Revista Española de Documentación Científica,
34, 2, abril-junio, 232-252, 2011
ISSN: 0210-0614. doi: 10.3989/redc.2011.2.779

Aplicación de algoritmos genéticos a la identificación de la estructura de enlaces en portales web

María del Rocío Martínez-Torres*, Beatriz Palacios-Florencio*,
Sergio L. Toral-Marín**, Federico José Barrero-García**.

Resumen: Este trabajo explora la estructura de enlaces de los portales web considerándolos como grafos interconectados y analizando sus características como una red social. A partir de cada dominio raíz se extraerán dos redes: la primera, una red de dominios y la segunda, una red de páginas accesibles desde el dominio raíz. Sobre ambas redes se evaluarán una serie de parámetros desde la perspectiva del análisis de redes sociales para caracterizar la estructura del portal. El análisis factorial proporciona la metodología estadística adecuada para extraer los principales perfiles de portales web a partir de sus características como grafo. No obstante, y debido al gran número de indicadores que se pueden obtener, la búsqueda exploratoria de los factores latentes implicaría contemplar un número de posibilidades extremadamente elevado que imposibilitaría la obtención de una solución óptima. Por ello, en este trabajo se propone la utilización de una búsqueda genética sobre el conjunto de indicadores de partida. Los algoritmos genéticos son capaces de proporcionar un subconjunto de indicadores que optimizan una función objetivo. Los resultados obtenidos categorizan los portales webs corporativos en cuanto a su estructura de enlaces y destacan las posibilidades de los algoritmos genéticos como herramienta para descubrir nuevo conocimiento.

Palabras claves: Análisis de enlaces, estructura de portales web, análisis factorial, algoritmos genéticos.

Applying genetic algorithms for the identification of Websites' structure

***Abstract**: This paper explores website link structure, whereby websites are considered as interconnected graphs and their features are analyzed as a social network. For each root domain, two different networks are extracted: the first being the domain network and the second, the page network. In each case, a series of indicators taken from social network analysis is evaluated in order to characterize the website structure. Factor analysis may provide an appropriate statistical methodology for extracting in graphic form the principal profile of the website in terms of its internal structure. However, the large number of indicators generated by such an exploratory search would lead to a prohibitive number of possibilities. Therefore, this work proposes the use of genetic*

* Escuela Universitaria de Estudios Empresariales, Universidad de Sevilla, Avda. San Francisco Javier s/n, 41018 Sevilla, España. rmtorres@us.es y beatriz@us.es.

** E. S. Ingenieros, Universidad de Sevilla, C/ Enriquez de Ribera, 1, 41092, Sevilla, España. toral@esi.us.es y fbarrero@us.es.

Recibido: 10-04-2010; 2.ª versión: 07-06-2010; aceptado: 15-02-2011.

algorithms. By using this guided search over a given space of possible solutions, genetic algorithms can provide a subset of indicators able to optimize a fitness function. The results categorize corporate websites in terms of their link structure and highlight the possibilities for using genetic algorithms as a tool for knowledge discovery.

***Keywords**: Link analysis, Website structure, factor analysis, genetic algorithms.*

1. Introducción

El análisis de los enlaces web es el estudio cuantitativo de hipervínculos entre páginas web. Por lo general, el análisis de los enlaces forma parte del denominado «*Webometrics*», que es el análisis cuantitativo (Almind e Ingwersen, 1997) y cualitativo (Pinto-Molina y otros, 2004) de los fenómenos web, ocupándose también del análisis de citas web, la evaluación de motores de búsqueda y los estudios puramente descriptivos de la web (Björneborn y Ingwersen, 2004; Thelwall, 2008). Los enlaces web han sido muy estudiados durante los últimos años con el fin de comprender la estructura y los patrones de crecimiento de la web (Thelwall, 2004), aplicándose en particular al desarrollo de los algoritmos de clasificación de páginas. Este rápido desarrollo experimentado por el análisis de enlaces web en cuanto a teorías, tecnologías y metodologías podría explicarse por el hecho de ser una disciplina estudiada desde distintos puntos de vista, como la informática, las ciencias de la información, los estudios de comunicación o la sociología (Thelwall, 2004). El análisis de las redes sociales (SNA, Social Network Analysis) ha sido frecuentemente utilizado para el estudio del análisis de enlaces (Park y Thelwall, 2004; Toral y otros, 2010). SNA es un conjunto de procedimientos de investigación para la identificación de las estructuras de los sistemas sociales basados en las relaciones entre los componentes del sistema, también conocidos como nodos. En la aplicación de métodos SNA para el análisis de enlaces, los dominios web y las páginas web dentro de cada portal se consideran los actores, representados por los nodos en el grafo de la red social, mientras los enlaces son modelados como la relación entre actores, representados por las líneas que unen esos nodos (Iacobucci, 1994; Martínez-Torres y otros, 2010). El grafo resultante será un grafo dirigido (con un sentido asociado a cada línea que une dos nodos), porque los vínculos están definidos por una etiqueta HTML que apunta a una nueva página, definiendo de este modo el sentido de cada línea (en los grafos dirigidos, las líneas suelen denominarse arcos). La mayoría de los estudios relacionados con enlaces web se centran en la estructura de la web considerada a gran escala. Así por ejemplo, en estudios previos se han analizado las relaciones entre los dominios web de instituciones académicas nórdicas (Ortega y Aguillo, 2008), o incluso de Universidades a nivel mundial (Ortega y Aguillo, 2009) desde la perspectiva del SNA. En Baeza-Yates y Castillo (2007), los dominios webs de países se analizan atendiendo a varios criterios, en particular, grados de los nodos y rankings. La reputación de la página es otro tema relacionado con el análisis

