF para ponderar los resultados por documentos. Se comentará a continuación la problemática específica de cada método, así como las mejoras introducidas.

Existen dos formas de trabajo para aplicar la ley de Zipf. La primera se aplica calculando la zona de transición documento a documento, y la segunda calcula la zona de transición para el total de documentos que forman el corpus. Cada una de ellas tiene sus ventajas y sus inconvenientes, que han quedado descubiertos después de realizar numerosos ensayos.

Al trabajar documento a documento aparece como principal problema el de la representatividad de la longitud de los documentos. Si los documentos son cortos, según los parámetros de homogeneidad de corpus definidos en (4), la zona de transición suele quedar muy desplazada hacia la derecha, hacia aquellos términos que presentan un número elevado de ocurrencias, con lo que el número de términos vacíos que salvan el proceso de filtrado y se incluyen como descriptores es mayor del deseado. La figura 1 presenta este caso. Ejemplos significativos (tomando valores medios) del valor del punto de transición de Goffman en función del número de términos que aparecen en cada documento se muestran en la tabla 1.

Figura 1
Función de distribución con documentos de longitud corta

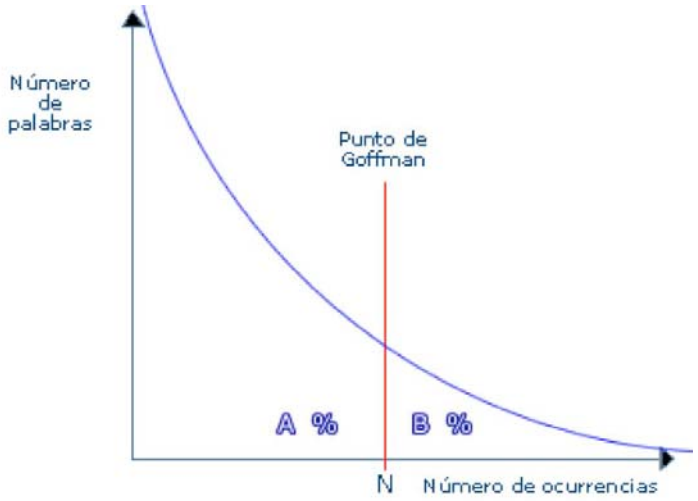
Tabla 1
Estadísticas del punto de transición de Goffman

Términos	Punto de Goffman
100	6
500	15
1000	22
2000	27
3000	32
4000	36
5000	39
10000	47
50000	67

Siendo n el valor del punto de transición de Goffman, los valores de frecuencia aceptados para los descriptores pertenecen al intervalo [n-d, n+d]. El valor del parámetro d se definirá posteriormente. En este caso específico, en el que los documentos son cortos y se considera la zona de transición desplazada hacia los términos con altos valores de ocurrencias, se define un factor de desplazamiento (f), de tal forma que el nuevo punto de Goffman (n2) sea igual a n-f. De esta forma los valores filtrados pertenecerían al intervalo [n-f-d, n-f+d].

Existe otro remedio a este desplazamiento, y consiste en definir, a priori, porcentajes mínimos de ocurrencias a ambos lados del punto de transición. Los estudios realizados en este trabajo muestran la tendencia de que, a ambos lados del punto de Gman, la suma de las ocurrencias de los términos que pertenecen a cada lado es similar (5). La definición de estos porcentajes se muestra en la figura 2. No es necesario que ambos porcentajes mínimos sean iguales.

Figura 2
Porcentaje de información según el punto de transición de Goffman



Otros parámetros de funcionamiento se refieren al número de términos filtrados, concretamente a este valor d, que hace que los términos aceptados tengan frecuencias de aparición próximas al punto de transición, a ambos lados de éste, formando la zona de transición. Los límites del intervalo que se forma son equidistantes del punto de transición de Goffman.

