Un modo de análisis de la infraestructura científica de las tecnologías de la información y de las comunicaciones

Xavier Polanco (xavier.polanco@lip6.fr)
Université Pierre et Marie Curie, Francia

El objetivo de este trabajo es presentar un modo de analizar el estado y la evolución de la investigación científica (que en el titulo llamamos "infraestructura científica") de donde se generan las tecnologías de la información y las comunicaciones (TICs), las cuales se encuentran en la base de la sociedad de la información y que darán paso a una sociedad del conocimiento. ¿Es posible en este campo prever, desde la investigación científica, cuál será la nueva generación de TICs? Con el fin de dar una respuesta, proponemos un método ilustrado por un estudio de caso sobre la "web semántica" (2004-2005), representada por 795 datos bibliográficos. El método se resume a la definición de un sistema de categorías, llamado clasificador, en el que los conceptos provenientes de los datos se disponen de acuerdo a un orden estadístico. Luego se procede a su modelización, a partir de la teoría de grafos. Sostenemos que este modo de análisis puede generalizarse al estudio de otros casos, además del ejemplo de la "web semántica" en el que nos apoyamos aquí.

Palabras clave: teoría de grafos, análisis de redes, previsión tecnológica, web semántica, sociedad de la información, sociedad del conocimiento.

The objective of this work is to present a way to analyze the state and the evolution of the scientific research (called "scientific infrastructure" in the title) from where are generated the information and communication technologies (ICTs), which can be found at the base of the information society and are essential to the rise of a knowledge society. Is it possible to anticipate, from the scientific research, which will be the new generation of ICTs? With the purpose of giving an answer, we propose a method illustrated by a case study on the "semantic Web" (2004-2005) represented by 795 bibliographical data. The method is based on the definition of a system of categories, called classifier, in which the concepts originating from the data are arranged according to a statistical order. The system of categories and concepts is then modelled within the parametres of the graph theory. We maintain that this way of analysis can be generalized to the study of other cases, aside from the example of the "semantic Web" that appears in this work.

Keywords: *graphs theory, network analysis, technological forecast, semantic web, information society, knowledge society.*

1. Introducción

Como se dice en el Manual de Lisboa (ML), el desarrollo de las tecnologías de la información y las comunicaciones (TICs) se encuentra en la base de la denominada "sociedad de la información".[1] La propuesta del ML tiene dos componentes: un marco conceptual general, para la medición de la sociedad de la información, y una definición acerca de "cómo abordar el desempeño de los agentes dentro de este nuevo paradigma", el cual se caracteriza por "un profundo cambio en la generación, la gestión y la circulación de la información y el conocimiento".

Con respecto al marco conceptual general, nuestro estudio se refiere solamente a dos de los cuatro sectores de actividad: "ciencia y tecnología", por una parte, e "informática", por otra. El objetivo de este trabajo es analizar y proponer indicadores que permitan apreciar el estado y la evolución de la investigación científica (que en el título llamamos infraestructura científica) a partir de la cual se generan las TICs que se encuentran en la base de la sociedad de la información.

Recordemos que los llamados "sectores o actividades de base" enmarcan en la propuesta del ML lo que allí se llama "la submatriz de difusión y aprovechamiento de la información y el conocimiento". En este nivel, el ML se concentra esencialmente en cómo producir indicadores validos del "uso y acceso" de las TICs en los diversos sectores de la sociedad comportando el prefijo "e-...". A nuestro juicio, el "acceso y uso" de las TICs supone la cuestión previa de saber cuál es el estado de la investigación científica y tecnológica que antecede al hecho de que las TICs se convierten en servicios y bienes económicos al nivel de la sociedad en general.

Por otra parte, nos importa subrayar que el empleo indistinto de los términos "información" y "conocimiento" plantea el problema de su diferenciación, que puede formularse de la manera siguiente: una cosa es la "teoría de la información" y otra la "teoría del conocimiento"; en otras palabras, información no es igual a conocimiento. En efecto, hay una asimetría entre los conceptos de "información" y "conocimiento" en el campo de las TICs: una tarea es procesar información y otra es producir conocimiento. De manera simplificada, digamos que el esfuerzo mayor de la investigación en curso apunta a que las TICs, tales como Internet y la web, pasen de la "información" al "conocimiento", en el sentido que vamos a presentar.