de enlaces frecuentemente recogido en la literatura. En este caso, el SNA ha sido aplicado también considerando el método de grados de entrada (Indegree), que considera el número de enlaces sobre una página como medida de su popularidad. Se trata de una alternativa a los métodos de Pagerank (Berlt y otros, 2010), introducido por Google para caracterizar numéricamente la popularidad de páginas web. Finalmente, el análisis de enlaces a través del SNA ha sido combinado con el análisis semántico del texto para mejorar los algoritmos de recuperación de información web (Almpanidis y otros, 2007). Aunque existen bastantes estudios acerca de la estructura de la web o entre dominios web, comparativamente poco se sabe a nivel de la estructura interna de los portales web como organización de información y como mecanismos de acceso a esa información.

En este trabajo se propone un estudio exploratorio para la identificar la estructura de enlaces web dentro de un portal usando el análisis factorial. Para este propósito, las estructuras de hipertexto de 80 portales web institucionales de Universidades españolas se han extraído tanto a nivel de subdominios y dominios externos como a nivel de páginas web. Frente a otros portales institucionales, los portales universitarios garantizan una amplia variedad en la muestra, gracias a que la autonomía universitaria permite que el diseño y evolución de su portal web sea decisión autónoma de los órganos de gobierno de cada Universidad. Asimismo, históricamente las Universidades han sido organizaciones con una presencia activa en la web desde prácticamente sus inicios (Goldfarb, 2006). Los portales web se modelarán como dos redes sociales. En la primera red, los nodos representan subdominios o dominios externos y los arcos los enlaces entre ellos. La segunda red es parecida pero considera páginas web en lugar de dominios o subdominios. A partir de estas dos redes se puede derivar un elevado número de indicadores de su estructura atendiendo a parámetros típicamente medibles desde la perspectiva del SNA. No obstante, y debido a la naturaleza exploratoria de este estudio, es difícil seleccionar un subconjunto de indicadores que proporcione una solución satisfactoria mediante la aplicación del análisis factorial. No se garantiza si un subconjunto diferente podría proporcionar una solución más coherente y la alternativa de considerar todos los posibles subconjuntos de soluciones resultaría computacionalmente prohibitiva. Como solución se propone el uso de una técnica de búsqueda guiada como los algoritmos genéticos, capaz de proporcionar un subconjunto de indicadores que optimice una función de coste multi-objetivo. El resultado obtenido proporciona nuevos conocimientos sobre los patrones estructurales de portales web y pone en relieve la utilidad de los algoritmos genéticos como herramienta para el descubrimiento de nuevos conocimientos. Así pues, el objetivo del trabajo es doble. En primer lugar, se pretende definir un sistema experto basado en algoritmos genéticos capaz de determinar un conjunto de indicadores óptimo en la aplicación del análisis factorial a un conjunto de indicadores que caracterizan redes sociales. Esta solución óptima se refiere a explicar un valor elevado de la varianza de los datos con un conjunto de factores que sean interpretables. El segundo objetivo consiste en aplicar este sistema experto a la identificación de patrones estructurales en los portales web corporativos de la Universidades españolas.

El resto del documento está estructurado de la siguiente manera: el apartado 2 proporciona una breve descripción de la metodología propuesta. En concreto, se describe cómo modelar un portal web como un grafo, las características medibles desde la perspectiva del SNA y la metodología del análisis factorial. El apartado 3 está dedicado a la aplicación de algoritmos genéticos al problema de extraer un subconjunto óptimo de variables capaz de explicar las dimensiones latentes de la estructura web. El caso de estudio y los resultados se discuten en el apartado 4. Finalmente, las conclusiones se detallan en el apartado 5.

