Artículo

Análisis de co–enlaces: una aproximación teórica

Por **José Luis Ortega** e **Isidro Aguillo**

José Luis Ortega Priego *es licenciado en documentación en 1999 por la Universidad de Granada y actualmente cursa doctorado en documentación en la Universidad Carlos III de Madrid. Forma parte del Grupo de Investigación de Cibermetría del Centro de Información y Documentación Científica (CSIC) en el ámbito de la cibermetría, minería web, visualización de información y usabilidad.*

Isidro Aguillo *trabaja en el Grupo de Investigación en Cibermetría del Cindoc–CSIC realizando estudios sobre indicadores web, revistas electrónicas y posicionamiento en motores de búsqueda. Es el editor de la revista electrónica Cybermetrics y coordina el Webometrics Ranking of World Universities. Es licenciado en zoología, practicando de forma activa la ornitología, y master en información y documentación por la Universidad Carlos III.*

Resumen: *Se hace un recorrido teórico y práctico por el análisis de co–enlaces, una de las técnicas clásicas de la visualización de información para la construcción de mapas de la ciencia en el ámbito de la cibermetría. Se analizan los pasos a seguir, la fuentes de datos más idóneas, los diferentes tipos de matrices y como ha de procesarse cada una. Los resultados nos permiten diferenciar los distintos pasos seguidos y los diferentes modos de procesamiento. Finalmente, se concluye que las matrices asimétricas aportan más información que las simétricas, el coseno tiene un poder de discriminación mayor que la correlación y los crawlers obtienen información más apropiada que los buscadores.*

Palabras clave: *Cibermetría, Análisis de co–enlaces, Visualización de información, Mapas de la ciencia, Escalamiento multidimensional, Buscadores.*

Title: Co–link analysis: a theoretical approach

Abstract: *This work provides a theoretical and practical walk–through of Co–link Analysis, one of the classic techniques for information visualisation that serves as a tool for constructing Maps of Science in Webometrics. An analysis is offered of the steps to be followed, the most suitable data resources, the different kinds of matrices, and how these have to be processed. The results permit distinguishing between the different steps taken and the differences in processing these matrices. The study finds that asymmetrical matrices provide more information than symmetrical ones, that there is greater discrimination with the cosine than with the correlation, and that crawlers obtain more accurate information than search engines.*

Keywords: *Webometrics, Co–link analysis, Information visualisation, Maps of science, Multidimensional scaling, Search engines.*

Ortega, José Luis; **Aguillo, Isidro**. «Análisis de co–enlaces: una aproximación teórica». En: *El profesional de la información*, 2006, julio–agosto, v. 15, n. 4, pp. 270–277.

Introducción

El análisis de co–enlaces o co–sitas (**Faba–Pérez**, et al., 2004; **Herrero–Solana**; **Morales del Castillo**, 2004) es una técnica de visualización de información que permite la construcción de mapas de la ciencia a través del patrón de co–ocurrencia de enlaces en un determinado ámbito del espacio web. Su origen, como muchas otras técnicas cibermétricas, proviene del campo de la bibliometría. En los 70, **Small** propuso la posibilidad de medir las relaciones científicas a través del grado de co–ocurrencia de palabras (**Small**; **Griffith**, 1974), co–ocurrencia de documentos (**Small**, 1978) o co–ocurrencia de autores (**White**; **Griffith**, 1981). Asumió que "si las palabras aparecen juntas, o co–ocurren, en múltiples documentos, entonces la comunidad de autores probablemente vería alguna conexión lógica entre ellos" (**Small**, 2003). Este grado de co–ocurrencia permitía agrupar distintos elementos (autores, artículos, publicaciones, etc.) a fin de representar gráficamente frentes científicos, su evolución en el tiempo y entidades prominentes en un determinado campo. Fue **McCain** (1990) quien formalizó lo que actualmente se conoce como *Author co–citation*

Artículo recibido el 10–03–06
Aceptación definitiva: 06–04–06

analysis (*ACA*). Esta técnica permitía agrupar a diferentes autores en función del grado de co–ocurrencia en un determinado corpus de citas bibliográficas.

En cibermetría los mapas de la ciencia han sido aplicados desde los inicios a través de análisis de *sitations* o enlaces externos o entrantes (**Prime**; **Bassecoulard**; **Zitt**, 2002; **Rousseau**, 1997). El carácter interconectado de la Web y la gran cantidad de información que se puede obtener han atraído el uso de estas técnicas de visualización. En este sentido, **Larson** (1996) estudió el grado de co–ocurrencias de enlaces externos en páginas de astronomía y ciencias de la tierra usando el buscador *AltaVista* como fuente de datos. Como resultado obtuvo distintos grupos de páginas relacionadas por áreas especializadas dentro de esta disciplina. **Boudourides** (2000) también desarrolló mapas basados en co–enlaces de distintas sedes web que participaban en el proyecto *Soeis*, detectando patrones geográficos en la distribución espacial, mientras que **Vaughan** (**Vaughan**; **Wu**, 2004; **Vaughan**; **You**, 2005) aplicó el análisis de co–enlaces al campo de la inteligencia competitiva.

