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

Xavier Polanco

Revista Iberoamericana de Ciencia, Tecnología y Sociedad, vol. 3, núm. 9, agosto 2007, p. 77-90

Un Modo de Análisis de la infraestructura científica de las Tecnologías de la Información y de las Comunicaciones

Xavier Polanco

Laboratoire d'Informatique de Paris 6
CNRS UMR 7606 - Université Pierre et Marie Curie - Paris 6
104 avenue Président Kennedy - 75016 Paris - France

1. Introducción

Como se dice en el *Manual de Lisboa* (ML)[1], el desarrollo de las tecnologías de la información y las comunicaciones (TICs) se encuentra a la base de la denominada "sociedad de la información". La propuesta del ML tiene dos componentes: un marco conceptual general (para la medición de la sociedad de la información) y luego define "cómo abordar el desempeño de los agentes dentro de esta 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, estos son "ciencia y tecnología" por una parte e "informática" por otra. El objetivo de este trabajo es de analizar y proponer indicadores que permitan apreciar el estado y la evolución de la investigación científica (y que en el titulo llamamos infraestructura científica) de donde se generan las TICs que se encuentran a 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 ahí se llama "la submatriz de difusión y aprovechamiento de la información y el conocimiento". Y a este nivel, el ML se concentra esencialmente sobre 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 cual es el estado de la investigación científica y tecnológica que antecede al hecho que las TICs se conviertan en servicios y bienes económicos al nivel de la sociedad en general.

Por otra parte, nos importa subrayar que el empleo de los términos "información" y "conocimiento" indistintamente (información = conocimiento) plantea el problema de su distinción y 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 (información ≠ 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 conocimiento. Simplificado, digamos que el esfuerzo mayor en la investigación en curso apunta a que las TICs, tales que 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 cual será la nueva generación de TICs? Con el fin de dar una respuesta a esta interrogación, proponemos un

[1] Véase http://www.ricyt.edu.ar/

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 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 investigación llamado “web semántica”[2] en donde se esta preparando la nueva generación internet-web. 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, y cinco años más tarde el mismo Berners-Lee firma con Shadbolt y Hall un nuevo trabajo titulado “The Semantic Web Revisited”, en el cual los autores hacen un balance de lo logrado (ver 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 donde 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, y es lo que aquí hacemos pero con la ambición de extender más tarde el estudio que aquí se presenta 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. ¿Como realizar esta extracción de conocimientos? El método que proponemos es una respuesta a esta interrogación.

Precisemos que se ha trabajo con la indización en inglés de los datos y que vamos a conservar 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; 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 para ello se utilizan métodos de clasificación automática no supervisada (en inglés “clustering” o “cluster análisis”) 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 de acuerdo a 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[3], se trata de un método de clasificación jerárquica ascendente del simple enlace (“single link”) basado en las palabras asociadas (“co-word analysis”).

En la segunda fase se propone un clasificador compuesto por un conjunto de categorías en donde los elementos clasificados o categorizados se disponen de acuerdo a un orden estadístico. Como resultado de la fase anterior “exploratoria”, fueron definidas seis categorías: ONTOLOGY, SEMANTICS, KNOWLEDGE, REASONING, LEARNING, NATURAL LANGUAGE. La estabilidad o permanencia del clasificador asegura su eficacia,

[2] Véase http://www.w3.org/2001/sw/
[3] Véase http://stanalyst.inist.fr

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 claves en las cuales se encuentra el nombre de la categoría respectiva.

Enseguida viene una nueva fase: la explotación del sistema de categorías y conceptos. En otras palabras, 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 cualquiera cantidad de datos. Pero para ello es necesario contar con la ayuda de programas informáticos adecuados para la categorización primero y la representación de las redes de categorías y de conceptos en segundo lugar. Digamos que la tarea propiamente informática o de programación está aun por realizarse.

En cuanto a las bases del método, nos apoyamos por una parte en la tradición de las "palabras asociadas" ("co-word analysis") (Callon et al, 1983, 1986; Courtial, 1991), y por otra en la del análisis de redes sociales ("social network analysis") (Wasserrman & Fausto, 1999). Dos tradiciones que hasta ahora se han desarrollado independientemente, y aun más, ignorándose.

4. Resultados

Los resultados de la fase 1 (exploratoria) los dejaremos de lado, 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 conceptos como grafos poniendo así en evidencia las redes implícitas en las categorizaciones.