¿Es posible en este campo prever desde la investigación científica cuál será la nueva generación de TICs? Con el fin de dar una respuesta a este interrogante, proponemos un estudio de caso: la "web semántica" (2004-2005) a partir de la base de datos bibliográficos PASCAL (INIST/CNRS), con la intención de poder prever, si es posible, la evolución de la "sociedad de la información" hacia la "sociedad del conocimiento" desde el punto de vista de las TICs. En concreto y para permanecer sobre una base empírica, la tarea consiste en hacer un mapeo del sector de

[1] Véase http://www.ricyt.org.

investigación llamado "web semántica", donde se está preparando la nueva generación internet-web.[2] Recordemos que el proyecto de la web semántica fue formulado por Berners-Lee hacia 1999 y retomado por Berners-Lee, Hendler y Lassila en 2001. Cinco años más tarde, el mismo Berners-Lee firmó con Shadbolt y Hall un nuevo trabajo titulado "The Semantic Web Revisited", en el cual los autores hacen un balance de lo logrado (Berners-Lee et al., 2001; Shadbolt et al., 2006).

2. Datos

En lo que se refiere a los datos, hemos utilizado por razones de comodidad la base PASCAL del INIST/CNRS, en la que la "web semántica" representa:

- 2004 = 330 datos indexados _ 809 palabras claves
- 2005 = 465 datos indexados _ 932 palabras claves

Como sabemos, las publicaciones se utilizan en general para medir y analizar la producción científica. Eso es lo que aquí hacemos, si bien con la ambición de extender más tarde el estudio al campo de las patentes, con el fin de analizar la relación entre ciencia publicada y tecnología patentada en las TICs.

Como se ha dicho en la introducción, el desafío es pasar de esta información -es decir, del hecho de saber que existen 795 datos- al conocimiento que dicha suma de datos representa acerca de la "web semántica". ¿Cómo realizar esta extracción de conocimientos? El método que proponemos es una respuesta a este interrogante.

Precisemos que se ha trabajado con la indización en inglés de los datos, que conservaremos sin traducir al castellano. En la aplicación no ha habido un "control de calidad" de las palabras claves, lo cual se impone cuando se quiere que ellas sean científicamente validas y pertinentes, para contar con conceptos "certificados" desde el punto de vista científico. Por cierto que tal certificación debe ser realizada por investigadores o profesionales calificados, como se hace corrientemente en la minería de datos (*data mining*). Desde ya la categorización que se muestra más abajo, en los cuadros 1 y 2, es una manera de facilitar el trabajo de control y de validación.

3. Metodología

La primera fase es exploratoria y en ella se utilizan métodos de clasificación automática no supervisada ("clustering" o "cluster analysis", en inglés) como técnicas de exploración y extracción de conocimientos a partir de los datos (información). En esta fase se trata de construir un número delimitado de clases agrupando los datos

[2] Véase http://www.w3.org/2001/sw/.

de acuerdo con la similitud de la información que ellos representan y, al mismo tiempo, separando las informaciones no similares en clases distintas. Este es un primer paso para analizar la información con el objetivo de obtener conocimientos, es decir, categorías de análisis. Para este efecto utilizamos el programa SDOC del módulo INFOMETRIA de STANALYST; se trata de un método de clasificación jerárquica ascendente del simple enlace ("single link") basado en las palabras asociadas ("co-word analysis").[3]

En la segunda fase se propone un clasificador compuesto por un conjunto de categorías en el que los elementos clasificados o categorizados se disponen de acuerdo con un orden estadístico. Como resultado de la fase anterior "exploratoria", fueron definidas seis categorías: ontology, semantics, knowledge, reasoning, learning y natural language. La estabilidad o permanencia del clasificador asegura su eficacia; lo que puede cambiar en el tiempo es la configuración o, si se quiere, el orden según el cual las categorías se disponen en el sistema. La operación de clasificación o categorización consiste en asignar a cada una de las categorías (celdas) los conceptos significados por las palabras clave en las cuales se encuentra el nombre de la categoría respectiva.