En (5) aparecen tres formas de definir el valor de este parámetro:

- Mediante un rango fijo, tomando un valor absoluto. Tiene el problema de que este valor no puede definirse igual para todo el conjunto de documentos que se filtra. Podría, a priori, establecerse mediante una función que tomase como parámetros el valor del punto de transición de Goffman y el número de palabras distintas del documento.
- Mediante la elección de un número total de términos filtrados por documento. Debe definirse uno diferente para cada documento, ya que no todos los documentos tienen la misma representatividad, ni el mismo número de términos. Este valor total de términos filtrados puede determinarse mediante una función similar a la del punto anterior. El valor de desplazamiento se establece por

tanteo, a partir del punto de Goffman.
- Mediante el establecimiento de un porcentaje de términos filtrados por documento. A partir del número de términos diferentes y en función, de heurísticas obtenidas en la observación de estudios sobre filtrados puede establecerse una función que determine el porcentaje óptimo de términos a filtrar, siempre a partir del punto de transición de Goffman.

Estudios realizados en (5) sobre el número de sustantivos que pueden considerarse relevantes en un documento, a partir del número de términos distintos que aparecen, indican que este número oscila entre el 15 % y el 35 %, de los cuales, aproximadamente, sólo el 15 % es significativo del dominio, centrado en la zona de transición.

Las dos primeras opciones son las que aparecen en los estudios publicados anteriormente (18, 17), pero en este trabajo se ha seleccionado también la tercera opción, al considerarse la más completa, integrándose en ella las otras dos.

Pese a todo, no es sencillo definir un número fijo o porcentaje de términos filtrados por documento, si se filtra documento a documento, porque existen dos factores que lo complican. Por un lado, la obtención de descriptores compuestos a partir de los descriptores simples hace variar el número de descriptores definitivos filtrados. Por otro lado, si un descriptor es filtrado como tal en un documento pasa automáticamente a serlo en el resto a efectos de referenciación. Si un descriptor es significativo y reprsentativo de un dominio lo será en todos los documentos relativos a ese dominio en los que aparezca. Para un descriptor dado, nada asegura que el proceso de filtrado lo haya considerado como tal en todos los documentos. Así, la función que establece el porcentaje de términos deseados debe tener en cuenta estos aspectos para tomar valores más bajos de los necesarios.

Otro problema importante es el del filtrado de términos compuestos. Al ser documentos relativamente cortos, la probabilidad de que cadenas de términos se repitan un número suficiente y representativo de veces en un único documento es baja.

En el caso de realizar el proceso de filtrado globalmente para el conjunto de todos los documentos desaparece alguno de los problemas anteriores pero surgen otros.

No existe el problema de definir a priori el total de términos deseado, salvo para el caso de los descriptores compuestos, que incrementan este número global y cuyo número no puede conocerse a priori ya que depende enormemente de cada dominio en particular.

Se soluciona también la escasez de términos compuestos filtrados, ya que al unirse todos los documentos en uno, la posibilidad de aumentar la frecuencia global de cada cadena de términos aumenta, hasta llegar en los casos significativos hasta la zona de transición.

La zona de transición sufre un desplazamiento respecto a los valores ideales pero, en este caso, hacia la izquierda, hacia los valores de ocurrencias más bajos, tal como muestra la figura 3, pero de forma más atenuada que en el caso en el que se desplazaba hacia la derecha.

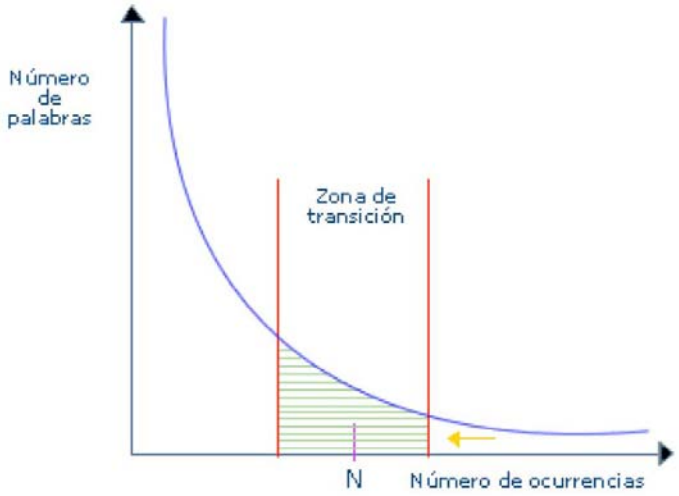
4 Fitrado multi-término

Se ha creado en este trabajo una modificación a la ley de Zipf para el tratamiento de palabras compuestas. La ley de Zipf no contemplaba filtrar por términos compuestos y la novedad introducida complementa el filtrado con éxito. De esta forma puede utilizarse este tipo de filtrado para obtener los descriptores ya que, en la mayoría de los tesauros, existen numerosos descriptores formados por más de un término.

