15

RECOLECCIÓN, DETECCIÓN DE COMUNIDADES Y VISUALIZACIÓN DE INFORMACIÓN WEB

José L. Alonso Berrocal
Carlos G. Figuerola
José Federico Medrano
Universidad de Salamanca

1. Introducción

El web es una colección de billones de documentos escritos de tal forma que pueden ser citados usando hiperenlaces y conformando el denominado hipertexto. Estos documentos, o páginas web, tienen unos pocos cientos de caracteres escritos en infinidad de idiomas y que cubren esencialmente todas las materias del saber humano.

Según (Baeza-Yates et al. 2005) una de las grandes ventajas de la Web es precisamente esa capacidad de relacionar información mediante vínculos o enlaces. Estas relaciones además van a permitir a los usuarios una gran flexibilidad en el momento de buscar la información de su interés. Por esto, el modelo Web se planteó ya desde sus inicios como un grafo dirigido. En este grafo, cada página es un

nodo y cada arco representa un enlace entre dos páginas. Estos enlaces no están puestos al azar, tienen una intencionalidad. Las páginas normalmente tienen enlaces hacia otras páginas con el mismo tema. Además, las mejores páginas tienden a ser más referenciadas que lo normal. La web como grafo, tiene una estructura que se puede clasificar como *red libre de escala*. Las redes libres de escala, al contrario de las redes aleatorias, se caracterizan por una distribución dispareja de los enlaces. Estas redes han sido el tema de una serie de estudios, entre los que cabe resaltar por su claridad los de (Barabási y Frangos 2002), y se caracterizan como redes en las cuales la distribución del número de enlaces sigue una ley de potencias (Baldi et al. 2003) y (Alonso Berrocal et al. 2001).

La recuperación de información es el área de la ciencia que nos permite obtener la información necesaria acerca de una materia a partir de una colección de datos. Esto no es lo mismo que recuperación de datos, en la que el contexto de los documentos consiste principalmente en determinar cuál de los documentos de la colección contiene las palabras de la consulta del usuario. El problema que se nos plantea en el web es el de la abundancia de información debido a la explosión documental en la que nos encontramos en la actualidad.

Una solución al problema planteado es el uso de la estructura hipertexto del web, empleando los enlaces entre las páginas, como citas en los mecanismos de la literatura clásica, para encontrar los documentos más importantes. La utilidad de este planteamiento ya fue demostrada por (Alonso Berrocal et al. 1999) y (Alonso Berrocal et al. 2004). Más recientemente se ha valorado este sistema como muy eficaz en el trabajo de (Cothey 2004).

2. Recoger la información del web

Estas páginas web se encuentran instaladas en un servidor web y son servidas ante las peticiones del cliente empleando el protocolo http y visionadas por los visores web. Para poder analizar esta enorme

cantidad de páginas es necesario elaborar programas automáticos que permitan analizar los documentos hipertexto recorriendo toda la red a través de los hiperenlaces que los conectan.

La bibliografía existente sobre este particular es extensa y variada destacando los trabajos de (Thelwall 2001), (Alonso Berrocal et al. 2003), (Chakrabarti 2002), (Castillo y Baeza-Yates 2005) que dan una idea de los mecanismos necesarios para el trabajo con este tipo de herramientas.

Un web crawler es un programa de ordenador que es capaz de recuperar páginas del web, extrayendo los enlaces desde estas páginas y siguiéndolos. Este trabajo de recorrer todas las páginas web recibe el nombre genérico de crawling y los programas desarrollados para hacerlo reciben nombres como crawler, spider, wanderer, robot, bot o recolector.

Hay varias formas de poder hacer éste recorrido del web, aunque básicamente existen tres:

1. Recorrido en anchura (breadth-first).
2. Recorrido en profundidad (depth-first).
3. El mejor posible (best-first).

Para el esquema de el mejor posible la decisión de cuáles son los enlaces a recorrer se toma en función de distintas técnicas, como por ejemplo la utilización del valor del PageRank, para decidir recorrer en primer lugar los que poseen un PageRank mayor y dejar para el final los que posean un PageRank menor.

El procedimiento básico de un robot consiste en suministrar una URL inicial o un conjunto de ellas (semillas), obtener la página web correspondiente y a continuación extraer todos los enlaces existentes en dicha página. Con los enlaces obtenidos es necesario realizar una serie de operaciones previas de normalización entre las que podemos indicar las siguientes: convertir URL a minúscula, eliminar anclas, adecuar el sistema de codificación, emplear la heurística para la determinación de la página por defecto, resolver los URL relativos, etc.