«El cálculo de similitud no pretende otra cosa que la transformación de las matrices asimétricas en simétricas, donde las distancias quedan representadas claramente»

El objetivo de este trabajo es hacer un recorrido por las posibilidades y problemáticas del uso de esta técnica para la construcción de mapas de la ciencia a través de datos web. Se mostrarán los pasos a seguir y las distintas formas existentes para su construcción. Además se discutirá la idoneidad de algunos procedimientos en función del tipo de análisis que se desea realizar. Para la consecución de estos objetivos se analizarán como ejemplo ocho universidades españolas, cuatro madrileñas y cuatro catalanas; se mostrarán los resultados que se obtienen según un proceso u otro.

Fuente de datos

El análisis de co–enlaces, como la mayoría de las técnicas cibermétricas, se basan en dos fuentes de datos diferentes: buscadores y *crawlers* o robots.

1. Buscadores

Su uso como fuente de obtención de datos se produjo en los primeros años del desarrollo de la cibermetría. El trabajo de **Rodríguez Gairín** (1997) ya anunciaba la posibilidad de utilizar *Altavista* como un *Citation Index* y a continuación una gran cantidad de trabajos se nutrieron de los datos aportados por *Altavista* (**Larson**, 1996; **Boudourides**, et al., 2000), *AllTheWeb* o *Yahoo! Search* (**Ortega Priego**; **Aguillo**, 2005). Sin embargo, diversos análisis posteriores cuestionaron el uso de los buscadores como fuente de datos ya que presentaban inconsistencias (**Rousseau**, 1998; **Bar–Ilan**, 1998). El orden de diversos operadores alteraba los resultados, además de variar en cortos espacios de tiempo (**Thelwall**, 2001). Otros estudios han encontrado problemas en la cobertura de escrituras no latinas (**Bar–Ilan**, 2005). Por todo ello, su uso como fuente de datos ha de realizarse teniendo en cuenta este hecho. Lo aconsejable, es la repetición de las consultas en dos momentos diferentes con la intención de detectar fallos en las respuestas y trabajar con el máximo de los resultados obtenidos. Sin embargo, pese a las desventajas, los buscadores permiten estudiar grandes poblaciones y contar con grandes repositorios de páginas y enlaces entre sí con poco esfuerzo.

Otro problema es que no en todos los buscadores es posible obtener esta información, esto es, el número de enlaces de una sede web a otra (**Aguillo**; **Arroyo**; **Ortega**; **Pareja**; **Prieto**, 2005). Por el momento *Yahoo! Search* y *MSN* nos la ofrecen a través de los mismos algoritmos de búsqueda:

—*+site:{dominioA} +linkdomain:{dominioB}*: el número de páginas indizadas en *MSN* y/o *Yahoo! Search* del dominio A (por ejemplo, *ucm.es*), que enlazan con el dominio B (por ej., *ub.es*).

—*+linkdomain:{dominioA} +linkdomain:{dominioB} –site: {dominioA} –site: {dominioB}*: número de páginas indizadas en *MSN* y/o *Yahoo! Search*, a excepción de las páginas de los propios dominios, que poseen un enlace al dominio A (*ucm.es*) y al dominio B (*ub.es*).

2. *Crawlers* (robots)

A medida que la cibermetría se consolidaba fueron apareciendo robots *ad hoc* como *Blinker* (**Cothey**, 2004) o *SocSciBot*, y aplicaciones comerciales (*Wikipedia*, 2006) que permitían obtener directamente información sobre un determinado grupo de páginas seleccionadas previamente. De esta forma, a través de los robots es posible obtener información de primera mano y con los criterios propios que la investigación requiere. Sin embargo, los robots plantean problemas de diseño y de tiempo de extracción de datos además de la forma en que los gestionan y el procedimiento que siguen a la hora de extraerlos (**Chakrabarti**. 2002). **Arroyo** (2004) ya advirtió este hecho al comparar las funcionalidades de diversos robots comercia-

	ucm.es	uam.es	uc3m.es	upm.es	ub.es	uab.es	upf.es	upc.es
ucm.es	907.000							
uam.es	18.400	216.000						
uc3m.es	11.100	8.310	191.000					
upm.es	12.400	9.790	8.000	413.000				
ub.es	19.300	12.900	3.100	8.100	405.000			
uab.es	15.100	11.200	3.120	7.880	22.600	309.000		
upf.es	8.330	2.620	2.400	2.340	11.700	12.200	217.000	
upc.es	9.200	3.270	2.980	11.400	14.800	15.000	9.440	347.000

Tabla 1. Matriz bruta simétrica obtenida de Yahoo! Search (06–03–06)

les y los resultados finales obtenidos al testar un grupo de páginas determinadas.