El algoritmo global creado tiene los siguientes pasos:

1 . Aplicar la cuenta de apariciones de términos simples en documentos.

Figura 3
Zona de transición desplazada hacia la izquierda



2. Aplicar la fórmula de Zipf (con IDF en su caso) para el calculo de la zona de transición.
3. El numero de términos filtrados depende de si se utiliza el rango para la desviación de la zona de transición, de si se hace una transformación de la zona de transición o de si existe un número fijo de términos a filtrar.
4. Se toma, en todos los casos, el valor mínimo del intervalo de la zona de transición como valor de corte para la búsqueda de términos compuestos.
5. Se realiza de nuevo la cuenta de apariciones añadiendo un término más a la última cuenta realizada (la primera vez será buscar compuestos de 2 palabras), partiendo sólo de aquellos términos que en el caso anterior superasen el valor mínimo de corte.
6. Aquellos términos compuestos que se sitúen en la zona de transición son considerados significativos.
7. Se calculan de nuevo todos los rangos para observar si existen nuevas variaciones en la zona de transición.
8. Se repiten los pasos 5, 6 y 7 hasta que ningún término compuesto supere el valor mínimo.
9. Se decrementan las ocurrencias globales en aquellos términos (compuestos o no) que estén incluidos en uno de longitud mayor, yendo desde n hasta 1, tomando sólo los más significativos. Esto asegura que para un término dado, forme parte de uno compuesto o no, la suma de sus apariciones será idéntica al caso del filtrado sin palabra compuesta.
10. Tras haber terminado el filtrado deben introducirse igualmente como descriptores aquellos sustantivos (si se conocen) que fórmen parte de un descriptor compuesto, en el caso de que no hubiesen sido ya proporcionados por el filtrado. Esto se hace de forma independiente del número de ocurrencias que tengan estos descriptores simples, ya que serán utilizados para la creación de las primeras jerarquías.

Se presenta a continuación un ejemplo de obtención de palabras compuestas, a partir de la aplicación de la ley de Zipf para un documento. Este caso es fácilmente extrapolable a un número cualquiera de documentos y al algoritmo IDF.

Imagínese un documento filtrado de acuerdo con el método anterior y cuya zona de transición, calculada a partir de los términos de frecuencia = 1, incluye aquellos términos cuyas apariciones están incluidas en el intervalo [20, 36].

De acuerdo con los siguientes valores, tomados parcialmente del conjunto de términos, sólo el término «economía» se considera como descriptor, al encontrarse dentro de la zona de transición.

	apariciones
de	1408
defensa	74
del	511
departamento	42
economía	36
ministerio	65

Tomando las composiciones de cadenas de términos de dos en dos se obtienen los siguientes valores:

	apariciones
de defensa	44
de economía	34

de ministerio	21
departamento de	41
economía del	19
ministerio de	59

Se aplica sucesivamente el algoritmo hasta llegar a los siguientes valores, obtenidos con cadenas de 3 y 7 términos:

	apariciones
departamento de economía	33
ministerio de defensa	34
departamento de economía y ministerio de defensa	14

Este último término («departamento de economía del ministerio de defensa») no se tiene en cuenta al no pertenecer a la zona de transición.

Globalmente, los resultados, una vez actualizados los cálculos, son éstos:

	apariciones
de	1341
defensa	42
del	511
departamento	9
departamento de economía	33
departamento de economía del ministerio de defensa	0
economía	3
ministerio	31
ministerio de defensa	34

La suma de las apariciones de cada término simple permanece constante. De los resultados que se presentan en esta última tabla se deduce que los posibles descriptores son: «departamento de economía», «ministerio» y «ministerio de defensa, al ser los únicos que se incluyen en la zona de transición.