A continuación viene una nueva fase: la explotación del sistema de categorías y conceptos. Se trata, en otras palabras, de pasar del clasificador al análisis de la red de categorías y conceptos basándonos en la teoría de grafos. En esta etapa trabajamos sobre una muestra de 38 artículos publicados en 2006. Por cierto que la metodología es extensible en principio a cualquier cantidad de datos. Pero para ello es necesario contar con la ayuda de programas informáticos adecuados para la categorización y la representación de las redes de categorías y de conceptos. Digamos que la tarea propiamente informática o de programación está aún por realizarse.

En cuanto a las bases del método, nos apoyamos tanto en la tradición de las "palabras asociadas" ("co-word analysis") (Callon et al., 1983, 1986; Courtial, 1990) como en el análisis de redes sociales ("social network analysis") (Wasserman y Faust, 1999), dos tradiciones que hasta ahora se han desarrollado de manera separada, e incluso ignorándose entre sí.

4. Resultados

Dejaremos de lado los resultados de la fase 1 (exploratoria) para concentrarnos sobre las operaciones de las fases 2 y 3, esto es, la organización de los datos en categorías y conceptos y la representación de las categorías y los conceptos como grafos, poniendo así en evidencia las redes implícitas en las categorizaciones.

[3] Véase http://stanalyst.inist.fr.

4.1 Organización de la información en categorías y conceptos

Los cuadros 1 y 2 muestran los sistemas de categorías y conceptos de los años 2004 y 2005, respectivamente. Se trata, como puede apreciarse, de seis celdas en las que figura una lista de conceptos que sigue un orden estadístico, la frecuencia del concepto en la colección de documentos, cuya cantidad y porcentaje indican la extensión del concepto.

Cuadro 1. Sistema de categorías y conceptos, datos 2004

	%	ONTOLOGY		%	SEMANTICS		%	KNOWLEDGE
175	53,03	Ontology	102	30,91	Semantics	43	13,03	Knowledge base
21	6,36	Description logic	7	2,12	Semantic network	40	12,12	Knowledge engineering
14	4,24	Description language	6	1,82	Semantic analysis	22	6,67	Knowledge representation
1	0,30	Ontology mapping	4	1,21	Formal semantics	8	2,42	Knowledge acquisition
1	0,30	Standard upper ontology	4	1,21	Semantic relation	6	1,82	Knowledge management
			1	0,30	Operational semantics	3	0,91	Knowledge based systems
						3	0,91	Knowledge discovery
						1	0,30	Knowledge
						1	0,30	Knowledge grid
						1	0,30	Knowledge representation languages
	%	REASONING		%	LEARNING		%	NATURAL LANGUAGE
8	2,42	Model-based reasoning	6	1,82	Learning algorithm	12	3,64	Natural language
7	2,12	Case based reasoning	3	0,91	Machine learning	7	2,12	Language processing
4	1,21	Automated reasoning	1	0,30	Inductive learning	4	1,21	Linguistics
2	0,61	Temporal reasoning	1	0,30	Learning (artificial intelligence)	2	0,61	Natural language processing
			1	0,30	Learning object metadata	1	0,30	Computational linguistics
			1	0,30	Probability learning	1	0,30	Linguistic model
			1	0,30	Reinforcement learning	1	0,30	Structural linguistics
			1	0,30	Supervised learning	1	0,30	Language generation

Para cada una de las categorías se utilizó una función de búsqueda del nombre de la categoría en la lista del vocabulario de indización, y que había sido previamente analizada estadísticamente utilizando el modulo Bibliometría de STANALYST. La única excepción es la inclusión en la categoría “Ontology” de “description logic” y “description language”, que no contienen en su composición la palabra “ontología”. Dicha inclusión está fundada en el conocimiento de que los “lenguajes o lógicas de descripción” se utilizan para la programación de ontologías (o sea, sistemas de representación de conocimientos de un dominio dado), al punto de que a veces se habla de ellos como “ontology languages” (Staab y Studer, 2004).

La información estadística que acompaña a los términos en los cuadros 1 y 2 define, como se ha dicho, la extensión del concepto, es decir, el número de documentos que ellos indexan. Este mismo criterio se aplica al nivel de la categoría como la suma de los documentos indexados por los conceptos de la categoría, si bien es necesario tomar en cuenta que esta suma no es la simple adición de los números que figuran en la celda de una categoría, dado que un mismo documento puede estar indexado a la vez por dos o más palabras clave de la misma categoría.