A continuación será necesario comprobar los URL que se habían seguido previamente y en caso de no haberlos recorrido introducirlos en una cola de URL a seguir. Después, normalmente, almacenamos la información, bien en bases de datos o en estructuras de ficheros con codificación ASCII. Finalmente obtenemos el URL del siguiente enlace a seguir y comienza de nuevo el proceso.

3. Representación de la información

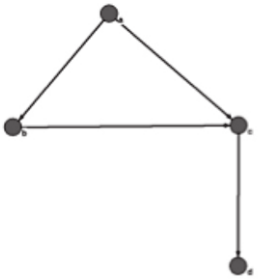
Una vez recogida toda la información con el recolector es necesario procesar toda esa información. Previamente creemos necesario realizar una breve introducción teórica para que se puedan comprender mejor los conceptos utilizados.

Cuando se trabaja con redes, se utiliza una rama de las matemáticas llamada Teoría de grafos para la definición de los conceptos. Aquí tratamos solamente de definir algunos de los conceptos que necesitamos para poder comprender mejor el objeto de nuestro estudio.

Comenzaremos definiendo un grafo como un conjunto de vértices y un conjunto de líneas entre pares de esos vértices.

Este grafo nos permite representar adecuadamente la estructura de una red, donde los vértices o elementos de la red se llaman genéricamente como nodos (siendo la unidad más pequeña de la red) y estos nodos se encuentran comunicados mediante líneas. Estas líneas pueden ser dirigidas, es decir el sentido de la conexión es importante, denominándose arcos; o bien líneas no dirigidas, la conexión indica un sentido bidireccional, que se denominan aristas.

Los grafos dirigidos finitos con *n* nodos, se representan como estructuras de datos por medio de una matriz de adyacencia: una matriz *n-por-n* cuyas entradas en la fila *i* y la columna *j* dan el número de arcos desde el nodo *i*-ésimo al *j*-ésimo. Veamos un ejemplo de esto en la figura 1:



a_{ij}	a	b	c	d
a	0	1	1	0
b	0	0	1	0
c	0	0	0	1
d	0	0	0	0

Figura 1: Matriz de adyacencia de un grafo.

Formalmente definimos la matriz de adyacencia como V = {v_1, v_2,…,v_n} de forma que

$$a_{ij} = \begin{cases} 1 & si\,(v_i, v_j) \in G \\ 0 & en\,otro\,caso \end{cases}$$

Como podemos ver ponemos un 1 cuando existe el enlace y un 0 en caso contrario.

A partir de esa representación en matriz podemos aplicar infinidad de cálculos, muchos de los cuales generan índices y algunos de ellos son los que van a ser empleados en este estudio.

Las estructuras de enlaces, una vez transformadas en grafos y matrices de adyacencia, permiten discernir los patrones estructurales del sitio. De esta forma las estructuras hipertextuales de un sitio web serán diferentes dependiendo de su funcionalidad dentro de la Red. Según sea la función del sitio web éste tendrá una estructura de enlaces determinada que compartirá con otros sitios similares.

Podemos así tener índices o medidas que afectan a toda la red, que afectan a los nodos individualmente, destacando las denominadas medidas de centralidad, o las medidas de posicionamiento, destacando entre ellas el PageRank.

4. Detección de comunidades

Una vez que tenemos la representación del grafo podemos aplicar Análisis de Redes Sociales a nuestra información (Wasserman y Faust 1998). El Análisis de Redes Sociales es una herramienta de medición y análisis estructural permitiendo conocer las interacciones existentes entre los actores de la red analizada (Molina 2001).

Hay un amplio conjunto de indicadores como la densidad, centralidad, centralización, intermediación, cercanía, etc. que nos permiten análisis tanto de nodos como de redes completas, aunque la detección de comunidades, grupos. cliques, etc., es un tema de alto interés.

En el contexto de las redes, al hablar de comunidad nos referimos a un conjunto de nodos de la red que están más densamente conectados entre sí que con el resto de la red.

Existen muchas técnicas para la detección de comunidades (Porter et al. 2009) (Fortunato 2010), como los algoritmos de agrupamiento jerárquico, métodos basados en cliques, agrupamiento por cortes, algoritmo Girvan-Newman,etc.

Un método ampliamente utilizado es el análisis de modularidad (el número de vínculos entre grupos es pequeño, dentro de grupos es alto), destacando el algoritmo Louvain (De Meo et al. 2011).