Con la variación proporcionada en el paso 10 del algoritmo de filtrado por palabra compuesta debieran también añadirse aquellos términos (sustantivos) que no hayan sido filtrados y que compongan los términos filtrados compuestos. En este caso debieran añadirse los términos «defensa», «departamento» y «economía».

5 Adquisición de interrelaciones por medio del filtrado multi-término

El sistema creado permite sugerir jerarquías por medio de las palabras compuestas, realizándose posteriormente un contraste con la obtención de relaciones entre descriptores En el ejemplo presentado anteriormente, se considera al descriptor «ministerio» como genérico del descriptor «ministerio de defensa».

Dentro del proceso global de construcción automática de tesauros se desea encotrar los siguientes tipos de relaciones (1, 9):

- Relaciones de equivalencia: sinónimos.
- Asociaciones permanentes: relacionados.
- Asociaciones circunstanciales: relacionados.
- Jerárquicas: genéricos y específicos.
- Asociaciones temáticas: cuasirrelacionados.

Se proponen clasificadores estadísticos y neuronales para obtener los distintos tipos de relaciones. Estos clasificadores trabajan principalmente con técnicas de cowording (19) y clustering. El proceso que debe realizarse para obtener la jerarquía final aparece definido en (5). Resultados significativos aparecen en (4, 13, 5).

Se llega al proceso de búsqueda de relaciones entre componentes con un apoyo añadido, que consiste en las jerarquías y relaciones obtenidas a partir del filtrado de términos por palabra compuesta. Este proceso proporciona las primeras jerarquías para cada descriptor compuesto y los descriptores simples de los que se compone (que suelen existir).

Por ejemplo, si se dispone de los descriptores «defensa», «ministerio» y «ministerio de defensa», se tienen ya dos (o incluso tres) relaciones entre componentes, reflejadas en la figura 4.

FIGURA 4
Ejemplo de relaciones provenientes del filtrado

Para el resto de relaciones puede decirse que se efectúan tres tipos globales de cálculos: un proceso que proporciona porcentajes de relación entre dos descriptores dados (representado por el método cienciométrico de Chen); un proceso de clusterización, que agrupa descriptores en función de similitud entre ellos (representados por el resto de clasificadores estudiados); y un proceso de extracción de componentes principales, que puede considerarse realizado también por los clasificadores estadísticos y neuronales mediante el calculo de los centroides de cada cluster, al que se añaden otras técnicas de extracción de componentes principales (5).

Existen globalmente dos procesos, que deben integrarse posteriormente, pero que trabajan en paralelo: el método de Chen por un lado y el proceso de clusterización, que incluye al resto de técnicas. por otro.

Se ha intentado definir un orden lógico de búsqueda de estas relaciones, independientemente de cada proceso, ya que los resultados serán posteriormente contrastados e integrados, comenzando por la búsqueda temática o de cuasirrelacionados, la búsqueda de equivalencias o sinonimias, la búsqueda de relaciones (permanentes y circunstanciales), para terminar finalmente con el proceso más complejo, el de generación de relaciones semánticas jerárquicas.

6 Experimentos

La experimentación presentada está englobada dentro del desarrollo de una tesis doctoral (5), contrastada con resultados aparecidos en (4) y (13). Esta investigación ha recibido financiación económica de programas nacionales (20) e internacionales (21, 22).

En este trabajo se presentan resultados relativos al filtrado de información, con la consiguiente adquisición de descriptores y a las relaciones obtenidas por medio del filtrado por palabra compuesta.

El proceso de adquisición de componentes está basado en los dos procesos de filtrado apuntados en el apartado 2. De ellos recibe mayor peso el filtrado estadístico IDF, basado en Ia ley de Zipf, ya que el filtrado por cadenas de caracteres o filtrado n-grams conlleva el problema de la definición del background global documental, y en esta experimentación se lo ha considerado como de menor fiabilidad, de tal forma que los resultados básicos del proceso de adquisición de componentes provienen del filtrado IDF, y son complementados por los resultados experimentales del filtrado n-grams. En estos momentos, en el grupo de investigación en el que se desarrolló esta tesis prosigue la investigación relativa al filtrado por cadenas. Se presentan a continuación los resultados obtenidos en los procesos de filtrado de información.