Matriz de co–enlaces

Una vez seleccionada la fuente de datos se realiza el proceso de construcción de la matriz de co–enlaces, que puede ser simétrica o asimétrica. La diferencia estriba en que la primera contiene los co–enlaces en sí entre dos nodos o sedes web. De esta forma la matriz es simétrica ya que su mitad inferior es idéntica a la superior, y por lo tanto contiene la misma información. Para su construcción se utiliza el siguiente algoritmo:

+linkdomain:{dominioA} +linkdomain:{dominioB} –site: {dominioA} –site: {dominioB}

La tabla 1 muestra una matriz simétrica de ocho dominios universitarios españoles. Se aprecia que la diagonal contiene el total de enlaces que recibe una sede web en el conjunto de la base de datos. También se aprecia el redondeo que el buscador ha realizado.

Sin embargo, la matriz asimétrica no recoge los co–enlaces directamente, sino tuplas de enlaces de una sede web hacia otra directamente. La tabla 2 muestra la matriz asimétrica de los ocho dominios anteriores. En ella se computa el total de enlaces que parten de una sede hacia otra. Las columnas recogen las sedes y sus enlaces salientes, mientras que las filas incluyen las sedes y sus enlaces entrantes. La diagonal recoge los enlaces internos de la propia sede. El algoritmo en este caso es:

+site:{dominioA} +linkdomain:{dominioB}

Normalización

Una vez que se ha obtenido la matriz cruda o bruta se realiza un proceso de normalización, esto es, procesar los datos para facilitar la comparación entre ellos y evitar otros factores ajenos a la investigación que puedan interferir en su análisis. **Leydesdorff** y **Vaughan** (2006) afirman que en matrices simétricas no era necesario ninguna normalización con datos bibliométricos. Sin embargo, **Vaughan** y **You** (2005) aportan una normalización en el caso de datos cibermétricos, basado en el índice de **Jaccard** (1912). El *Cálculo normalizado de co–enlaces* o *Normalized colink count* (*NCC*) se obtendría del total de co–enlaces entre la sede A y B partido entre la suma de coenlaces de A más la suma de los co–enlaces de B menos los co–enlaces de A y B.

Enlace/sede	ucm.es	uam.es	uc3m.es	upm.es	ub.es	uab.es	upf.es	upc.es
ucm.es	341.000	413	306	254	470	224	98	78
uam.es	657	89.700	165	374	191	153	68	71
uc3m.es	234	144	111.000	267	73	62	69	60
upm.es	226	158	270	207.000	131	85	56	233
ub.es	319	193	59	100	147.000	694	251	389
uab.es	224	118	60	105	625	133.000	207	395
upf.es	104	59	40	40	259	180	79.500	121
upc.es	188	58	173	307	498	387	284	245.000

Tabla 2. Matriz asimétrica obtenida de Yahoo! Search (06–03–06)

$$NCC_{ij} = \frac{C_{ij}}{(C_i + C_j - C_{ij})}$$

Por otra parte **Moya–Anegón** (2005) sugiere otra medida de normalización, la *Medida de cocitación normalizada* o *Normalized cocitation measurement*, siendo:

$$NCM_{ij} = \frac{C_{ij}}{\sqrt{C_i\, C_j}}$$

Para las matrices asimétricas, **Ortega** (en prensa) aportó una normalización sustentada en el cálculo de pesos utilizados en los sistemas de recuperación de información basados en modelos vectoriales. Se trata de una adaptación del *IDF/TF* (**Spark–Jones**, 1972; **Larson**; **Hearst**, 1998), donde el peso $w_{(ik)}$ de un enlaces es:

$w_{(ik)} = of_{(ik)} * Log$ *(total de sedes web / número de sedes que hacen un enlace a la sede* (S_k)*)*

En este caso la frecuencia de enlaces externos (*outlink frequency*) $(of_{(ik)})$ es:

$of_{(ik)}$ = número de enlaces externos desde la sede (S_i) a la sede (S_k)/total de enlaces externos de la sede (S_i)

Sin embargo esta medida es válida cuando la matriz es mucho más grande y no todas las sedes son siempre enlazadas por otras ya que, cuando todas las sedes reciben enlaces, sus pesos son 0. En nuestro ejemplo no se podría utilizar porque todas las sedes reciben enlaces.

Comparación

Una vez normalizadas las matrices, siempre y cuando se pueda realizar, medimos la similitud de las distintas sedes. La similaridad trata de medir la semejanza de dos variables acorde a los valores que adquieren y se expresa en un rango de [0,1], donde [1] expresa la semejanza total y [0] la disimilaridad total. En el caso de las matrices simétricas no es necesario utilizar ninguna medida de similaridad, ya que la similitud de dos sedes web ya viene expresada en la matriz por el número de co–enlaces (**Leydesdorff**; **Vaughan**, en prensa). Cuanto más sean co–enlazadas dos sedes más próximas se hallarán entre sí.