Un método que se está mostrando eficaz es el algoritmo de agrupamiento VOS y algunos trabajos están demostrando su mayor eficacia frente a otros sistemas, sobre todo da mejor rendimiento que los sistemas basados en la modularidad en la detección de agrupamientos pequeños (Eck 2011).

En una red de entrenamiento, formada por cerca de 36.000 nodos y aproximadamente 476.000 enlaces hemos calculado las comunidades con más de 10 nodos en cada una de ellas y hemos aplicado el algoritmo Louvain y el algoritmo VOS.

Los resultados obtenidos en ambos casos han sido los siguientes:

Louvain. Modularidad de 0,65				
Cluster	**Freq**	**Freq%**	**CumFreq**	**CumFreq%**
1	8142	22,6576	8142	22,6576
2	5370	14,9436	13512	37,6012
3	4444	12,3668	17956	49,9680
4	3893	10,8334	21849	60,8014
5	3683	10,2491	25532	71,0505
6	2391	6,6537	27923	77,7042
7	2386	6,6398	30309	84,3440
8	2373	6,6036	32682	90,9475
9	1969	5,4793	34651	96,4269
10	117	0,3256	34768	96,7525
11	94	0,2616	34862	97,0141
12	62	0,1725	34924	97,1866
13	56	0,1558	34980	97,3424
14	42	0,1169	35022	97,4593
15	19	0,0529	35041	97,5122
16	12	0,0334	35053	97,5456

VOS Clustering. Calidad de 0,85				
Cluster	**Freq**	**Freq%**	**CumFreq**	**CumFreq%**
1	3388	9,4281	3388	9,4281
2	2976	8,2816	6364	17,7098
3	2098	5,8383	8462	23,5481
4	2004	5,5767	10466	29,1248
5	1696	4,7196	12162	33,8444
6	1648	4,5861	13810	38,4305
7	1215	3,3811	15025	41,8116
8	984	2,7383	16009	44,5499
9	953	2,6520	16962	47,2019
10	932	2,5936	17894	49,7955

11	795	2,2123	18689	52,0078
12	788	2,1928	19477	54,2006
13	784	2,1817	20261	56,3824
14	729	2,0287	20990	58,4110
15	655	1,8227	21645	60,2338
16	633	1,7615	22278	61,9953
17	613	1,7059	22891	63,7011
18	607	1,6892	23498	65,3903
19	575	1,6001	24073	66,9904
20	550	1,5305	24623	68,5209
21	520	1,4471	25143	69,9680
22	482	1,3413	25625	71,3093
23	471	1,3107	26096	72,6200
24	466	1,2968	26562	73,9168
25	460	1,2801	27022	75,1969
26	446	1,2411	27468	76,4380
27	436	1,2133	27904	77,6513
28	428	1,1910	28332	78,8424
29	419	1,1660	28751	80,0083
30	413	1,1493	29164	81,1576
31	391	1,0881	29555	82,2457
32	382	1,0630	29937	83,3088
33	365	1,0157	30302	84,3245
34	362	1,0074	30664	85,3318
35	361	1,0046	31025	86,3364
36	327	0,9100	31352	87,2464
37	307	0,8543	31659	88,1007
38	281	0,7820	31940	88,8827
39	258	0,7180	32198	89,6007
40	248	0,6901	32446	90,2908
41	217	0,6039	32663	90,8947
42	210	0,5844	32873	91,4791
43	196	0,5454	33069	92,0245
44	184	0,5120	33253	92,5365
45	180	0,5009	33433	93,0374

46	156	0,4341	33589	93,4715
47	138	0,3840	33727	93,8556
48	124	0,3451	33851	94,2006
49	120	0,3339	33971	94,5346
50	112	0,3117	34083	94,8463
51	90	0,2505	34173	95,0967
52	89	0,2477	34262	95,3444
53	83	0,2310	34345	95,5753
54	76	0,2115	34421	95,7868
55	67	0,1864	34488	95,9733
56	63	0,1753	34551	96,1486
57	59	0,1642	34610	96,3128
58	41	0,1141	34651	96,4269
59	39	0,1085	34690	96,5354
60	31	0,0863	34721	96,6217

El número de comunidades detectadas por el algoritmo VOS es sensiblemente mayor, con un menor número de documentos en cada comunidad y en un primer análisis obtenemos comunidades más coherentes que con el algoritmo Louvain.