Partiendo de un corpus documental relativo al área de conocimiento de biología, compuesto por 69 documentos, se procede a la adquisición de componentes. Mediante el filtrado IDF (completándose los datos con el filtrado n-grams y con la intervención manual del experto), y a partir del total de 97.568 términos (simples), se ha obtenido un conjunto de 938 descriptores, reflejándose en la tabla 2 los cálculos globales.

TABLA 2
Resultados globales del proceso de filtrado

Número de términos filtrados por ambos métodos	692
Número de añadidos manualmente	244
Número total definitivo de descriptores	938

Una vez aplicados a los resultados obtenidos con la ley de Zipf y la fórmula de ponderación JDF se logran los resultados definitivos del proceso de filtrado estadístico. Pese a determinar mediante porcentajes el número de términos que se quiere filtrar, al aparecer los descriptores en distintos documentos varía este porcentaje si existen descriptores que sí son resultado del filtrado en unos casos pero no lo son en otros. En estos últimos deben añadirse a los resultados finales. Es muy complejo determinar a priori los porcentajes teniendo en cuenta este hecho.

Como resultado del filtrado aparece un conjunto de términos que no se considera perteneciente al dominio, siendo en la mayoría de los casos palabras vacías. A la inversa, existe una serie de términos que no supera el proceso de filtrado y que se considera representativo de ésté. Automáticamente no es posible solucionar, de momento, este doble problema, lo que implica la necesidad de la intervención manual del experto.

La posterior inclusión en el proceso de filtrado del módulo específico de obtención de palabras compuestas hace variar el resultado, ya incluido en la tabla anterior.

La aparición de descriptores en documentos oscila entre casos de descriptores que aparecen solamente en un único documento (un total de 49) y un grupo de descriptores que aparecen en más de 10 documentos distintos (el máximo es 15), entre los que se encuentran descriptores más generales como «vida» y «naturaleza».

Los descriptores que presentan, globalmente, un número mayor de ocurrencias son descriptores generales, como biología, que a su vez aparecen en un número mayor de documentos. Los descriptores más específicos pueden también presentar un número alto de ocurrencias pero aparecen en un menor número de documentos. El máximo de apariciones para un descriptor en el total del corpus es 68, y el mínimo es 6, siendo la media aproximadamente igual a 21.

Entre los términos representativos del dominio que fueron introducidos manualmente se encuentran los que figuran en la tabla 3.

La mayoría de estos términos aparecía un número de veces menor al valor inferior de la zona de transición, pero muy cercano a este valor. Igualmente, el número de documentos en los que aparecían tampoco era elevado. En caso contrario, podían haber salvado el proceso de filtrado mediante el documento conjunto del corpus. Es obvio que los términos arriba indicados pertenecerían a un dominio relacionado con la naturaleza.

Palabras vacías y sustantivos que no son significativos del dominio, que fueron resultado del proceso de filtrado y que por tanto han debido ser eliminados manualmente son, por ejemplo, los que figuran en la tabla 4.

Son términos que aparecen en un número elevado de documentos pero que no son filtrados en todos ellos. Recuérdese que si un término se selecciona como descriptor en un documento ya es considerado como tal en todo el corpus.

No se han introducido manualmente palabras compuestas pese al escaso número de éstas obtenido, ya que no existían más combinaciones de palabras compuestas, que siendo relativas al dominio, apareciesen un número significativo de veces. Podría haber sido el caso de los términos compuestos «glóbulo rojo», «hemisferio sur» o «aparato circulatorio». Las palabras compuestas que han superado el proceso de filtrado aparecen en la tabla 5.