Pero en el caso de las matrices asimétricas sí es necesario calcular una medida de similitud que compare los enlaces que reciben las distintas sedes en el conjunto de la matriz. Así, si dos sedes son enlazadas por una tercera se encontrarán más próximas que aquellas que no lo son.

Existen distintas medidas de similitud que pueden ser usadas, como el *Coseno de Salton* (**Salton**; **Wong**; **Yang**, 1975; **Salton**, 1971); el *Coeficiente de Jaccard* (**Jaccard**, 1912; **Rorvig**, 1999), el *Coeficiente de correlación de Pearson* (**White**, 2003) o *Chi–cuadrado* (**Ahlgren**; **Jarneving**; **Rousseau**, 2003). Pero, en este caso nos centraremos en las dos más usadas: el *Coseno de Salton* y el *Coeficiente de correlación de Pearson*.

«El uso de la matriz simétrica provoca una gran distorsión en los resultados debido al ruido que generan todos los enlaces que indizan los buscadores»

Ahlgren (2003) mostró que el *Coeficiente de correlación de Pearson* no era una medida adecuada para el *ACA* puesto que es sensible al número de ceros y a la adición de nuevas variables. Sin embargo, **White** (2003) argumentó que esta incoherencia no afectaba a los resultados finales del *ACA*, la agrupación de datos y la representación gráfica. **Bensman** (2004) expresó que los argumentos del mencionado artículo son demasiados teóricos y no son significativos en el mundo real. **Ahlgren** propuso distintas medidas alternativas (*coseno*, *chi–cuadrado*) mientras **Leydesdorff** (2005) sugirió una teoría de la información como una metodología aceptable para calcular la similaridad.

El *coseno* puede ser definido como el coseno del ángulo entre dos vectores. Esta medida no es sensible al número de ceros y además no se basa en la media de la distribución. La fórmula es:

$$Sim(S_i, S_j) = \frac{\sum e_i e_j}{\sqrt{\sum (e_i)^2 \sum (e_j)^2}}$$

donde S_i y S_j son dos sedes web, y (e) es el número de enlaces a otras sedes web.

El *Coeficiente de correlación de Pearson* (*–r*) no es estrictamente una medida de similitud sino que describe la fuerza y la dirección de una relación lineal entre dos variables *X* e *Y*, definiéndose como:

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 (Y - \overline{Y})^2}}$$

donde $\overline{X}$ e $\overline{Y}$ son las medias de las dos variables.

Finalmente, la matriz de similaridad se transforma en una matriz de distancias, siendo para el coseno:

$\delta(S_i, S_j)=1–Sim(S_i, S_j)$

y para la correlación

$\delta(–r)= 1–(–r)$

Como se ha podido ver, el cálculo de similitud no pretende otra cosa que la transformación de las matrices asimétricas en simétricas, en la que las distancias quedan representadas claramente. De esta forma ya se puede pasar a la visualización de los datos.

Visualización

Una vez calculadas las distancias entre las distintas sedes web que componen la matriz, se procede a la representación gráfica. Tradicionalmente el análisis de co–enlaces ha utilizado el escalamiento multidimensional (*Multidimensional scaling*, *MDS*) como la principal vía para la representación de este tipo de datos. Consiste en "un conjunto de técnicas de análisis de datos cuya finalidad es mostrar la distancia de los datos a través de una representación geométrica" (**Young**, 1985). Su origen está en los estudios de **Torgerson** en el *MDS* métrico (**Torgerson**, 1952) y **Kruskal** en el no–métrico (**Kruskal**, 1964). Mediante este algoritmo se reduce un espacio vectorial de n–dimensiones a otro de 2 ó 3 dimensiones, lo cual permite la representación gráfica de estos vectores y ver su posicionamiento en el espacio. El *MDS* genera un mapa de puntos a partir de una matriz de distancias, y al margen de error entre el mapa calculado y las distancias reales se le denomina *stress* (φ). Cuanto mayor sea el grado de *stress* más distorsión habrá entre los valores observados y los calculados. En torno a 0,0 y 0,2 son niveles aceptables para un buen modelo (**Conchillo Jiménez**; **Ruiz Gallego–Largo**, 1993).

Interpretación

Una técnica que posibilita interpretar los resultados obtenidos es la agrupación (*clustering*), ya que agrupa los casos estudiados en función de sus similitudes. Existen dos métodos diferenciados: los jerárquicos y los partitivos (**Sneath**; **Sokal**, 1973), pero en este tipo de análisis, la agrupación jerárquica acumulativa (*agglomerative hierarchical clustering*, *AHC*) es la más común. Esta técnica posee múltiples procedimientos de agregación: *Método de Ward* (*Ward's method*), *Vecino más cercano* (*Single link*), *Vecino más lejano* (*Complete link*), etc., siendo el último de ellos el más apropiado por su poder de discriminación y de diferenciar poblaciones heterogéneas. El resultado es un dendrograma que muestra relaciones jerárquicas entre objetos en forma de un árbol de igual longitud en sus ramas.

