



Evaluación a través del análisis de las palabras asociadas (I)

Aplicación a la evaluación de la investigación científica y técnica

E. Jurado-Alameda, R. Bailón-Moreno
 Dpto. de Ingeniería Química
 Universidad de Granada
 y R. Ruiz-Baños
 Biblioteca de Andalucía (Granada)

1. El conocimiento científico-técnico y los documentos escritos

La investigación científica y técnica se caracteriza, entre otras cosas, por la continua producción de documentos escritos. La difusión de los nuevos conocimientos requiere siempre el soporte documental para llegar al resto de la comunidad científica y en última instancia al conjunto de la sociedad. Según el Centro de Sociología de la Innovación de la Escuela de Minas de París, la actividad científica y técnica puede analizarse según cinco dimensiones principales, que ellos denominan "la rosa de los vientos" de la investigación (Fig. 1). Cada una de estas dimensiones

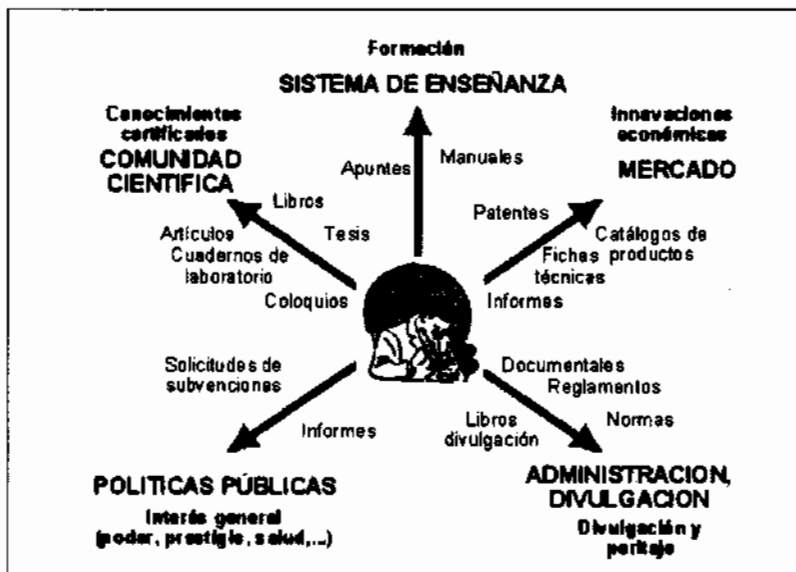
o "puntos cardinales" se dirige a distintos estamentos de la sociedad, produce conocimientos específicos para cada estamento y emplea registros documentales adaptados a cada fin.

a) *Producción de conocimientos certificados.* Son aquellos conocimientos, procedentes de la denominada investigación académica, dirigidos exclusivamente a la comunidad científica y cuya calidad e interés son evaluados por ella misma. Los documentos más usuales son artículos, libros, tesis, comunicaciones a congresos y cuadernos de laboratorio.

b) *Producción de innovaciones con repercusión económica.* Conocimientos orientados a la creación de nuevos productos comerciales

La evaluación de la investigación en ciencia y tecnología mediante el empleo del análisis de las palabras asociadas presenta numerosas ventajas respecto a los análisis clásicos basados en índices de actividad. Se estudia el contenido informativo de los documentos que los científicos y tecnólogos escriben, siendo posible determinar la estructura íntima de las redes tecnocientíficas, así como de sus actores constituyentes: temas de investigación, laboratorios, investigadores, etc. También se puede determinar la posición estratégica de estos actores, sus relaciones y su evolución temporal.

Figura 1.
 "Rosa de los vientos" de la investigación



o nuevos procedimientos industriales. Los resultados se orientan a conseguir ventajas competitivas de las empresas y sus marcas dentro del mercado. La actividad literaria se centra en las patentes, catálogos de productos, fichas técnicas, informes, etcétera.

c) *Investigaciones de interés público.* Acciones financiadas mediante programas públicos para la realización de objetos técnicos complejos (satélites espaciales, armamento, etc.) o bien objetivos más difusos, pero no exentos de interés general (estudios ecológicos, epidemiológicos, sociales, etcétera). No se buscan ventajas competitivas comerciales, sino más bien de carácter político, de bienestar social, de prestigio y de poder. Estas actividades requieren la elaboración de solicitudes de subvenciones y la continua elaboración de informes.

d) *Formación y enseñanza.* Los científicos y los técnicos reelaboran los conocimientos para que puedan ser transmitidos a otros investigadores y profesionales. De esta forma se renueva y se incrementa el potencial humano que genera nuevos conocimientos. Para conseguir esto es necesaria la redacción de manuales y apuntes que se distribuyen en forma de fotocopias.

e) *Divulgación y peritaje.* La sociedad, no experta en ciencia y tecnología, reclama a sus científicos su opinión sobre temas de interés. Por un lado, deben satisfacer la curiosidad de la población en general publicando revistas y libros de divulgación, emitiendo documentales o realizando programas científicos en radio y televisión. O bien entrar en el debate social elaborando leyes, normativas o reglamentos y ayudando, como peritos, a su aplicación efectiva.

Es evidente que la evaluación de la actividad científica pasa necesariamente por el análisis de la producción documental de los científicos y tecnólogos. Además, cuanto más se profundice en el contenido de los documentos, más pertinentes y precisos serán los re-

sultados obtenidos y su consiguiente valoración.

Dos son los enfoques fundamentales a la hora de hacer esta evaluación:

- a) Métodos basados en índices de actividad
- b) Métodos basados en índices de relación:

- Índices relacionales de 1ª generación.
- Índices relacionales de 2ª generación.