Tabla 3
Descriptores introducidos manualmente

Asia	espora	hipófisis	molla	polimerasa
caballo	fanerógama	larvario	nocivo	procariota
cucaraha	garrapata	linfocito	ocelote	renal
desmosoma	gavilán	mielina	peciolo	repollo
díptero	gaviota	miriápodo	piojo	testículo

Tabla 4
Términos no significativos eliminados manualmente

algunos	efectivo	muchas	serie
cerca	estar	paralelo	sobre
claros	forma	prácticas	sufrido
contra	manejos	principal	tales
depender	Manuel	pueden	uso

Tabla 5
Listado de descriptores compuestos filtrados

AMPLIFICACION PROTO, CONTROL BIOLOGICO, ENEMIGOS NATURALES, GLANDULA PINEAL, IN VITRO, INDICE RNA, SELECCION NATURAL
SINDROME DE DOWN, INDICE RNA DNA, CONDICIONES IN VITRO, ENRAIZAMIENTO IN VITRO, CRECIMIENTO EN LARVAS, TRANSPORTE DE AMINOACIDOS, EDAD DE PIEDRA
CRECIMIENTO DE LAS LARVAS

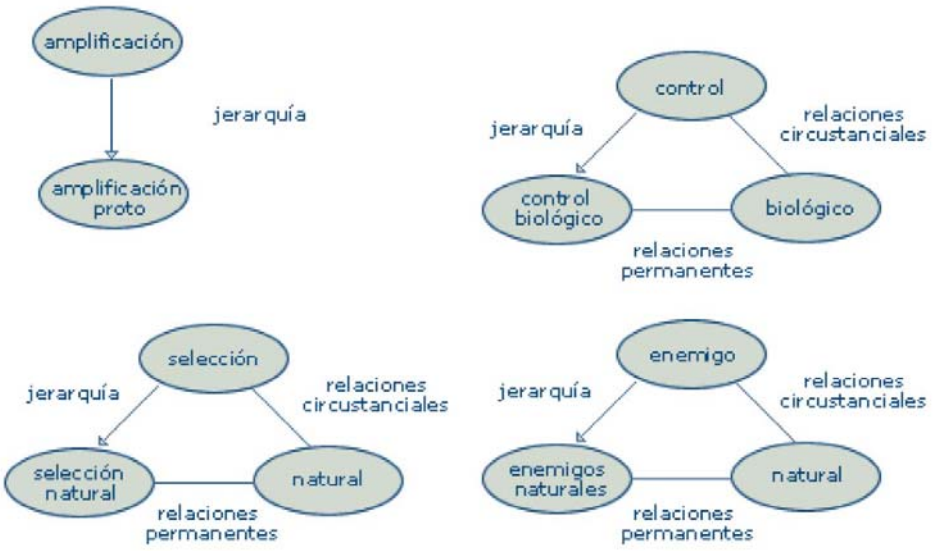
¿Por qué existen tan sólo 15 palabras compuestas como resultado final del proceso? La razón principal que puede deducirse viene derivada de la corta longitud de la mayoría de los documentos, lo que implica que existe baja probabilidad de que exista en un mismo documento varias repeticiones de una cadena de términos. La probabilidad es aún menor si se tienen en cuenta distintos documentos.

En otros estudios parciales realizados, el algoritmo de palabras compuestas funciona sin problemas, pero en este caso, debido a que no se tiene disponibilidad de documentos fáciles de interpretar, se prefiere incluir este ejemplo, únicamente con 15 términos, poco más del 1,5 % del total, para ilustrar la técnica, pero c o n la ventaja principal de poder interpretar los resultados.

Al proponer la modificación del filtrado estadístico IDF para la obtención de descriptores compuestos se estableció que las primeras relaciones jerárquicas se construirían a partir de este proceso, obteniéndose también relaciones permanentes y circunstanciales.

Las relaciones ya obtenidas mediante este proceso son presentadas en las figuras 5, 6, 7 y 8:

FIGURA 5
Relaciones obtenidas por el filtrado mediante palabra compuesta (1)



Recuérdese que, dado un descriptor compuesto formado por dos descriptores simples más un número indeterminado de partículas de unión, se establecían tres relaciones: una relación jerárquica entre el primer descriptor simple y el descriptor, compuesto, una asociación permanente entre el segundo descriptor simple y el descriptor compuesto y una asociación circunstancial entre los dos descriptores simples.