Por otra parte, podemos comparar la misma estructura en dos momentos distintos, como aquí se hace, y de esta manera apreciar la evolución sin recurrir a una encuesta de los actores comprometidos en el campo científico considerado. En este caso, solamente se consideran las publicaciones que estos actores han producido acerca de la web semántica dentro de un periodo dado.

Cuadro 2. Sistema de categorías y conceptos, datos 2005

	%	ONTOLOGY
272	58,50	Ontology
15	3,23	Ontology mapping
31	6,67	Description logic
13	2,80	Description language

	%	SEMANTICS
187	40,22	Semantics
11	2,37	Semantic analysis
9	1,94	Formal semantics
5	1,08	Semantic network
4	0,86	Semantic relation
3	0,65	Operational semantics
1	0,22	Semantic memory

	%	KNOWLEDGE
43	9,25	Knowledge base
42	9,03	Knowledge representation
24	5,16	Knowledge engineering
10	2,15	Knowledge discovery
10	2,15	Knowledge management
4	0,86	Knowledge acquisition
1	0,22	Knowledge
1	0,22	Knowledge based systems
1	0,22	Knowledge representation languages
1	0,22	Knowledge transfer

	%	REASONING
8	1,72	Case based reasoning
3	0,65	Automated reasoning
2	0,43	Model-based reasoning
2	0,43	Temporal reasoning
1	0,22	Common-sense reasoning
1	0,22	Qualitative reasoning
1	0,22	Reasoning
1	0,22	Spatial reasoning

	%	LEARNING
6	1,29	Learning algorithm
2	0,43	Inductive learning
1	0,22	Concept learning
1	0,22	Learning
1	0,22	Unsupervised learning

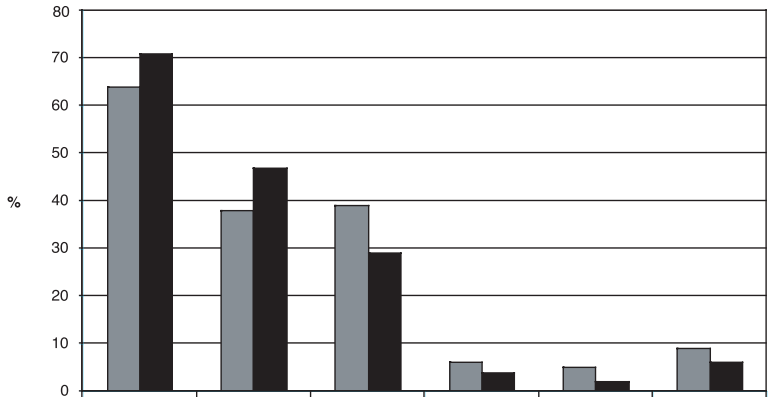
	%	NATURAL LANGUAGE
13	2,80	Natural language
9	1,94	Linguistics
3	0,65	Language family
2	0,43	Language class

Las categorías significan objetos de investigación, al mismo tiempo que determinan un campo científico, una especialidad importante para llevar adelante el proyecto tecnológico de la “web semántica”. En efecto, no hay ontología [1] sin semántica [2] y es así como el conocimiento [3] puede ser introducido en los ordenadores y en consecuencia en la web, al mismo tiempo que la capacidad de razonamiento [4] en los servidores y motores de búsqueda, se busca igualmente que servidores y motores afinen su funcionamiento mediante aprendizaje [5]. Por otra parte, como el lenguaje es la forma natural de la comunicación humana, el desafío es procesar el lenguaje natural [6] que se encuentra en los datos y usarlo en el dialogo entre usuario y computador. En suma, seis propiedades necesarias (pero tal vez no suficientes) para que la web que conocemos actualmente evolucione en el sentido de la llamada “web semántica”. Es decir, la web será “semántica” si incorpora al menos estas seis propiedades.

Los cuadros 1 y 2 muestran la “infraestructura científica” de la innovación tecnológica como resultado de la capacidad de procesar e incorporar conocimientos en los sistemas de información y comunicación, cuyo “impacto social” se traducirá en la emergencia de la “sociedad del conocimiento”.