Resultados

Fueron obtenidos con el paquete estadístico *SPSS 13.0* y presentados gráficamente con el programa de redes *Ucinet 6.1* y *NetDraw 2.28*. El tamaño de los nodos representa el número de páginas obtenidos de *Google* el 6 de mayo de 2006 y el color representa a la región a la que pertenecen: amarillo para las catalanas y rojo para las madrileñas.

http://www.spss.com
http://www.analytictech.com

Las figuras 1 y 2 representan los mapas obtenidos a partir de la matriz asimétrica convertida a distancias a través de las dos medidas de similitud vistas antes: el *coseno* y el *coeficiente de correlación*. Ambas poseen un buen ajuste, siendo φ=0,026 para la *correlación* y φ=0,035 para el *coseno*. En ambos modelos no se aprecian diferencias significativas. Las universidades están agrupadas por su relación geográfica, siendo sus diferencias o distancias bastantes acentuadas en los dos modelos. Se aprecia que aquellas con una mayor cantidad de páginas se encuentran fuertemente conectadas entre sí. Sin embargo, en el *coseno* se aprecia que las distancias y las proximidades son más acentuadas que en la correlación, y los nodos se acercan o se repelen con mayor fuerza.

La figura 4 muestra el mapa obtenido a través de la matriz simétrica bruta, sin ninguna transformación, mientras que en la figura 5 se observa la matriz simétrica bruta con la normalización realizada. Notar que

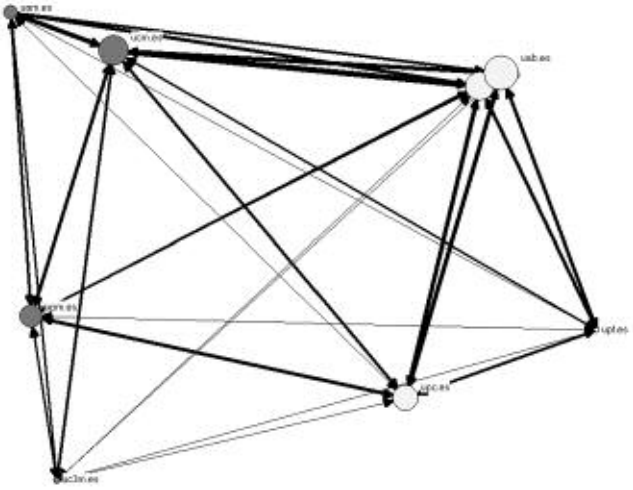

Figura 1. Correlación (stress = 0,026)

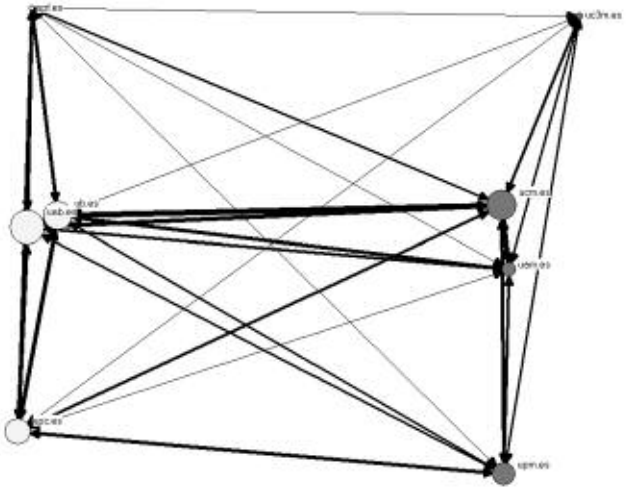

Figura 2. Coseno (stress = 0,035)

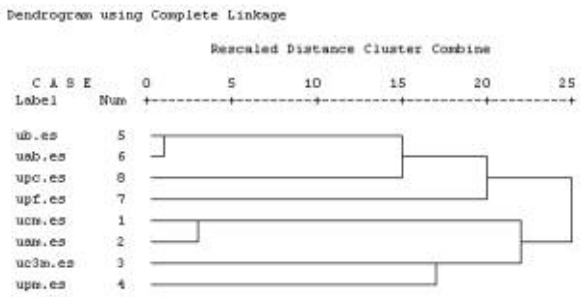

Figura 3. Dendograma de la matriz asimétrica (método: vecino más lejano)