5. Visualización

La Web presenta un dinamismo increíble, tanto en estructura (cada día se observan cambios en sitios web, páginas que desaparecen, restructuraciones, nuevas páginas y secciones, etc.) como en contenido (cada día nuevos sitios web son agregados) (Medrano y Alonso Berrocal 2011). Por ello autores como (Fry 2000) se preguntan: ¿Cómo podría representarse la estructura de Internet, que se encuentra en constante cambio?" Es casi imposible dar respuesta a estas y muchas preguntas relacionadas sin pensar en técnicas de Visualización de Información dinámica.

El rápido y descontrolado crecimiento de la información que reside en la Web, ha conllevado al surgimiento de numerosos campos

de investigación enfocados en la búsqueda de herramientas de análisis que permitan representar y comprender vastas cantidades de información. Actualmente existen sistemas capaces de almacenar grandes volúmenes de datos. Sin embargo, la representación textual de los datos no permite realizar exploraciones adecuadas sobre los mismos, ocasionando de esta forma la pérdida del potencial de la información almacenada (Keim 2002).

Como consecuencia de esto, han surgido áreas de investigación como la Analítica Visual, la cual se define como la ciencia del razonamiento analítico con ayuda de interfaces visuales altamente interactivas, o la Visualización de Información definida como el proceso de pasar de representaciones gráficas a representaciones perceptivas, eligiendo las técnicas de codificación que maximicen la comprensión humana y la comunicación. El enfoque de la exploración de datos a través de la visualización busca combinar flexibilidad, creatividad y conocimiento general con grandes volúmenes de datos almacenados, a fin de facilitar la interacción directa con la información a través de la extracción de conocimientos y la realización de análisis y conclusiones.

Existen numerosas técnicas de visualización, empleando diferentes algoritmos como el Kamada-Kawai, Fruchterman-Reingold, escalado multidimensional (MDS), etc.

En relación con la técnica de agrupamiento VOS, podemos emplear la visualización VOS mapping que se muestra muy eficaz frente al MDS (Van Eck et al. 2010). En este mapa los colores indican la densidad dentro de cada comunidad, variando desde el azul (densidad más baja) hasta el rojo (densidad más alta). Nos permite ver las comunidades más importantes y puestas en relación unas con otras.

Analizadas las comunidades en rojo podemos observar la fuerte correlación existente entre ellas y que la proximidad visual se corresponde que la cercanía entre las comunidades detectadas

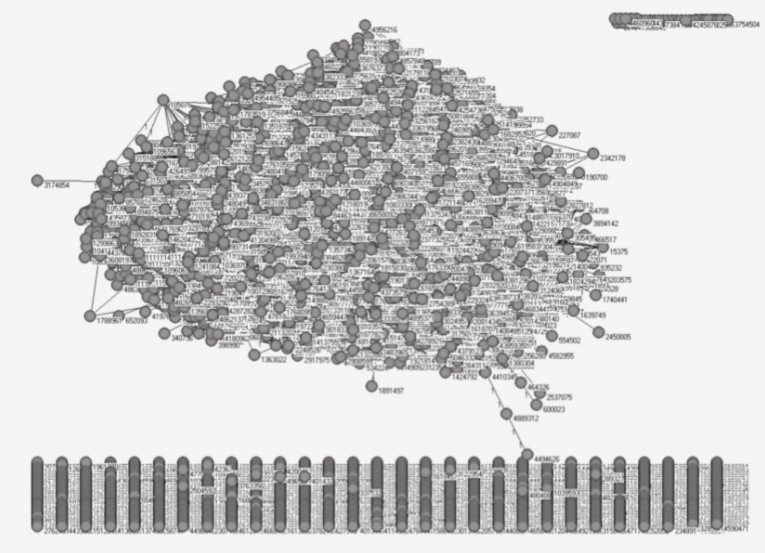


Figura 2: Visualización Fruchterman-Reingold del grafo completo.

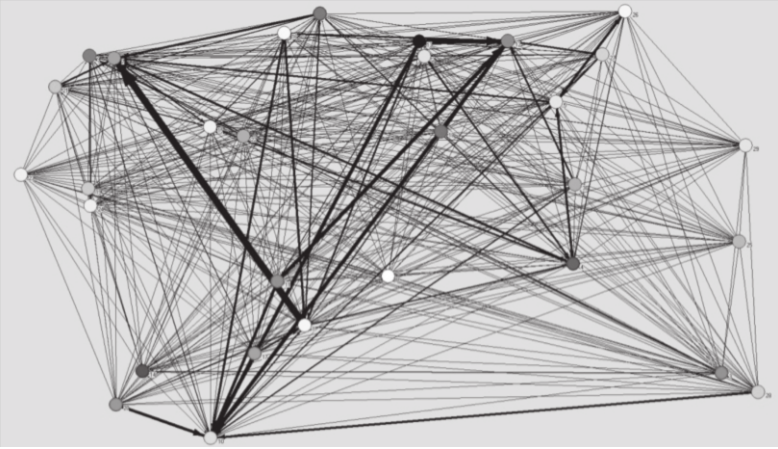


Figura 3: Visualización Fruchterman-Reingold de 30 comunidades.