2. Índices de actividad

Consisten en determinar el volumen y el impacto de la actividad científica. Se parte de la suposición de que la separación entre disciplinas, temas de investigación e instituciones es clara y está definida de forma aproximadamente constante a lo largo del tiempo. Por ejemplo, el campo de la química de los tensioactivos es perfectamente conocido y se diferencia nítidamente del campo de la toxicología. Es decir, se presupone que se sabe *a priori* cuáles son los documentos, las revistas, los autores o los centros de investigación pertinentes sin hacer un sondeo previo sobre la adecuación de esta supuesta pertinencia.

La ciencia se considera según un modelo análogo a los modelos económicos del tipo entrada/salida (*input/output* en la literatura anglo-

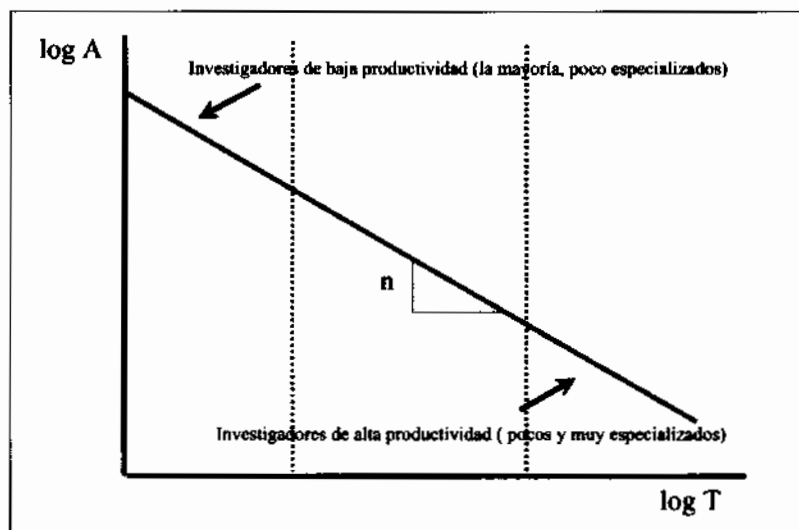
sajona), llevándose a cabo una cuantificación del tipo coste-beneficio o inversión-resultado (Moed, 1984; Martín *et al.*, 1983; Sancho, 1990; Callon *et al.*, 1995). Se tiende a identificar cantidad con calidad. Así, un trabajo es de mayor calidad cuantas más veces haya sido citado. O un grupo de investigación es más rentable si la cantidad de publicaciones por unidad monetaria invertida es superior al de otros grupos.

El índice de actividad más sencillo y más usado es el cómputo de publicaciones en revistas académicas. Se aplica tanto a investigadores, como laboratorios o incluso países. Nos da indicaciones valiosas sobre aspectos como el dinamismo (según crezca, decrezca o se mantenga constante el número total), o la productividad. La ley que suele expresar mejor la distribución de productividades es la de Lotka (similar a la de Pareto en la economía)(Fig. 2):

$$A = \frac{A_j}{T^n}$$

Donde A es el número de autores que publican T trabajos en un determinado tiempo en una disciplina dada. A_j es el conjunto de autores que publican un solo trabajo y n un valor empírico que suele ser del orden de 2 (Lotka, 1926). En este caso, los índices de calidad son T (mayor o menor productividad de cada investigador) y el ex-

Figura 2.
Ley de Lotka



ponente n, que cuanto menor sea, mayor productividad media se asigna al conjunto de investigadores (o laboratorios) estudiados.

Otro índice de actividad muy extendido es el de la productividad de las revistas científicas, regido por la ley de Bradford. Según esta ley empírica, si contamos el número de artículos sobre una temática dada, encontramos que las revistas que los publican se agrupan fundamentalmente en tres zonas (Fig. 3). La primera, denominada núcleo, agrupa a un conjunto muy reducido de revistas que publican mucho sobre el tema en cuestión, considerándose como las altamente especializadas. Hay otro conjunto, mucho más amplio de especialización media y que puede expresarse por la siguiente ecuación:

$$T = a + b \log R$$

Donde R es el número de revistas que publican T trabajos en el campo considerado. Por último, llegando a la denominada inflexión de Groos, a veces aparece un tercer grupo de revistas que publican de vez en cuando algún que otro trabajo sobre la temática en estudio. Son revistas ocasionales. (Bradford, 1934; Ferreiro-Alález *et al*, 1980, Gras, 1967).

A partir del análisis de las citas bibliográficas, se pueden establecer numerosos índices. El más popular es el factor de impacto de una revista. Se define como el cociente entre el número de citas recibidas por la revista durante los dos años posteriores a su publicación, C_{1-2} y el número de artículos publicados, T_0 .

$$FI = \frac{C_{1-2}}{T_0}$$

El factor de impacto se recoge en el Journal Citation Report (JCR), publicado por el ISI y está referido a las revistas recogidas en las bases de datos de esta misma empresa (Sen *et al.*, 1992). Hoy día en España se valoran positivamente sólo aquellos artículos publicados en revistas recogidas por el ISI y con un elevado factor de impacto.