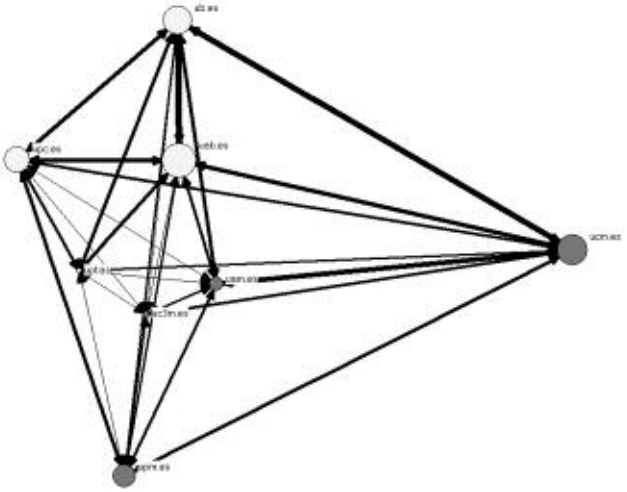

Figura 4. Bruta (stress = 0,037)

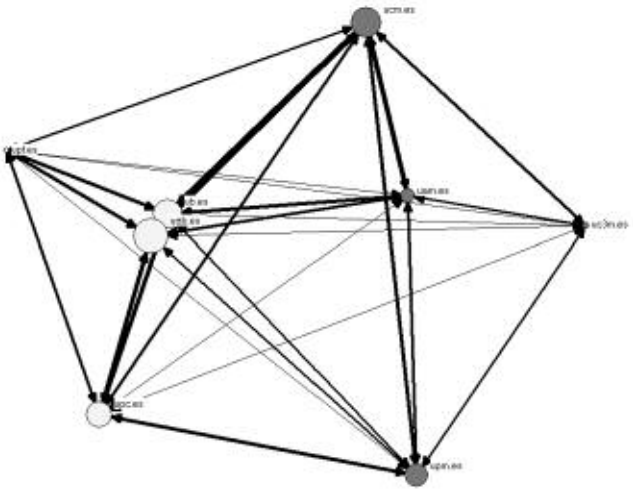

Figura 5. Bruta normalizada (stress = 0,019)

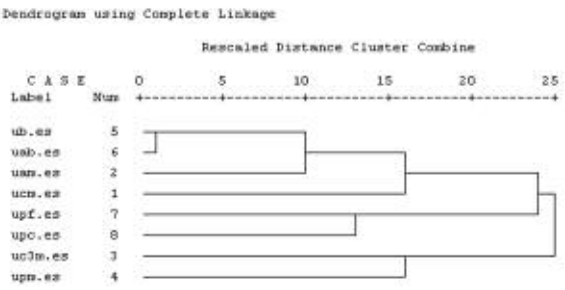

Figura 6. Dendograma (método: vecino más lejano)

tanto la *NCM* como la *NCC* producen el mismo resultado. Se aprecia claramente que la normalización ayuda a detectar los nodos más próximos y a tener una mejor comprensión de los modelos, además de conseguir la reducción del stress (φ=0,037 para la bruta no normalizada y φ=0,019 para la normalizada). Como se ve, en la matriz no normalizada las distancias apenas son significativas, y quitando los nodos *ucm.es* y *upm.es* todos los demás están a la misma distancia.

Respecto a los resultados obtenidos entre las matrices asimétricas y no asimétricas se notan claras diferencias. En las primeras el mapa resultante tiene una mayor discriminación y las relaciones son más explícitas, lo que permite una buena clasificación de los casos y por lo tanto una mejor interpretación del modelo. Esta diferencia también es significativa en los dendogramas. En el primero, los dos grupos quedan bien diferenciados y el dendograma se ajusta a la represen-

tación gráfica. Sin embargo, en el segundo los dos grupos se difuminan y las relaciones resultantes no quedan claras.

Discusión

Una de las principales diferencias entre una matriz asimétrica y simétrica estriba en la computación de los co–enlaces. Una matriz simétrica computa los co–enlaces en referencia a todo el corpus utilizado, en nuestro caso toda la base de datos de *Yahoo! Search*, con billones y billones de enlaces. Sin embargo, la matriz asimétrica computa sus co–enlaces sobre la co–ocurrencia de los enlaces extraídos del conjunto de nodos estudiados. De esta forma la matriz simétrica muestra la relación de dos sedes web acorde a billones de páginas en toda la Web, mientras que en la asimétrica, la relación se establece a partir de las ocho sedes web seleccionadas. Así pues, el uso de la matriz simétrica provoca una gran distorsión en los resultados debido al ruido que generan los enlaces que indizan los buscadores. Esto dificulta la interpretación de los resultados y complica la detección de otros patrones significativos en los datos, además de generar resultados demasiados obvios.

Por otro lado el uso de los valores de la diagonal en el modelo asimétrico ocasiona grandes problemas en la representación, ya que ésta contiene todos los enlaces del dominio A al dominio A, o sea, los vínculos internos. En el caso del *ACA*, el uso de la diagonal es interesante por que ésta contiene todas las autocitas de un autor, pero en el análisis de co–enlaces la desproporción entre enlaces salientes e internos es tan grande que distorsiona enormemente los resultados (**Ortega**; **Aguillo**; **Prieto**, 2006).