La figura 1 es un ejemplo de cómo podemos apreciar la evolución y comparar las categorías de acuerdo con el porcentaje de publicaciones que cada una de ellas representa en 2004 y 2005, destacándose “Ontology”, seguida de “Semantics”, luego “Knowledge”, en una menor medida “Natural language” y seguido de forma pareja por “Reasoning” y “Learning”.

Figura 1. Las categorías en porcentaje de documentos 2004 y 2005



	Ontology	Semantics	Knowledge	Reasoning	Learning	Natural language
2004	64	38	39	6	5	9
2005	71	47	29	4	2	6

Por cierto que considerar el desarrollo entre un año y el siguiente no permite concluir sobre la evolución de las categorías. Para ello se necesita tomar en cuenta un periodo, digamos 2000-2006, si consideramos que el proyecto de la “web semántica” fue enunciado en 2000 y revisado por su autor en 2006, como se ha citado en la introducción. En la figura 1, sólo dos categorías aumentan su porcentaje de un año al otro y las cuatro restantes disminuyen en proporciones más o menos parecidas.

4.2 Grafo de la red de conocimientos

A fin de representar la red implícita a los sistemas de categorías y conceptos (o clasificadores) nos basamos en la teoría de grafos. Para la comodidad de la demostración, hemos constituido una muestra de 38 publicaciones sobre la “web semántica” de fecha 2006. Por cierto que la metodología se aplica en principio a cualquier cantidad de datos. A título de ejemplo, el cuadro 3 expone el sistema de categorías y conceptos de la muestra.

Es a partir de este ejemplo que la tarea es ahora traducir el sistema en un grafo. Para ello es necesario apoyarse sobre dos matrices: la matriz de incidencia

categorías-documentos, que permite conocer la distribución de los documentos en las categorías, y la matriz de adyacencia categorías-categorías, que da cuenta de las relaciones (aristas) entre las categorías. Si al menos existe un documento común, esta relación puede ser anotada 1 (presencia) y 0 (ausencia), o bien considerar el número de documentos que soportan la existencia de la relación. Entonces se habla de relaciones valuadas, como vemos en el grafo de la figura 2. El cuadro 4 expone la matriz de adyacencia entre las categorías de donde se construye el grafo de la figura 2. Las relaciones entre categorías están indicadas en cada celda por el número de documentos que ellas representan. La relación más fuerte, = 10, es entre "Ontology" y "Knowledge", seguida por la relación = 4 entre "Ontology" y "Semantics".

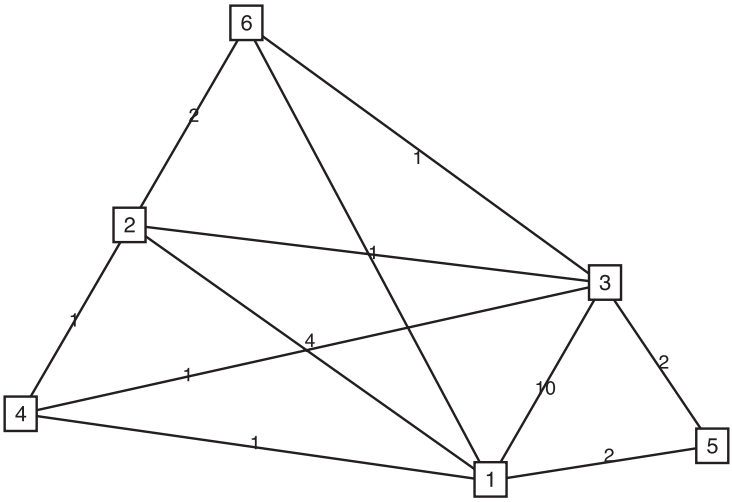
Cuadro 3. Ejemplo basado en 38 publicaciones 2006

	%	ONTOLOGY		%	SEMANTICS		%	KNOWLEDGE
18	44,74	Ontology	7	18,42	Semantics	3	7,90	Knowledge base
2	5,26	Description language	3	7,90	Semantic análisis	3	7,90	Knowledge engineering
2	5,26	Description language	1	2,63	Semantic network	3	7,90	Knowledge management
1	2,63	Data description language				3	7,90	Knowledge representation
						2	5,26	Knowledge acquisition
	%	REASONING		%	LEARNING		%	NATURAL LANGUAGE
1	2,63	Automated reasoning	1	2,63	Concept learning	2	5,26	Linguistic
1	2,63	Case based reasoning	1	2,63	Machine learning	1	2,63	Linguistic tool