El análisis de citas nos permite también evaluar el grado de envejecimiento u obsolescencia de la literatura científica en un campo determinado. Si analizamos la edad de las referencias que aparecen en un conjunto de artículos científicos o de patentes, se observa que el número de referencias se incrementa desde el año actual hasta alcanzar un valor máximo a los dos o tres años de edad. A partir de aquí las referencias disminuyen exponencialmente hasta anularse virtualmente en un tiempo infinito (Fig. 4). Este decaimiento en la citación se rige por la ley de la utilidad de Brookes (Brookes, 1970; Egghe, 1993; Ruiz-Baños, 1997):

$$U = U_0 a^t$$

Donde U es la utilidad, definida como el tanto por uno de referencias dadas o citas recibidas con la edad t, siendo la base de la función exponencial, a, la tasa o factor de envejecimiento anual. Esta tasa es un parámetro indicativo del grado de obsolescencia y, por tanto, de la utilidad residual que le resta a un cuerpo documental. Relacionado con el factor de envejecimiento está la vida media, h, que representa la edad a la que la utilidad de los documentos se ha reducido a la mitad (o bien la edad en la que recibe la mitad de todas las citas).

$$a^h = \frac{1}{2}$$

El factor de impacto y la tasa de envejecimiento son tan sólo dos de los índices de actividad más representativos de los que emplean citas o referencias. Durante los últimos años se ha usado y en nuestra opinión "abusado" del empleo del análisis de citas para intentar llegar a conclusiones más allá de las pu-

Figura 3. Ley de Bradford

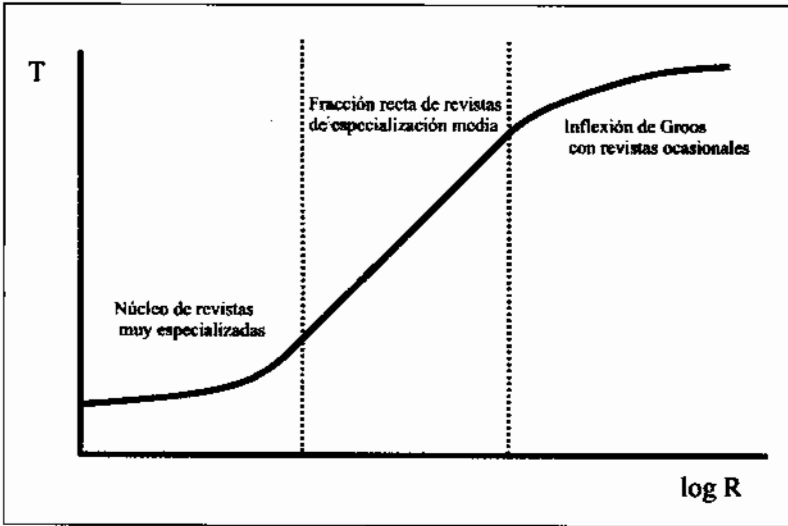
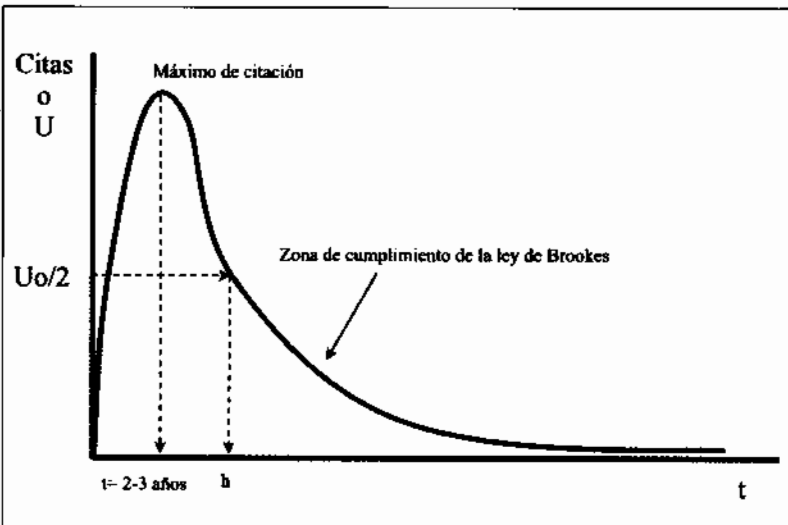


Figura 4. Ley de Brookes



ramente razonables si tenemos en cuenta las serias limitaciones que presenta este tipo de análisis. Desde que McRobert y McRobert publicaron su conocido artículo, el análisis de citas está siendo puesto en entredicho, entre otras, por las siguientes razones (MacRobert *et al.*, 1989; Gaillard, 1989 y Gómez-Caridad, *et al.*, 1996):

1. En la práctica sólo puede emplearse con artículos de revista recogidos por las bases de datos del ISI, que por su propia política de selección, sólo recogen una pequeña parte de toda la producción científica, presentando además un fuerte sesgo hacia revistas biomédicas y aquellas escritas exclusivamente en inglés.

2. Es imposible analizar documentos científicos que no contienen referencias o son éstas muy escasas, tales como muchos informes, fichas técnicas, catálogos, manuales, documentales, libros y artículos de divulgación, y un largo etcétera.

3. Nombres homógrafos: no se puede distinguir entre dos investigadores que tienen el mismo apellido y las mismas iniciales. Por ejemplo, en un análisis de citas poco cuidadoso podríamos encontrar que J. González es un superhombre que produce en un solo año tanto como cien investigadores juntos (probablemente J. González sea realmente 100 investigadores "normales" y no un "supercientífico").

4. En los índices elaborados por el ISI, sólo se retiene el nombre del primer autor. Los demás autores "no existen" a efectos de indicadores.

5. Las mujeres están especialmente discriminadas, ya que en la mayoría de los países cambian de nombre al casarse, por lo que no es posible hacer un seguimiento temporal correcto.

6. Lo más usual es que los autores abusen de la autocitación, es decir, citarse a sí mismos, a veces sin una justificación convincente. También es corriente citar en abundancia "al superior" o al amigo, y ni siquiera nombrar al "enemigo".

La citación está tergiversada por muchos sesgos de esta naturaleza.