Se obtienen, por tanto, como relaciones de partida, 14 relaciones jerárquicas, 14 asociaciones permanentes y 12 asociaciones circunstanciales.

FIGURA 6
Relaciones obtenidas por el filtrado mediante palabra compuesta (2)

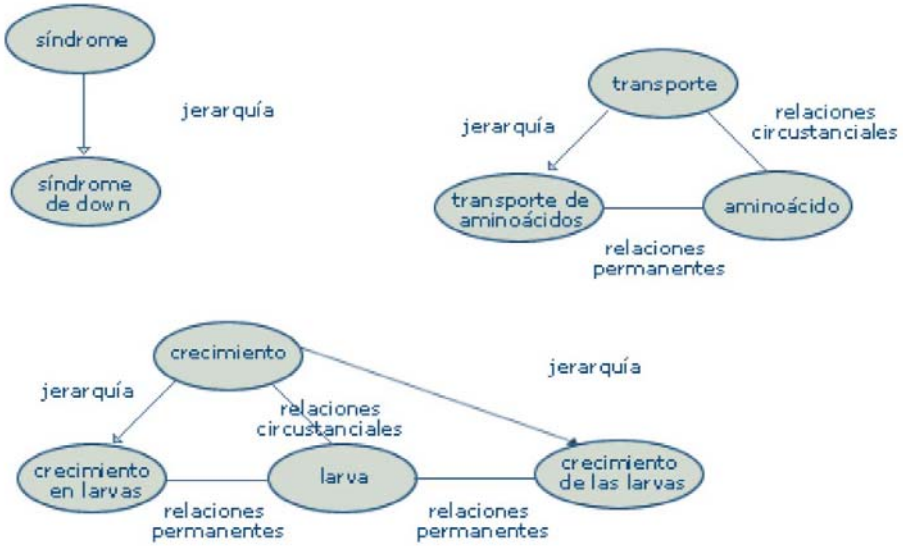


FIGURA 7
Relaciones obtenidas por el filtrado mediante palabra compuesta (3)

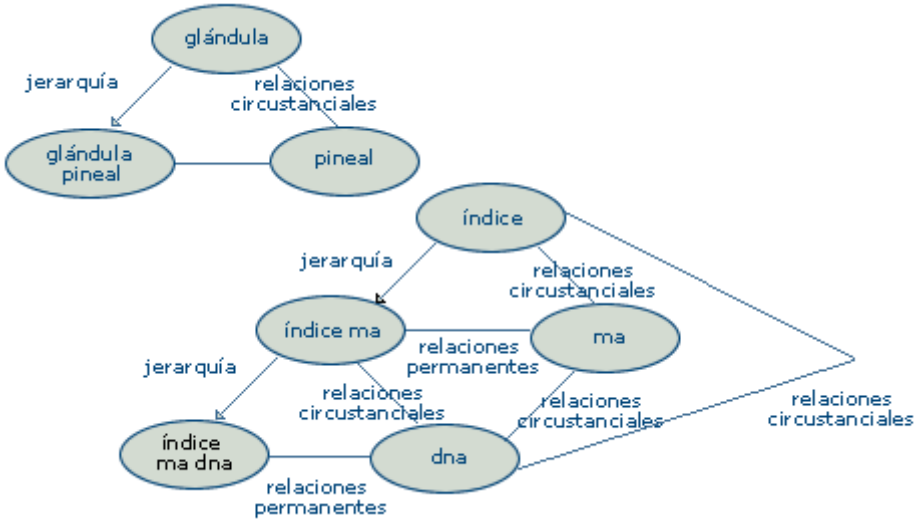


FIGURA 8
Relaciones obtenidas por el filtrado mediante palabra compuesta (4)

7 Conclusiones

Como conclusiones principales que pueden deducirse de este trabajo se destacan las siguientes:

- Se ha definido una metodología específica de análisis de dominios que emplea como repositorio una estructura, el tesauro de software, basada en los tesauros de descriptores de las ciencias documentales. Esta metodología propone la generación automática de representaciones de un dominio y puede, por lo tanto, utilizarse en la construcción automática de tesauros.
- Se utilizan técnicas estadísticas de filtrado en el proceso de adquisición de componentes.
- Se propone la modificación de estos procesos con una doble intención: desviar la zona de transición en función de la longitud de cada documento y proporcionar un algoritmo para aplicar la ley de Zipf a la obtención de descriptores compuestos.