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 siguiendo 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 BIBLIOMETRIA de STANALYST. La sola 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 que los "lenguajes o lógicas de descripción" se utilizan para la programación de ontologías (= sistemas de representación de conocimientos de un dominio dado), al punto que a veces se habla de ellos como "ontology languages" (Staan & 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, pero atención, 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 claves de la misma categoría.

Por otra parte, podemos comparar la misma estructura a 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		%	SEMANTICS		%	KNOWLEDGE
272	58,50	Ontology	187	40,22	Semantics	43	9,25	Knowledge base
15	3,23	Ontology mapping	11	2,37	Semantic analysis	42	9,03	Knowledge representation
31	6,67	Description logic	9	1,94	Formal semantics	24	5,16	Knowledge engineering
13	2,80	Description language	5	1,08	Semantic network	10	2,15	Knowledge discovery
			4	0,86	Semantic relation	10	2,15	Knowledge management
			3	0,65	Operational semantics	4	0,86	Knowledge acquisition
			1	0,22	Semantic memory	1	0,22	Knowledge
						1	0,22	Knowledge based systems
						1	0,22	Knowledge representation languages
						1	0,22	Knowledge transfer
	%	REASONING		%	LEARNING		%	NATURAL LANGUAGE
8	1,72	Case based reasoning	6	1,29	Learning algorithm	13	2,80	Natural language
3	0,65	Automated reasoning	2	0,43	Inductive learning	9	1,94	Linguistics
2	0,43	Model-based reasoning	1	0,22	Concept learning	3	0,65	Language family
2	0,43	Temporal reasoning	1	0,22	Learning	2	0,43	Language class
1	0,22	Common-sense reasoning	1	0,22	Unsupervised learning			
1	0,22	Qualitative reasoning						
1	0,22	Reasoning						
1	0,22	Spatial reasoning						

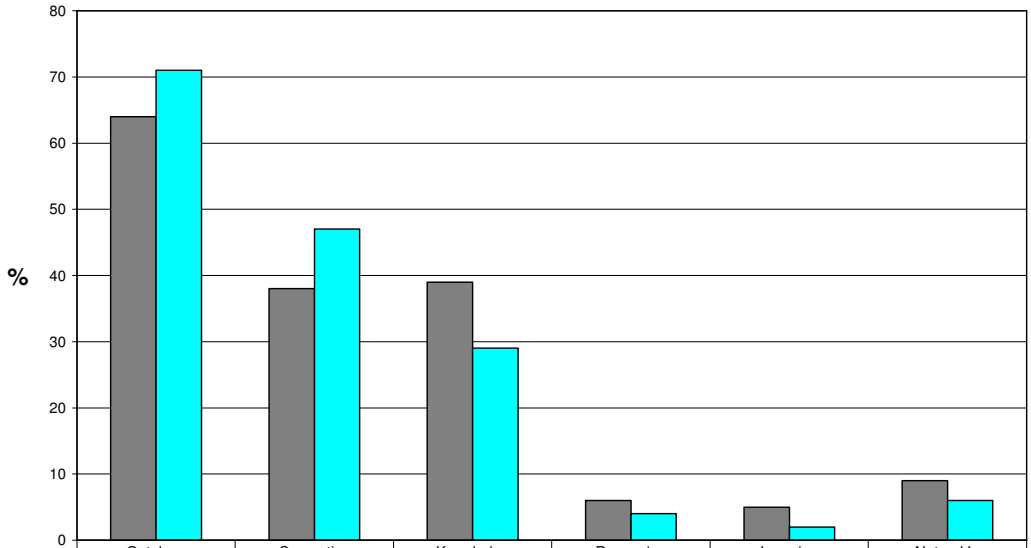
Las categorías están significando objetos de investigación al mismo tiempo que determinan un campo científico, una especialidad importante para llevar adelante el proyecto tecnológico de la web semántica. En efecto, no hay ontología [1] sin semántica [2] y es así como el conocimiento [3] puede ser introducido en los computadores y por vía de consecuencia en la web, asimismo 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 digamos necesarias (pero tal vez no suficientes) para que la web que conocemos

actualmente evolucione en el sentido de la llamada web semántica. La web será “semántica” si incorpora al menos estas seis propiedades.