7. No hay una relación directa entre citación y calidad. Los artículos metodológicos, por su naturaleza, son más citados que el resto, cuando no son en realidad de mayor o menor importancia. Hay muchos ejemplos de trabajos de una calidad excepcional que, precisamente por ello, son rápidamente absorbidos por el acervo científico común y apenas son citados. Este es el caso del trabajo original de Watson y Crick sobre la estructura del ADN, de importancia capital, y que apenas se cita porque se "sobreentiende".

Estos inconvenientes obligan a ser prudentes en el empleo de los análisis de citas. Es conveniente seguir tres principios de precaución:

a) No emplear este indicador para analizar comportamientos individuales, sino aplicarlo a comunidades lo más amplias posible. Es totalmente erróneo e injusto aplicarlo para evaluar la calidad de un científico aislado (en estos casos, es necesario emplear al número máximo de parámetros y, sobre todo, la opinión de los expertos que sí pueden dar resultados fiables).

b) Sólo debe hacerse comparaciones entre grupos comparables. Es correcto analizar campos científicos iguales entre sí, pero no sacar conclusiones comparando las citas entre áreas no homogéneas. No es válido equiparar la investigación arqueológica del antiguo Egipto, con la investigación aerospacial, por poner un ejemplo extremo.

c) A la hora de llegar a conclusiones, deben proscribirse algunas nociones, y en particular la de calidad. Por supuesto que es lícito hablar de la calidad de un investigador, pero sería peligroso transformar el índice de citas en medida de esta calidad.

3. Indicadores relacionales de 1ª generación

Están basados en el análisis de

co-citas. Dos documentos están "co-citados" o citados conjuntamente cuando aparecen simultáneamente en la lista de referencias de un trabajo científico. Parece razonable que si dos documentos (o dos investigadores, o dos laboratorios) aparecen juntos con frecuencia es porque de alguna forma están relacionados. Por extensión, si consideramos un conjunto elevado de documentos, es posible establecer una red compleja que ponga de manifiesto las relaciones entre documentos, científicos e instituciones (Small, H. *et al.*, 1985). Esta red estará formada por zonas más enlazadas, denominadas grupos ("Clusters" en la literatura anglosajona) que representarán los denominados "Colegios Invisibles" o grupos de investigadores relacionados entre sí y que se citan entre ellos de forma asidua. Los resultados son mapas de la ciencia y de la técnica y que dan una idea global de la estructura y de la dinámica de la Ciencia basada en las citas.

De una forma indirecta podemos conocer qué investigadores y qué instituciones trabajan en la misma disciplina y así establecer paralelismos de gran interés. Se emplean técnicas matemáticas avanzadas extraídas de otras ciencias, fundamentalmente las Ciencias Sociales, Economía y Mercadotecnia: análisis escalar multidimensional (MDS), análisis de componentes principales (PCA), análisis de grupos (*clusters analysis*), dendrogramas, etcétera.

Los índices procedentes del análisis de co-citas aportan más información que los índices de actividad clásicos, pero siguen adoleciendo de los inconvenientes de todos los análisis de citas, tal como se expuso más arriba. Además se les puede achacar que no analizan el contenido del documento, sino su parte más accesoria, que es la lista de las referencias. Este tipo de análisis se olvida del contenido informativo del título, del resumen, de las palabras clave e incluso del cuerpo del documento en sí. Es decir, sólo muestran una visión parcial del conocimiento.

Además presentan un problema

muy grave: ofrecen resultados a "toro pasado", es decir cuando ha transcurrido un número suficiente de años para que se produzca el fenómeno de la citación (como poco 3 y a veces hasta 10 ó 15). Cuando se tienen entonces resultados, la ciencia ha cambiado tanto, que éstos no son válidos para el momento actual. Por esta misma razón, la predicción de comportamientos para el futuro es prácticamente imposible.

4. Índices relacionales de 2ª generación: El análisis de palabras asociadas

El análisis de palabras asociadas se centra en el estudio del contenido verbal del documento, o de cualquiera de sus partes. El análisis puede reducirse por comodidad a las palabras clave o bien extenderse al título, al resumen o a todo el documento. Puede aplicarse a artículos de revista, patentes, libros, tesis, informes, fichas técnicas, catálogos y por extensión y sin limitación alguna a cualquier tipo documental que contenga palabras.

Genera, al igual que al análisis de co-citas, mapas de la ciencia, pero en este caso con una serie de diferencias muy ventajosas, ya que:

1. Lo que se muestra es la es-

tructura y la dinámica del conocimiento científico y técnico en sí: Las teorías emergentes, los métodos empleados, las categorizaciones y las relaciones existentes entre ellas.

2. El análisis se puede realizar en "tiempo real": No hace falta que pase un tiempo, como en las análisis de citas y co-citas, sino que es aplicable en el momento en que se plasma el conocimiento científico en los documentos.

3. Los métodos matemáticos son genuinos y diseñados especialmente para este análisis. Además proporcionan un conjunto de índices propios suficientemente compactos y fiables.

4. El objetivo es determinar la posición estratégica. No se pretende poner de manifiesto la calidad como una cuestión de cantidad, sino que se pretende identificar la situación más ventajosa dentro del mundo científico y técnico: "Cómo ser la persona adecuada en el momento adecuado", para así cosechar un mayor éxito.

5. Permite hacer predicciones a corto y medio plazo. Sobre todo por la ventaja añadida de trabajar con el conocimiento en sí, sin retraso alguno y sabiendo dónde se está situado.

6. Está justificado por dos teo-

orías sociológicas: la teoría actor-red y la teoría de la traducción. El análisis de palabras asociadas no es simplemente un artificio matemático, sino la cuantificación razonada del comportamiento social y antropológico de los científicos.