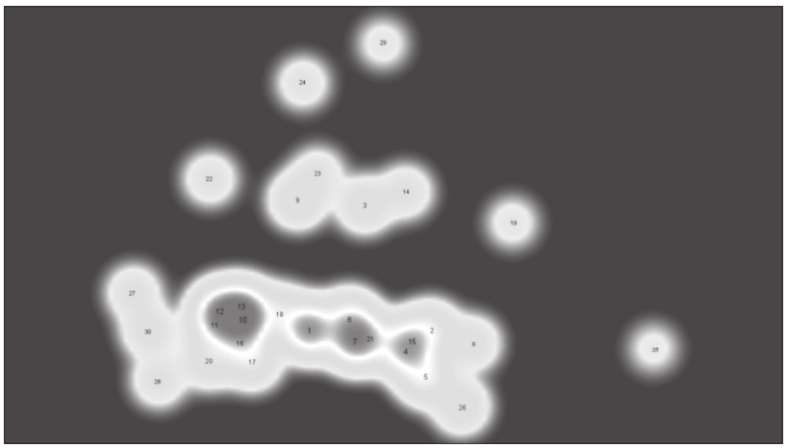


Figura 4: VOS mapping de 30 comunidades.

6. Conclusiones

Los mecanismos de recogida y representación de la información web están muy bien definidos desde hace muchos años y con una gran cantidad de investigaciones, en infinidad de ámbitos, que han demostrado la eficacia de los mecanismos de recogida y la aplicación de la teoría de grafos para su representación.

La utilización del Análisis de Redes Sociales, ha permitido aplicar un elevado número de índices, y han permitido caracterizar la información web de forma adecuada. La aplicación de las técnicas de detección de comunidades han facilitado la comprensión de los flujos internos de información, y facilitando la reducción del grafo. Las nuevas técnicas de detección son eficaces desde el punto de vista algorítmico y con unos índices de efectividad altos.

La visualización de grafos complejos es siempre problemática, pero aplicando la técnica del VOS clustering y del VOS mapping, obtenemos representaciones muy acertadas de la realidad de las comunidades obtenidas y se ponen de manifiesto las relaciones existentes de una forma novedosa y muy cercana a la realidad.

7. Bibliografía

ALONSO BERROCAL, J. L., C. G. Figuerola, et al. (2003). "Agentes inteligentes: recuperación automática de información en la web." *Revista española de documentación científica* 26(1): 11-20.

ALONSO BERROCAL, J. L., C. G. Figuerola, et al. (2004). *Cibermetría: nuevas técnicas de estudio aplicables al web*. Gijón, Ed. Trea.

ALONSO BERROCAL, J. L., C. G. Figuerola, et al. (2001). "Cibermetría del Web: Las leyes de exponenciación." *Revista general de información y documentación* 11(1): 201-209.

ALONSO BERROCAL, J. L., C. G. Figuerola, et al. (1999). "Representación de páginas web a través de sus enlaces y su aplicación a la recuperación de información." *Scire: representación y organización del conocimiento* 5(2): 91-98.

BAEZA-YATES, R., C. Castillo, et al. (2005). "Characteristics of the Web of Spain." *Cybermetrics* 9(1).

BALDI, P., P. Frasconi, et al. (2003). *Modeling the Internet and the Web*. Chichester, Wiley.

BARABÁSI, A. L. and J. Frangos (2002). *Linked: The New Science Of Networks*. [s.l], Perseus Publishing.

CASTILLO, C. and R. Baeza-Yates (2005). *Effective web crawling*. ACM SIGIR Forum.

COTHEY, V. (2004). "Web-crawling reliability." *Journal of the American Society for Information Science and Technology* 55(14): 1228-1238.

CHAKRABARTI, S. (2002). *Mining the Web: Discovering knowledge from hypertext data*. San Francisco, CA, Morgan Kaufmann.

DE MEO, P., E. Ferrara, et al. (2011). *Generalized louvain method for community detection in large networks.* Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on, IEEE.

ECK, N. J. P. (2011). *Methodological Advances in Bibliometric Mapping of Science*. Rotterdam, The Netherlands, Erasmus University Rotterdam.