Los cuadros 1 y 2 nos están mostrando la “infraestructura científica” de la innovación tecnológica como resultado de la capacidad de procesar y de incorporar conocimientos en los sistemas de información y comunicación, y 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 al porcentaje de publicaciones que cada una de ellas representa en 2004 y 2005, destacándose ONTOLOGY seguida de SEMANTICS y luego KNOWLEDGE, y en una menor medida NATURAL LANGUAGE seguido a igualdad por REASONING y LEARNING.

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



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

Por cierto que considerar el desarrollo de un año al 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 revisto por su autor en 2006 como se ha citado en la introducción. En la figura 1 solo 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 cualquiera cantidad de datos. A titulo 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 de traducir el sistema en un grafo. Para ello es necesario apoyarse sobre dos matrices: la matriz de incidencia categorías-documentos que permite de 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 si no lo hay (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 mas fuerte = 10 es entre ONTOLOGY y KNOWLEDGE, seguida por la relación = 4 entre ONTOLOGY y SEMANTICS.

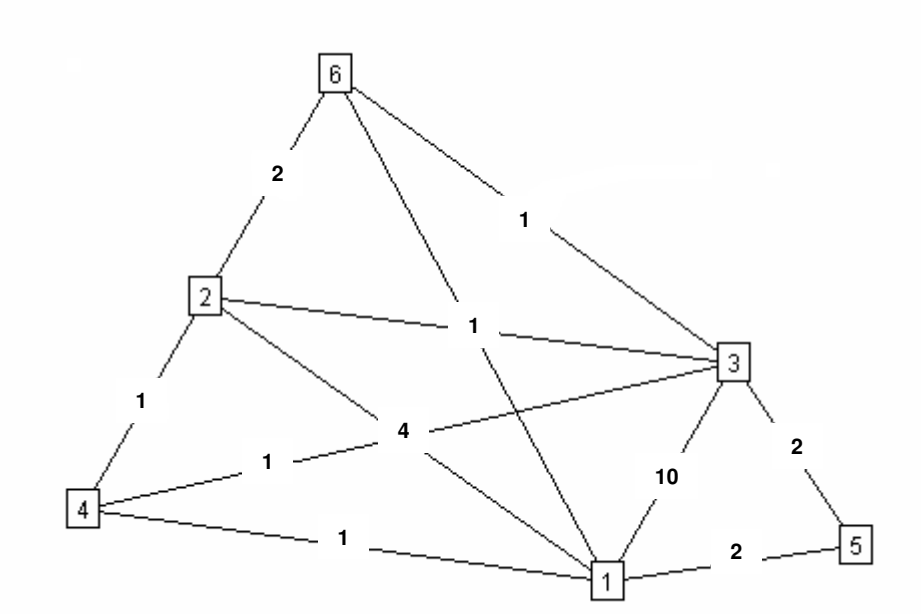
Cuadro 3: Ejemplo basado en 38 publicaciones 2006

	%	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 analysis	3	7,90	Knowledge engineering
2	5,26	Description logic	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	Linguistics
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



1. 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 y que 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 $N = 6$ y $R = 11$. El número 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, de 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 y que no vemos en la matriz de adyacencia categorías-categorías pero que la figura 2 pone en evidencia, y en donde pueden apreciarse tres subgrafos completos, es decir que 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

Los cuales pueden entonces analizarse separadamente con el fin de afinar aun más nuestro análisis. Si $N = 4$ el número máximo de relaciones posibles es igual a 6 (un equilátero con sus diagonales), si $N = 3$ solo son posibles 3 relaciones (un triángulo). Este ejemplo de configuración estructural sugiere que podemos compararla en el tiempo, es decir, 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 la centralidad (*C*) como vemos en el cuadro 5. La densidad se mide como hemos dicho anteriormente, solo que ahora no se trata de la densidad global del grafo de categorías sino que 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*), así $C(g) = \sum(r)$; 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 $C(g) = \sum r / N-1$. 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)

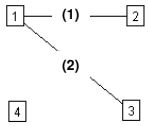
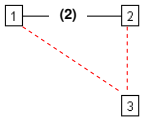
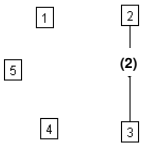
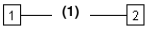
		D	C(g)	C(g)/N -1	C(v)
ONTOLOGY	[1]	0,14	5	1	18
SEMANTICS	[2]	0,10	4	0,8	8
KNOWLEDGE	[3]	0,07	5	1	15
REASONING	[4]	0,03	4	0,8	3
LEARNING	[5]	0,03	2	0,4	4
NATURAL LANGUAGE	[6]	0,04	3	0,6	4

Vemos en el cuadro 5 que ONTOLOGY es de 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.