En la figura 5 se muestra una visión global y simplificada del análisis de las palabras asociadas. Como se indica, parte del contenido de los documentos técnicos y científicos: artículos, patentes, catálogos de productos comerciales, informes, etcétera. Mediante programas informáticos adecuados, tal como Leximappe del Centro de Sociología de la Innovación de la Escuela de Minas de París, se procede al estudio de estos documentos. Los dos aspectos principales son, por una parte, la descripción de la estructura de la ciencia y la técnica y, por otra, su evolución temporal. En cualquier caso, las aplicaciones son múltiples y entre ellas podemos citar:

1. Búsqueda de documentos sobre un tema de investigación concreto en una base de datos, de forma muy eficaz y pertinente.

2. Puede servir como medio para encontrar las fronteras divisorias entre disciplinas como paso previo al empleo de indicadores de actividad clásicos: productividad, factor de impacto, obsolescencia, etcétera.

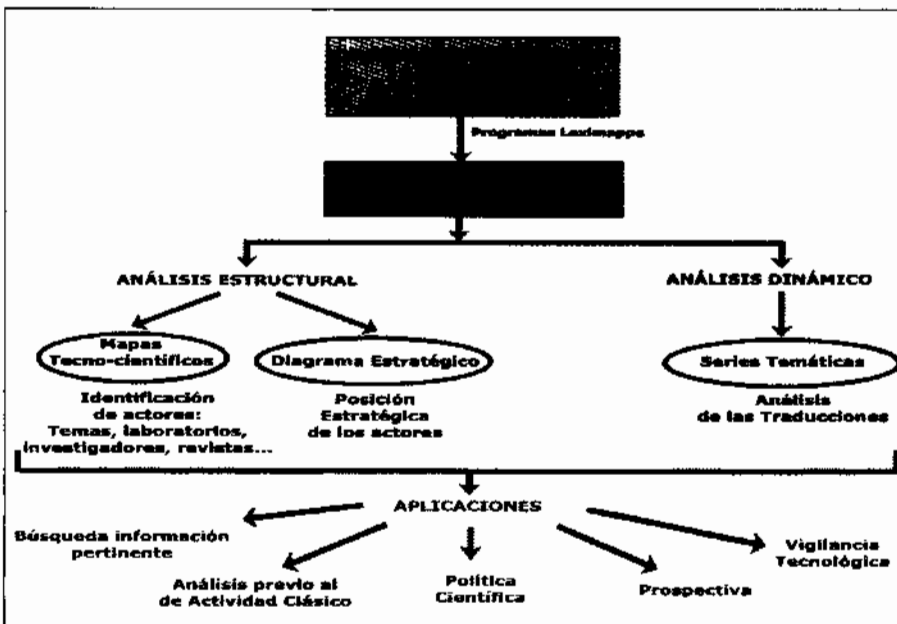
3. Es una herramienta insustituible para una política científica racional. Muy útil para grupos de investigación, universidades, instituciones, o empresas.

4. En la línea con lo anterior, permite realizar prospectivas que pueden ser de interés para decidir próximos pasos que dar en investigaciones futuras.

5. En el caso de empresas industriales, es un medio ideal para su política de vigilancia tecnológica.

En los próximos apartados se explicarán someramente los elementos fundamentales del análisis de palabras asociadas y las teorías que lo sustentan.

Figura 5. Análisis de las palabras asociadas



5. Teoría actor-red

La construcción de esta teoría ha sido llevada a cabo por un equipo multidisciplinar y multiinstitucional de investigadores procedentes, básicamente, de varios países europeos (Callon *et al.*, 1986; Callon, 1989) y que desde finales de los años 70 hasta la actualidad ha sido adoptada por un buen número de laboratorios y centros de investigación de toda Europa y Estados Unidos.

La teoría actor-red considera la ciencia y la técnica como una red sociocognitiva en la que los aspectos sociales y cognitivos del conocimiento o la técnica se entremezclan íntimamente. Ciencia y tecnología no se estudian sólo en sí mismos, sino que son consecuencia de la relación interactiva social y cognitiva de un conjunto de actores. Los pilares básicos de la teoría actor-red son:

1. La existencia de un conjunto de actores.
2. La asociación de estos actores en un entramado o red y
3. La continua transformación de los actores y de la red: proceso denominado traducción.

Los actores se caracterizan porque pueden ser descritos mediante textos escritos, artículos científicos, libros, patentes, etcétera, y consecuentemente mediante palabras. Estos actores así descritos no poseen una definición constante, sino cambiante en función de las palabras que en cada momento los describan. Un aforismo del tipo "dime qué escribes y te diré quién eres" podría muy bien representar este modelo. Además, como la ciencia y la técnica producen textos escritos constantemente modificados y evolucionados, los actores que los producen y las relaciones que los unen, deben cambiar de igual forma.

Según M. Callon *et al.* (1986), "un actor se define como alguien o algo creador de asociaciones: un científico, un texto, un aparato, un concepto, etc.". Hay que subrayar que no se hace distinción alguna

entre actores humanos y no humanos, ya que son definidos por la red sociocognitiva que crean y por el cambio a lo largo del tiempo de las palabras que los forman. "Un científico que escribe un texto es un actor que cambia constantemente: él nunca permanece lo mismo que anteriormente, ya que investiga continuamente con eficacia e incrementa el conocimiento" (Courtil *et al.*, 1994).

El concepto de actor se puede abordar desde dos puntos de vista complementarios, como las dos caras de la misma moneda, actor-entorno y actor-red. Entendemos por actor-entorno el conjunto de entes que forman ese actor. Todas estas entidades no están jerarquizadas, de tal forma que todas son de igual importancia. Si faltara alguna de ellas, es decir, que se encontrara fuera del mundo o entorno del actor, el actor no podría existir.