Cuadro 4. Matriz de adyacencia categorías-categorías

		[1]	[2]	[3]	[4]	[5]	[6]
ONTOLOGY	[1]		4	10	1	2	1
SEMANTICS	[2]			1	1	0	2
KNOWLEDGE	[3]				1	2	1
REASONING	[4]					0	0
LEARNING	[5]						0
NATURAL LANGUAGE	[6]						

Figura 2. El grafo de las categorías



i) Análisis del grafo de categorías. En la figura 2, las categorías están representadas por un número y sobre las relaciones figura el número de documentos que ellas representan, el cual podemos leer en la matriz de adyacencia categorías-categorías del cuadro 4. Como se observa en la figura 2, el grafo G es un conjunto de nodos N y un conjunto de relaciones R, en donde N es igual a 6 y R es igual a 11. El numero total de relaciones posibles del grafo esta dado por N (N - 1) / 2, es decir 15. Sobre esta base se calcula la densidad del grafo de acuerdo con la formula D = R / N (N - 1) / 2 = 2R / N (N - 1). Entonces: D = 11 / 15 = 0,73 (73%). Esta medida permite comparar la consistencia de dos grafos, en nuestro caso, comparar la estructura de los grafos (de las seis categorías) de diferentes años o periodos, y saber si se fortalece o debilita desde el punto de vista de su densidad.

Un segundo elemento de análisis es la estructura del grafo que no vemos en la matriz de adyacencia categorías-categorías, pero que la figura 2 pone en evidencia. Allí pueden apreciarse tres subgrafos completos, es decir, los nodos están completamente ligados entre ellos (D = 1). En efecto, podemos distinguir en el grafo dos equiláteros con sus respectivas diagonales y un triángulo:

* 1-2-3-4	=	Ontology, Semantics, Knowledge, Reasoning
* 1-2-3-6	=	Ontology, Semantics, Knowledge, Natural language
* 1-3-5	=	Ontology, Knowledge, Learning

Estos agrupamientos pueden, entonces, ser analizados separadamente, con el fin de

afinar aún más el análisis. Si N = 4 el número máximo de relaciones posibles es igual a 6 (un equilátero con sus diagonales), si N = 3 sólo son posibles 3 relaciones (un triangulo). Este ejemplo de configuración estructural sugiere que podemos compararla en el tiempo, esto es, comparar las formas según las cuales se configura el grafo de categorías en el tiempo.

A su vez, los nodos del grafo, o categorías, tienen dos valores estructurales: una es la densidad (D) y la otra es la centralidad (C), como vemos en el cuadro 5. La densidad se mide como hemos dicho anteriormente, sólo que ahora no se trata de la densidad global del grafo de categorías sino de los nodos, es decir, de cada una de las categorías, puesto que, como veremos, cada nodo es a su vez un grafo. Por su parte, la centralidad o importancia de los nodos en el grafo se mide por el número de relaciones r que cada uno presenta o grado (g), de modo que $C(g) = \sum(r)$. El problema con esta medida es que ella depende de la talla del grafo, cuyo valor máximo es N - 1. En consecuencia, su medida estandarizada es $C(g) = \sum g / N - 1$. Otra forma de medir C es haciendo la suma del peso o valor v de las relaciones. Este valor es aquí igual al número de documentos d que soportan estas relaciones, $C(v) = \sum(d)$.

Cuadro 5. Densidad y centralidad de las categorías (nodos del grafo de la figura 2)

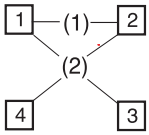
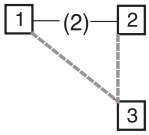
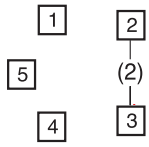
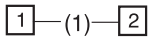
		D	C (g)	C (g)/N -1	C (v)
ONTOLOGY	[1]	0,14	5	1	18
SEMANTICS	[2]	0,10	4	0,8	8
KNOWLEDGE	[3]	0,07	5	1	15
REASONING	[4]	0,03	4	0,8	3
LEARNING	[5]	0,03	2	0,4	4
NATURAL LANGUAGE	[6]	0,04	3	0,6	4