8 Referencias

1. AITCHISON, J. Thesaurus construction: a practica1 manual. ASLIB, 1987.
2. ISO 2788-1986 (E). Guidelines for the Establishement and Development of Monolingual Thesauri. 2.” edición, 11-15 udc, 025.48, ISO 2788, 1986.
3. VAN SLYPE, G. Les Langages d’lndexation: Conception, Construction et Utilisation dans les Systèmes Documentaires. París, Les Editions d’organisation, 1991.
4. LLORÉNS J.; VELASCO, M.; AMESCUA, A.; MOREIRO, J. A.; MARTÍNEZ, V. Automatic Domain Analysis using Thesaums Structures. Aceptado para publicación en Journal of the Ameritan Society for Information Science, 1998.
5. VELASCO, M. Generación Automática de Representaciones de Dominios. Tesis Doctoral. Universidad Politécnica de Madrid, 1998.
6. NEIGHBORS, J. Software Construction using Components. Thesis, Department of Information and Computer Science. University of California, Irvine, 1981.
7. PRIETO-DÍAZ, R. Domain Analysis: An Introduction. ACM Sigsoft. Software Engineering Notes, 1990, vol 15(2).
8. PRIETO-DÍAZ, R.; FRAKES, B. Introduction to Domain Analysis and Domain Engineering. Tutorial 4”’ International Conference on Software Reuse, IEEE Computer Society, 1996.
9. LLORENS, J. Definición de una metodología y una estructura de repositorio orientadas a la reutilización: el Tesauro de Software. Tesis Doctoral, Universidad Carlos III de Madrid, 1996.
10. LLORENS, J.; AMESCUA, A.; VELASCO, M. Software Thesaurus: a Tool for Reusing Software Objects. Proceedings of the Fourth IEEE Assessment on Software Tools. Toronto, Canada, 1996.
11. BEGHTOL, C. Domain Analysis, Literary Warrant, and Consensus: The Case of Fiction Studies. Journal of the American Society for Information Science, 1995, vol 46(l), pp,30- 44.
12. HJORLAND, B.; ALBRECHTSEN, H. Toward a New Horizon in Information Science: Domain-Analysis. Joumal of the American Society for Information Science, 1995, vol46(6), pp 400-425
13. VELASCO, M.; MARTÍNEZ, V.; LLORENS, J.; AMESCUA, A. Automatic Domain Analysis: Generation of Domain Representations. IT-Knows, Austria-Hungría, septiembre, 1998.
14. SALTON, G. Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989.
15. COHEN, J. Highlights: Language and Domain-Independent Automatic Indexing Terms for Abstracting. Journal of the American Society for Znformation Science, ‘1995, vol 46(3).
16. ZIPF, G. K. Human Behaviour and the Principle of Least Effort: An Zntroduction to Human Ecology. Haffner, Nueva York, 1972.
17. MUÑOZ, A. Redes Neuronales para la Organización Automática de Información en Bases Documentales. Tesis Doctoral, Universidad de Salamanca, 1994. ,
18. CLEVELAND, D. B.; CLEVELAND, A. D. Zntroduction to Zndexing ana’ Abstracting. Libraries Unlimited, 1990.
19. CHEN,H.; YIM, T.; FYE, D.; SCHATZ, B. Automatic Thesaurus Generation for an Electronic Community System. Journal of the American Society for Information Science, 1995, vol 46(3).
20 GATOAC. Generación automática de tesauros orientada a las arquitecturas de componentes. Proyecto CICYT, 1997.
21. AUTOSOFT. High Leve1 Software Project Reuse Based on Automatic Domain Generation. Proyecto Esprit 25762, 1998.
22. IMDEX. Integrated System for Multimedia Indexing, Matching and Retrieval. Proyecto Esprit 23011, 1996.