Conclusión

Como hemos podido comprobar, es más aconsejable trabajar con matrices asimétricas que con simétricas, ya que las primeras expresan la co–ocurrencia a partir de los enlaces que se generan dentro del grupo de análisis y no en el conjunto de la base de datos o del repositorio de donde se extraen los datos. Además, los modelos resultantes permiten optar por las matrices asimétricas como la forma más optima de desarrollar análisis de co–enlaces.

El *coseno* provoca una mayor apreciación de las distancias y favorece en mayor medida las similitudes que en el caso de la correlación, aunque estas diferencias no son muy significativas. De esta forma, el *coseno* posee una mayor capacidad discriminatoria que en grandes análisis permiten diferencias grandes cantidades de casos.

La elección de distintas fuentes de obtención de datos debe realizarse en función de los medios que se posean. La forma más fiable es el uso de un *crawler* o robot ya que puede extraer la información precisa para el estudio además de no atenerse a los problemas de sesgo y cobertura de los buscadores. Desde nuestro punto de vista, los robots son aconsejables en estudios de gran envergadura donde las limitaciones de los buscadores son más apreciables. Sin embargo, para estudios pequeños como el nuestro, los buscadores son una fuente rápida de datos y los sesgos apenas afectan al modelo.

«Para estudios pequeños los buscadores son una fuente rápida de datos y los sesgos apenas afectan al modelo»

El uso de técnicas de agrupamiento ayuda a la interpretación y agrupación de los casos en función de sus similitudes. En nuestro caso, el agrupamiento ha actuado como complemento a la información suministrada por el mapa, pero en modelos con un alto número de casos la técnica de agrupamiento se hace altamente necesaria para la interpretación.

Pese a todo, estas apreciaciones dependerán del tipo de análisis que vayamos a realizar, de la cantidad y tipología de casos a analizar, y la significación del corpus que se utilice para computar los co–enlaces.

Bibliografía

1. **Aguillo, I.**; **Arroyo, N.**; **Ortega, J. L.**; **Pareja, V.**; **Prieto, J. A.** «Análisis cibermétrico de los principales motores de búsqueda». En: *Fesabid, 9as. Jornadas españolas de documentación*, 2005.

2. **Ahlgren, P.**; **Jarneving, B.**; **Rousseau, R.** «Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient». En: *Journal of the American Society for Information Science and Technology*, 2003, v. 54, n. 6, pp. 550–560.

3. **Arroyo, N.** *Métodos y herramientas para la extracción de datos en cibermetría. El software académico y comercial*. Universidad de Salamanca: 2004.

4. **Bar–Ilan, J.** «Search engine results over time – a case study on search engine». En: *Cybermetrics*, 1998, v. 1, paper 1.
http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

5. **Bar–Ilan, J.** «Expectations versus reality – search engine features needed for Web research at mid 2005». En: *Cybermetrics*, 2005, v. 9, p. 1.
http://cybermetrics.cindoc.csic.es/cybermetrics/articles/v9i1p2.html

6. **Bensman, S. J.** «Pearson's R and author cocitation analysis: a commentary on the controversy». En: *Journal of the American Society for Information Science and Technology*, 2004, v. 55, n. 10, pp. 935.

7. **Boudourides, M.**; **Sigrist, B.**; **Alevizos, Ph.** «Webometrics and the self–organization of the European information society». En: *Triple helix conference*, 2000.
http://hyperion.math.upatras.gr/webometrics/

8. **Chakrabarti, S.** *Mining the Web: discovering knowledge from hypertext data*. San Francisco: Morgan–Kaufmann Publishers, 2002.

9. **Conchillo Jiménez, A.**; **Ruiz Gallego–Largo, T.** «Escalamiento multidimensional: una metodología de análisis en el campo de los factores humanos». En: *Boletín digital FH*, 1993, v. 2.
http://www.tid.es/presencia/boletin/boletin2/art003.htm

10. **Cothey, V.** «Web–crawling reliability». En: *Journal of the American Society for Information Science and Technology*, 2004, v. 55, n. 14, pp. 1.228–1.238.

11. **Faba–Pérez, C.**; **Guerrero–Bote, V. P.**; **Moya Anegón, F. de**. *Fundamentos y técnicas cibermétricas*. Mérida: Junta de Extremadura, 2004.

12. **Herrero–Solana, V.**; **Morales del Castillo, J.** "Mapas geopolíticos de internet: aplicación de las nuevas técnicas de representación de la información". En: *Ciência da informação*, 2004, v. 33, n. 3.

13. **Jaccard, P.** «The distribution of flora in the Alpine zone». En: *The new phytologist*, 1912, v. 11, n. 2, pp. 37–50.

14. **Kruskal, J. B.** «Nonmetric multidimensional scaling». En: *Psychometrika*, 1964, v. 29, n. 1–27, pp. 115–129.