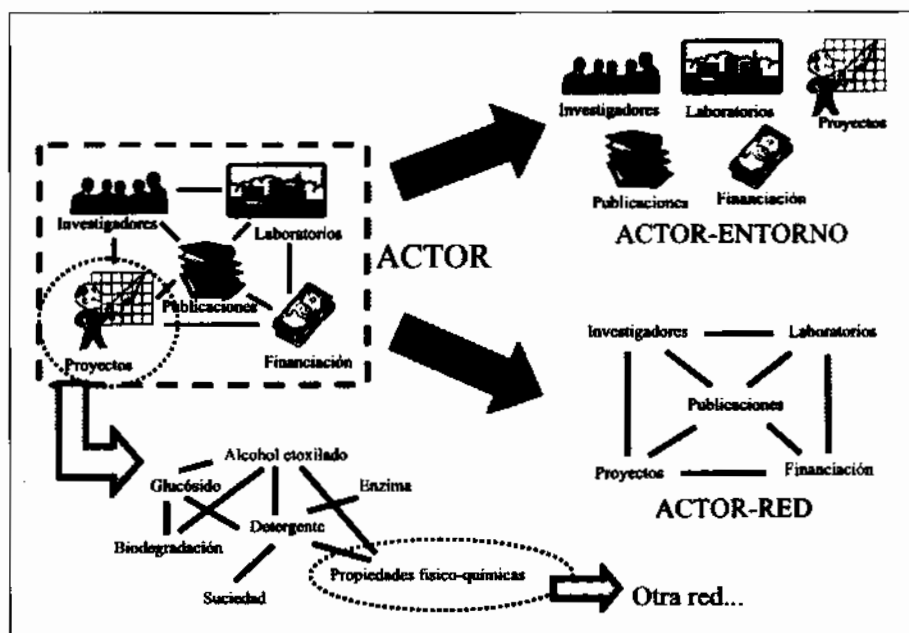
Supongamos un actor dedicado a la investigación de detergentes formado, básicamente, por un conjunto de investigadores, unos laboratorios, proyectos de investigación, publicaciones y la financiación necesaria (Fig. 6). El actor-entorno estaría formado por cada una de las entidades citadas, la forma en que éstas han sido reclutadas y la distribución de los roles o funciones que deban cum-

plir cada una de ellas. Indudablemente, si faltara alguna de estas entidades, por ejemplo, los investigadores, no existiría el actor y sería imposible llevar a cabo los proyectos prefijados. Esto es lo que induce a mantener al mismo nivel cualquier actor, sea humano o no humano. Tan importantes son los investigadores, como la financiación, los proyectos o los medios instrumentales: todos son imprescindibles.

El actor-red representa, sin embargo, la estructura de relaciones entre los entes que componen el actor. Esta estructura es susceptible de cambiar, es más, la esencia del progreso científico consiste en el cambio continuo y en la redefinición permanente de los actores de la red sociocognitiva.

Los entes que forman un actor son a su vez actores, con su entorno y su red. En el ejemplo de la figura 6 se observa que el ente "proyectos" es un actor temático cuya red se especifica debajo. Los entes de este actor son las palabras clave glucósido, alcohol etoxilado, enzima, biodegradación, toxicidad, suciedad, propiedades físico-químicas, que a su vez son actores temáticos. Cada uno de ellos posee su propia red y su propio entorno, expresable con nuevas palabras.

Figura 6. Actores y redes



6. Teoría de la traducción

El cambio continuo de la estructura de las redes científicas y técnicas, en definitiva de los actores-red y los actores-entorno, es debido a lo que el antropólogo de la ciencia Bruno Latour llama traducción (Latour, 1992). La traducción ("translation" en inglés) representa dos ideas: por un lado, consiste en dar nuevas interpretaciones de los intereses de los actores (significado "lingüístico") y, por otro, canalizar a los actores hacia otras direcciones (significado "geométrico"). Es decir, la traducción de los actores implica un cambio en las palabras que definen el actor y un cambio en la posición estratégica dentro de la red.

Supongamos un grupo de investigación dedicado al estudio de las enzimas para su aplicación en la industria alimentaria y cuyo interés es abrir nuevas líneas de investigación en el campo de los productos lácteos. Sea también una empresa que se dedica a comercializar productos lácteos y que necesita desarrollar un nuevo producto que le proporcione ventajas económicas en el mercado. Si ambos actores se acercan y entran a colaborar en un proyecto común, se habrá producido una traducción del tipo "aproximación por convergencia". Se generará un nuevo actor, heredero de los anteriores, cuyas palabras clave serán la conjunción de las palabras clave de los actores precedentes. Este es el significado lingüístico de traducción. A su vez, ambos actores han cambiado su posición dentro de la red acercándose uno al otro. Este es el significado geométrico. En la figura 7 se hace un esquema de lo ocurrido. En el trasfondo de la traducción ha habido una connivencia entre los intereses de ambos actores.

Como se ve, en esta concepción de la dinámica de la ciencia y la técnica, los factores sociales son los que la hacen progresar. Además se rompe el antiguo mito de que la investigación consiste en "descubrir las leyes ocultas de la natura-

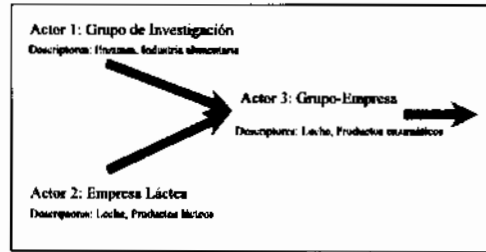


Figura 7.
Ejemplo de traducción de aproximación por convergencia entre un grupo de investigación y una empresa

leza", como si se tratara de un tesoro oculto. Simplemente consiste en hacer uso de ella para generar nuevos conceptos, nuevos productos, nuevas entidades. En definitiva, los actores formados por redes simplemente extienden sus redes cada vez un poco más allá con el objetivo de crear nuevos actores.