Vemos en el cuadro 5 que “Ontology” es por lejos la categoría más densa y, al mismo tiempo, la más central por el nombre de documentos, C(v), y a igualdad con “Knowledge” en lo que respecta a la centralidad de grado, C(g). Y si además consideramos que la extensión de “Ontology” es de 45% (es decir, que ella cubre aproximadamente el 45% de las publicaciones de la muestra), mientras que la extensión de “Knowledge” solo es de 15%. En otras palabras, los indicadores están señalando que “Ontology” constituye, de acuerdo con el ejemplo, la categoría central y más importante.

ii) Análisis de los nodos o conceptos internos de las categorías. Además, cada nodo del grafo de categorías (figura 2) puede representarse como un grafo de conceptos (señalados en el cuadro 3). Como se observa en el cuadro 6, dos clases de nodos se destacan a primera vista: nodos enlazados por una arista, o relación cuyo valor es igual al número de documentos que la soportan, y nodos aislados. La

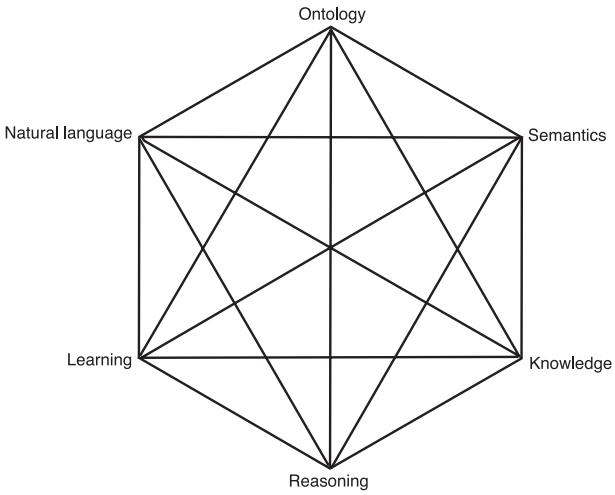
densidad de un nodo, esto es, de una categoría, es función de los tres factores siguientes: (i) el número n de conceptos que la constituyen a un momento dado, (ii) el número r de relaciones y (iii) el número d de documentos que soportan las relaciones con respecto al total de documentos de la categoría (∑d).

Cuadro 6. Grafos internos a los nodos-categorías

[1] Ontology	[2] Semantics	[3] Knowledge
		
[4] Reasoning	[5] Learning	[6] Natural Language
		

iii) Un instrumento de previsión. Llamamos la atención sobre el hecho de que con el cuadro 6 se dispone de un instrumento de pronóstico o previsión, en el sentido de que donde no hay un enlace entre dos nodos existe la posibilidad de que posteriormente, t+1, se cree un enlace entre esos nodos (conceptos), sobre la base del principio que una red tiende a hacerse más densa, completando las relaciones entre los nodos que la componen y no solamente aumentando el número de nodos. En el cuadro 6, los enlaces posibles están señalados por líneas discontinuas a título de ejemplos.

Figura 3. El grafo completo de la infraestructura científica de la "web semántica"



Lo que se ha dicho en el párrafo precedente se aplica igualmente al grafo de categorías de la figura 2. En otras palabras, la hipótesis es que el grafo que vemos en la figura 2 tendría como tendencia interna devenir en un grafo completo. Se llama grafo completo al grafo cuyos nodos (o vértices, de acuerdo con el lenguaje de grafos) están todos ligados de dos en dos por una relación (o arista), como se observa en la figura 3. En general, un grafo completo de N nodos contiene N (N - 1) / 2 relaciones. En el caso de nuestro grafo, N = 6 entonces R (es decir el conjunto de relaciones) = 15. Lo que da por resultado el hexágono completamente ligado que vemos en la figura. Podríamos decir que ella define de una manera gráfica la infraestructura científica de las TICs que están por venir y que son necesarias para que se desarrolle la llamada "sociedad del conocimiento". Dicho esquemáticamente, la web actual es a la "sociedad de la información" lo que la "web semántica" será a la "sociedad del conocimiento": el paso de una a la otra supone entonces la puesta en marcha tecnológica, en la práctica social, del paradigma científico que el grafo completo de las categorías representa.