15. **Larson, R.** «Bibliometrics of the world wide web: an exploratory analysis of the intellectual structure of cyberspace». En: *Proceedings of ASIS96*, 1996, pp. 71–78.
http://sherlock.berkeley.edu/asis96/asis96.html

16. **Larson, R. R.**; **Hearst, M.** *Term weighting and ranking algorithms*. Consultado en: 17–05–05.
http://www.sims.berkeley.edu/courses/is202/f98/Lecture17/index.htm

17. **Leydesdorff, L.** «Similarity measures, author cocitation analysis, and information theory». En: *Journal of the American Society for Information Science and Technology*, 2005, v. 56, n. 7, pp. 769–772.

18. **Leydesdorff, L.**; **Vaughan, L. W.** "Co–occurrence matrices and their applications in information science: extending ACA to the Web environment". En: *Journal of the American Society for Information Science and Technology*, 2006 (en prensa).

19. **McCain, K.** «Mapping authors in intellectual space: a technical overview». En: *Journal of the American Society for Information Science and Technology*, 1990, v. 41, n. 6, pp. 433–443.

20. **Moya Anegón, F. de**; **Vargas–Quesada, B.**; **Chinchilla–Rodríguez, Z.**; **Corera–Álvarez, E.**; **Herrero–Solana, V.**; **Muñoz–Fernández, F. J.** "Domain analysis and information retrieval through the construction of heliocentric maps based on ISI–JCR category co–citation". En: *Information processing and management: an international journal*, 2005, v. 41, n. 6, pp. 1.520–1.533.

21. **Ortega, J. L.**; **Aguillo, I. F.**; **Prieto, J. A.** «Longitudinal study of contents and elements in the scientific Web environment». En: *Journal of Information Science*, 2006.
http://internetlab.cindoc.csic.es/cv/11/Ortega2006.pdf

22. **Ortega Priego, J. L.** "Mapping web relationships in US National Laboratories: the vector space model vs. co–link analysis". En: *Journal of the American Society for Information Science and Technology*, 2006, (en prensa).

23. **Ortega Priego, J. L.**; **Aguillo, I. F.** «A web map of the CSIC research centres: a comparative study of the cosine and Pearson's r». En: *Proceedings of the 10th international conference of the International Society for Scientometrics and Informetrics*, 2005.
http://internetlab.cindoc.csic.es/cv/11/Ortega_Aguillo.pdf

24. **Prime, C.**; **Bassecoulard, E.**; **Zitt, M.** «Co–citations and co–sitations: a cautionary view on an analogy». En: *Scientometrics*, 2002, v. 54, n. 2, pp. 291–308.

25. **Rodríguez Gairín, J. M.** «Valoración del impacto de la información en internet: Altavista, el 'Citation Index' de la Red». En: *Revista española de documentación científica*, 1997, v. 20, n. 2, pp. 175–181.

26. **Rorvig, M.** «Images of similarity: a visual exploration of optimal similarity metrics and scaling properties of TREC topic–document sets». En: *Journal of the American Society for Information Science*, 1999, v. 50, n. 8, pp. 639–651.

27. **Rousseau, R.** «Sitations: an exploratory study». En: *Cybermetrics*, 1997, v. 1, p. 1.
http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

28. **Rousseau, R.** «Daily time series of common single word searches in AltaVista and NorthernLight». En: *Cybermetrics*, 1998, n. 2, p. 2.
http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

29. **Salton, G.** «The Smart retrieval system–experiments». En: **Salton, G.** (ed.). *Automatic document processing*. New York: Prentice–Hall, Englewood Cliffs, 1971.

30. **Salton, G.**; **Wong, A.**; **Yang, C. S.** «A vector space model for automatic indexing». En: *Communications of the Association for Computing Machinery*, 1975, v. 18, n. 11, pp. 613–620.

31. **Small, H.** «Cited document as concept symbols». En: *Social studies of science*, 1978, v. 8, pp. 327–340.

32. **Small, H.** «Paradigms, citations, and maps of science: a personal history». En: *Journal of the American Society for Information Science and Technology*, 2003, v. 54, n. 5, pp. 394–399.

33. **Small, H.**; **Griffith, B. C.** «The structure of the scientific literatures: identifying and graphing specialities». En: *Science studies*, 1974, v. 4, n. 1, pp. 17–40.

34. **Sneath, P.**; **Sokal, R.** *Numerical taxonomy*. San Francisco: Freeman, 1973.

35. **Spark–Jones, K.** "A statistical interpretation of term specificity and its application in retrieval". En: *Journal of documentation*, 1972, v. 28, n. 5, pp. 111–121.

36. **Thelwall, M.** «The responsiveness of search engine indexes». En: *Cybermetrics*, 2001, n. 5, p. 1.
http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html

José Luis Ortega, Isidro Aguillo, *Grupo de Investigación en Cibermetría., Cindoc-CSIC, Joaquín Costa, 22, 28002 Madrid.*
jortega@cindoc.csic.es
isidro@cindoc.csic.es