Según el concepto de la traducción, Newton "no descubrió" la Ley de la Gravitación Universal, sino que "creó" una teoría que intentaba explicar el porqué caen los objetos. El modelo creado fue simplemente una construcción matemática de carácter puramente humano, una abstracción semántica que puede ofrecer un explicación plausible de lo observado. Como los términos y relaciones creados por Newton fueron convincentes, quedó como "un hecho incontrovertible" que la Ley de la Gravitación Universal era una ley natural, oculta y preexistente, dada a la luz por Newton. No obstante, ya en los primeros tiempos el mundo científico planteó preguntas que Newton jamás pudo constatar de forma contundente y que ponían seriamente en entredicho su teoría:

- ¿Cuál era la naturaleza íntima de la interacción entre las masas? ¿Partículas, ondas,...

- ¿A qué velocidad se transmite la interacción entre las masas? ¿Es finita o infinita? ¿Cuánto tiempo tarda en llegar la atracción de la tierra a la luna o de la tierra a la manzana que cae de un árbol?

- La atracción debería "viajar" a través del universo. ¿Para ello era necesario apoyarse en un soporte material, como una especie de éter de propiedades extraordinarias, o bien viajaba sin soporte a través del vacío?

- ¿Esa interacción duraba eternamente o se gastaba? Si no se gastaba nunca, ¿qué la compensa-

ba eternamente? ¿Cuándo se le acabará la gravedad al universo?

A pesar de estas objeciones, la capacidad explicativa de la Teoría de la Gravitación Universal era, y es hoy en día tan grande, que podían pasar inadvertidas y lo más conveniente consiste en aceptarla plenamente como un "hecho". Para la teoría de la traducción un "hecho" es aquello que tras un periodo previo de controversias es aceptado mayoritariamente por la comunidad científica y deja de discutirse. No obstante, todos los hechos, aun los más asentados como el de la Teoría de la Gravitación, pueden en cualquier momento ser revisados, entrar en un nuevo periodo de controversias y ser modificados. Es decir, pueden traducirse de nuevo.

En nuestro ejemplo, el hecho permaneció incontrovertible durante varios siglos hasta que una nueva traducción se cernió sobre el edificio construido por Newton: la Teoría de la Relatividad de Einstein. Para este científico, la Ley de la Gravitación es una red de conceptos mal construida y propone una nueva red que cambia términos como "fuerza", "acción a distancia" o la ecuación $F=m \cdot a$, por "velocidad de la luz", "relatividad del tiempo", "espacio geométrico" y $E=m \cdot c^2$. No obstante, permanecen muchos otros términos como "masa" o "campo gravitatorio" que nos indican que la red temática de Einstein es la sucesora de la red temática de Newton. Lo que preexiste en la naturaleza es, por ejemplo, la caída de los objetos, en cambio las leyes que explican el fenómeno son tan sólo redes de conceptos que satisfacen necesidades humanas y sociales. En el caso de la "gravitación / espacio geométrico" (depende de la visión explicativa que se emplee), la investigación responde a la necesidad teleológica de explicar la naturaleza última del Universo. Los intereses pueden ser muy variados y siempre consisten en satisfacer necesidades sociales. En definitiva, Newton y Einstein no "descubrieron" un conjunto de leyes físicas sino que las "inventaron". Es el mismo fenómeno de Henry Ford que "no descubrió" el automóvil con motor de gasolina, sino que lo

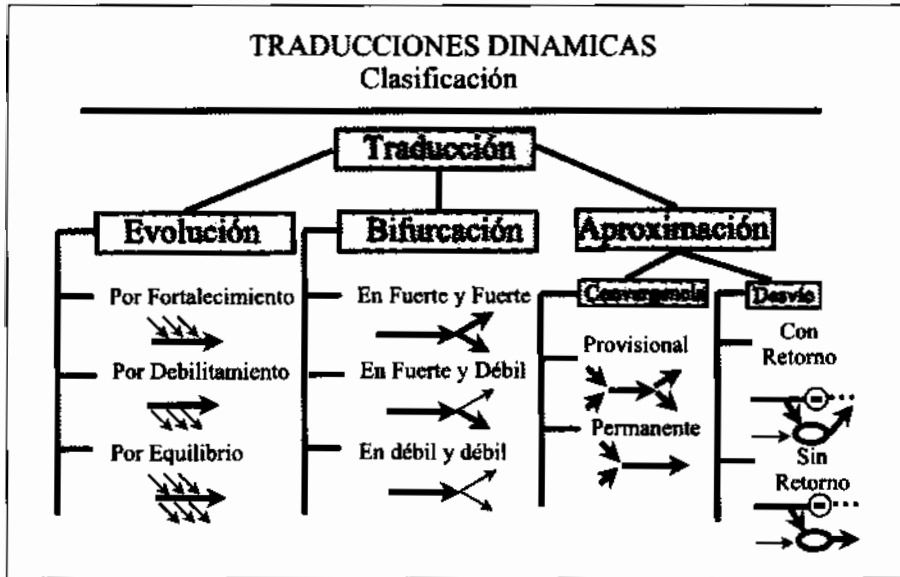


Figura 8. Clasificación de las traducciones

"inventó" y su objetivo era poder disponer de un medio de transporte más eficaz que los anteriores. La ciencia y la tecnología, tal como se pone de manifiesto en el ejemplo de la figura 7 de la empresa láctea y el grupo de investigación, no se dirigen a "descubrir" sino a "inventar" conceptos, objetos o artilugios. Ciencia y tecnología son una construcción social e indisociables de la propia sociedad.