2. 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

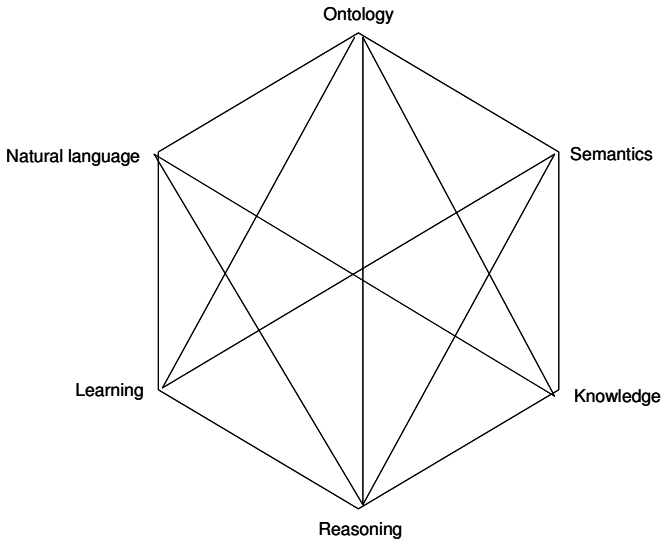
(señalados en el cuadro 3). Como se observa en el cuadro 6 dos clases de nodos se destacan a primera vista: nodos enlazados por una arista o relación cuyo valor es igual al número de documentos que la soportan y nodos aislados. La densidad de un nodo, esto es de una categoría, es función de los tres factores siguientes: (i) el número *n* de conceptos que la constituyen a un momento dado, (ii) el numero *r* de relaciones, y (iii) el numero *d* de documentos que soportan las relaciones con respecto al total de documentos de la categoría ($\sum d$).

Cuadro 6: Grafos internos a los nodos-categorías

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

3. Un instrumento de previsión. Llamamos la atención sobre el hecho que con el cuadro 6 se dispone de un instrumento de pronóstico o previsión, en el sentido que donde no hay un enlace entre dos nodos existe la posibilidad 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 titulo de ejemplos.

Figura 3: El grafo completo de la infraestructura científica de la “web semántica”



Lo que se ha dicho en el parágrafo 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 un grafo completo, se llama grafo completo el 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 grafica la infraestructura científica de las TICs à venir y 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) y que nos permitió de 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 de 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 integrar el LENGUAJE NATURAL [6] en el procesamiento de los datos y en el dialogo usuario-computador. 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".

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 que el 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 hipótesis que una red dada de conocimientos tiende a hacerse más densa completando las relaciones entre los nodos que la componen, y no solo a crecer en talla aumentando el número de nodos. En lenguaje de grafos: el grafo inicial tiende a devenir 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. Es más bien un tipo ideal.

El punto crítico del método de análisis propuesto es que falta el programa automatizando las tareas que el método supone, mas exactamente las fases 2 y 3. Por ahora se trabajó a la "mano" con la ayuda de una hoja de cálculo y de un editor interactivo de grafos. La ambición es de 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 directamente 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" y es lo que se llama "clustering" o "cluster análisis" en inglés, y "clasificación supervisada" o simplemente "clasificacion", la cual supone la acción algorítmica de un clasificador. La primera produce clases (o clusters) de datos y la segunda categorías de conceptos.

Referencias

M. Callon, J-P.Courtial, W. A.Turner, S. Bauin. (1983) From translations to problematic networks: An introduction to co-words analysis. *Social Science Information*, vol. 22, p. 191-235.

M. Callon, J. Law and A. Rip (eds) (1986) *Mapping the Dynamics of Science and Technology*. London: Macmillan Press.

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

T. Berners-Lee, J. Hendler, O. Lassila (2001) The Semantic Web. *Scientific American*, May, p. 34-43.

N. Shadbolt, W. Hall, T. Berbers-Lee (2006) The Semantic Web Revisited. *IEEE Intelligent* Systems, May/June, p. 96-101.

S. Staab & R. Studer (editors) (2004) *Handbook on Ontologies*. Berlin: Springer.

S. Wasserman, K. Faust (1999) *Social Network Analysis. Methods and Applications*. Cambridge: Cambridge University Press.