5. Conclusión

El objetivo que nos propusimos fue proponer un método de análisis y, al mismo tiempo, indicadores de la investigación científica de donde se generan las TICs. Para ello nos apoyamos en un ejemplo concreto, real, la "web semántica" (2004-2005), que nos permitió resumir la infraestructura científica en seis categorías principales y que tal vez podamos considerar como el paradigma científico de las nuevas TICs.

En efecto, la “Ontología” [1] permite al computador disponer de una “Semántica” [2], con el fin de representar y manejar “Conocimientos” [3], de poder además realizar inferencias o “Razonamientos” [4] y, en la ejecución de sus tareas, “Aprender” [5] a hacerlas cada vez mejor, además de integrar el “Lenguaje natural” [6] en el procesamiento de los datos y en el diálogo usuario-ordenador. Cuando esta infraestructura científica se difunda y generalice al nivel de las TICs y éstas en la practica social, entonces podremos hablar con propiedad de “sociedad del conocimiento”.

Un punto importante es la generalización del modo de análisis que se ha presentado. El modo de análisis que venimos de proponer puede generalizarse al estudio de otros casos diversos del ejemplo de la “web semántica” en el que nos hemos apoyado aquí. El modo de análisis se resume a la categorización, mediante la definición de un clasificador, y luego a su modelización, de acuerdo con la teoría de grafos. Aquí hemos enunciado el principio o la hipótesis que una red dada de conocimientos tiende a hacerse más densa, completando las relaciones entre los nodos que la componen, y no sólo a crecer en talla, aumentando el número de nodos. En lenguaje de grafos: el grafo inicial tiende a devenir en un grafo completo, pasando por la etapa de subgrafos completos. La hipótesis es demasiado fuerte, puesto que ella supone (como se observa en la figura 3) que todas la categorías tienen una misma importancia o centralidad, lo cual difícilmente ocurre. Se trata más bien de un tipo ideal.

El punto crítico del método de análisis propuesto es que falta un programa automatizando para las tareas que el método supone, más exactamente las fases 2 y 3. Por ahora se trabajó “a mano”, con la ayuda de una hoja de cálculo y de un editor interactivo de grafos. La ambición es programar el método propuesto. Un aspecto importante es el paso del análisis de datos basado en la clasificación no supervisada a la matriz de categorías y conceptos, sabiendo que ésta no se deriva directa o automáticamente de la primera; ella supone el empleo del clasificador. La fase del análisis de datos basado en una clasificación automática no supervisada (o fase 1) aparece como una poderosa ayuda en el trabajo de determinar las categorías y el contenido conceptual de ellas. Esta observación nos obliga a precisar que la clasificación automática se divide en dos grandes ramas: “clasificación no supervisada” -es lo que se llama “clustering” o “cluster analysis”, en inglés- y “clasificación supervisada” o simplemente “clasificación”, la cual supone la acción algorítmica de un clasificador. La primera produce clases (o “clusters”) de datos, mientras que la segunda produce categorías de conceptos.

Bibliografía

BERNERS-LEE, T. HENDLER, J. y LASSILA, O. (2001): "The Semantic Web", *Scientific American,* Mayo, p. 34-43.

CALLON, M., COURTIAL, J. P., TURNER, W. A. y BAUIN, S. (1983): "From translations to problematic networks: An introduction to co-words analysis", *Social Science Information,* vol. 22, p. 191-235.

CALLON, M., LAW, J. y RIP, A. (eds) (1986): *Mapping the Dynamics of Science and Technology,* London, Macmillan Press.

COURTIAL, J. P. (1990): *Introduction à la scientométrie,* Paris, Anthropos-Economica.

SHADBOLT, N., HALL, W. y BERNERS-LEE, T. (2006): "The Semantic Web Revisited", *IEEE Intelligent Systems,* Mayo/Junio, p. 96-101.

STAAB, S. y R. STUDER, R. (eds.) (2004): *Handbook on Ontologies,* Berlin, Springer.

WASSERMAN, S. y FAUST, K. (1999) *Social Network Analysis. Methods and Applications,* Londres, Cambridge University Press.