Las traducciones que subyacen en la actividad científico-técnica son muy diversas y se han descrito multitud de ellas (Latour, 1992; Bastide *et al.*, 1989; Callon, 1995; Ruiz-Baños, 1997). Ruiz-Baños clasifica las traducciones según sean de evolución, de bifurcación o de aproximación. Incluye en su clasificación la influencia de la fortaleza de los actores, así como la duración de la traducción (Fig. 8).

7 Los índices del análisis de palabras asociadas

El análisis de palabras asociadas considera que el contenido de un documento viene definido por sus descriptores, obtenidos de las palabras clave, el título, el resumen o de todo el conjunto documental. Se parte, por tanto de una matriz de datos "documentos x descriptores", denominada matriz de ocurrencias, que representa el contenido con-

ceptual del campo científico en estudio (Courtial, 1990). La lista de descriptores puede ser muy extensa, del orden de varios miles, por lo que las dimensiones de esta matriz de ocurrencias son extraordinarias. Cuando un documento "i" contiene el descriptor o palabra clave "j" en la celdilla "i,j" colocamos un 1 y en caso contrario se coloca un 0. El número de veces que una palabra clave "i" aparece u ocurre se denota por C_i . Según la ley de Zipf, la ocurrencia de los descriptores es muy baja para la mayoría de ellos, por lo que los aproximadamente 700 primeros son suficientes para recopilar los verdaderamente relevantes. El resto se desprecian, ya que forman parte del "ruido" de la información.

Decimos que dos palabras co-ocurren cuando aparecen simultáneamente en el mismo documento. Dos palabras estarán más ligadas o asociadas entre sí cuanto mayor sea la co-ocurrencia entre ellas. Es posible igualmente construir una matriz de asociaciones, consistente en una matriz de adyacencia cuadrada de co-ocurrencias de descriptores x descriptores. Cada elemento representa la asociación entre los descriptores. En la celdilla C_{ij} colocamos el número de documentos en los que la palabra "i" y la palabra "j" aparecen simultáneamente. Para evitar que los valores de la matriz de asociaciones dependan del tamaño de la muestra, se recurre a normalizarla emplean-

do el llamado índice de asociación o índice de equivalencia, e_{ij} :

$$e_{ij} = \frac{C_{ij}^2}{C_i C_j}$$

El significado de este índice es muy claro si expresamos la ecuación anterior de la siguiente forma:

$$e_{ij} = \frac{C_{ij}}{C_i} \frac{C_{ij}}{C_j}$$

Donde el primer factor es la probabilidad de tener la palabra i cuando se tiene la palabra j, y el segundo es la probabilidad de tener la palabra j cuando se tiene la palabra i. Este índice es una medida de la relación "Y" entre las palabras i y j

Los valores de e_{ij} oscilan entre 0 y 1. Cuando dos palabras no aparecen nunca juntas, su co-ocurrencia es nula y el índice de equivalencia vale cero. En cambio, cuando dos palabras siempre que aparecen lo hacen juntas en los mismos documentos, el índice de equivalencia es la unidad. Este índice es independiente del tamaño de la muestra.

Esta matriz de asociaciones representa la red científico-técnica del cuerpo documental analizado. Ahora a partir de ella es posible extraer aquellas subredes o actores-red que la conforman recurriendo a dos tipos de algoritmos:

- a) Algoritmo de clasificación por enlace simple.
- b) Algoritmo de agrupación sobre centros simples.

Detalles sobre estos algoritmos pueden encontrarse en Ruiz-Baños, 1998.

Una vez identificados los actores o temas, definidos por sus descriptores y por los enlaces que los unen, es conveniente poder establecer parámetros numéricos que de alguna forma nos hagan referencia a sus estructuras internas y a su relación con la globalidad de la red. Se definen los siguientes índices:

- a) **Densidad.** También llamado índice de cohesión interna, es la in-

tensidad de las asociaciones internas de un tema y representa el grado de desarrollo que posee. Se determina como el cociente entre la suma de los índices de equivalencia internos y el número de palabras que definen el tema multiplicado por 100.

$$d = 100 \frac{\sum_{i=1}^L e_i}{P}$$

donde:

e_i : Índice de equivalencia del enlace interno i

L : Número de enlaces internos del tema.

P : Número de palabras del tema.

b) **Centralidad.** La centralidad

o índice de cohesión externa es la suma de los índices de equivalencia de todos los enlaces externos que posee un tema. Usualmente el valor de la centralidad suele venir multiplicada por el factor 10:

$$c = 10 \sum_j^T e_j$$

donde:

e_j : Índice de equivalencia del enlace externo j

T : Número total de enlaces externos.

Un tema con elevada centralidad está situado en el centro de la red y se relaciona muy bien con los demás actores.

c) **Índice de similitud dinámica, SD.** Mide el grado de parecido

que conserva un tema de investigación cuando pasa de un periodo de tiempo a otro. Es un parámetro relacionado con el grado de traducción que sufre el tema de investigación con el transcurso del tiempo.

Para ilustrar algunas de las posibilidades de evaluación que permite el análisis de palabras asociadas, se ha estudiado el campo científico de los tensioactivos. Los siguientes apartados son una muestra de algunas de sus posibilidades, en concordancia con lo esquematizado en la figura 5.

IQ

Más información gratuita y rápida marcando en la última página el nº 75