2. Metodología de análisis de la estructura de portales web usando SNA

Las redes que representan portales web se extraen comenzando a partir de un dominio raíz (dominio propio de un portal web institucional) y continuando luego con los enlaces de salida a otras páginas. Para cada portal web se consideran dos tipos de redes diferentes. La primera es la llamada red de dominios, en la cual los nodos representan subdominios o dominios externos diferentes al dominio raíz de partida. Los arcos representan el enlace entre ellos. La segunda red es la llamada red de páginas, que contiene todas las páginas web del portal web institucional y los enlaces entre ellas. Obviamente, ambas redes son grafos dirigidos y pueden ser extraídos hasta un nivel profundidad deseada (entendiendo por profundidad el número de enlaces necesarios para alcanzar una página desde el domino raíz). En ambos casos, la construcción de la red está limitada al dominio raíz. Esto significa que aunque se tienen en cuenta enlaces a otros dominios o páginas fuera del dominio raíz (y por tanto se incluyen en las redes consideradas), los enlaces de salida desde estos últimos no serán seguidos y no formarán parte de la red a analizar.

2.1. SNA

Una red social se puede representar como un grafo $G = (V, E)$ donde V denota un conjunto finito de vértices y E denota un conjunto de líneas de modo que $E \subseteq V \times V$. Matemáticamente, los grafos se suelen conceptualizar como matrices (Nooy y otros, 2005), como se muestra en la Ecuación (1).

$$M = (m_{i,j})_{n \times n} \qquad \text{donde } n = |V|, \qquad m_{i,j} = \begin{cases} 1 & si\ (v_1, v_j) \in E \\ 0 & en\ otro\ caso \end{cases} \tag{1}$$

En caso de un grafo valuado, la función de peso $w(e)$ está definida en el conjunto de líneas entre nodos, i.e. $w(e) = Ex\Re$, y la matriz anterior queda por tanto definida como se muestra en la Ecuación (2).

$$m_{i,j} = \begin{cases} w(e) & if\ (v_1, v_j) \in E \\ 0 & en\ otro\ caso \end{cases} \tag{2}$$

En el contexto de análisis de enlaces web, la red de dominios es una red en forma de estrella con el dominio raíz en el centro de la estrella y el resto de los dominios enlazados a ella. Varios indicadores relativos al tamaño de la red de dominios se han extraído en términos de número de nodos y arcos. Normalmente, los portales web institucionales suelen incluir subdominios que deberían tratarse aparte de los dominios externos. Esta distinción se ha tenido en cuenta en el tamaño medido en número de nodos. Finalmente, la densidad y el grado medio de los nodos se han considerado también como posibles indicadores. La densidad hace referencia al número de líneas y el grado, al número de arcos en los que cada vértice está involucrado. Por lo que respecta a la red de páginas, se trata de una red mucho más compleja, con mucho mayor tamaño y número de enlaces que la red de dominios. Esto también permite obtener mayor riqueza de información relativa a sus características como red social:

- Tamaño: el número de nodos representa el número de páginas web incluidas en el portal (o referenciadas si se trata de otros dominios) y los arcos representan las interrelaciones entre esas páginas. Un parámetro importante que determina el tamaño de la red es el nivel de profundidad para el que se extraen las páginas pertenecientes a cada portal web. En este estudio hemos usado una profundidad de siete. Este valor es considerado suficiente para captar la información esencial de la estructura del sitio web y es mayor que la profundidad de cinco usada en algunos estudios previos (Yang y Qin, 2008).
- Densidad: es una media del número de líneas en una red simple, expresada como una proporción del número máximo posible de líneas. Las figuras 1.*a*) y 1.*b*) detallan una red de baja y alta densidad, respectivamente. El principal problema de esta definición es que no tiene en cuenta las líneas valuadas con valor superior a 1 y que depende del tamaño de la red. Una medida diferente de densidad se basa en la idea del grado de un nodo, que es el número de líneas que inciden (grado de entrada) o salen (grado de salida) de él (Toral y otros, 2009a). Mayores grados de nodos producen redes más densas, porque los nodos involucran más arcos, y el valor medio del grado de los nodos de una red no es una medida dependiente del tamaño de la red**.** Como la red de páginas es un grafo dirigido, varias medidas estadísticas sobre la distribución del grado de salida de los nodos serán consideradas. Finalmente, la densidad puede ser también medida desde la perspectiva del SNA usando un punto de vista egocéntrico. La densidad egocéntrica de un nodo es la densidad de sus conexiones entre sus vecinos (Nooy y otros, 2005).
- Componentes: Un componente fuerte es una subred fuertemente conectada de tamaño máximo. Se dice que una red está fuertemente conectada si cada par de vértices está conectado por un camino teniendo en cuenta el sentido de los arcos (Nooy y otros, 2005). En el contexto de este estudio, el análisis de los componentes de la red permite la identificación de subes-

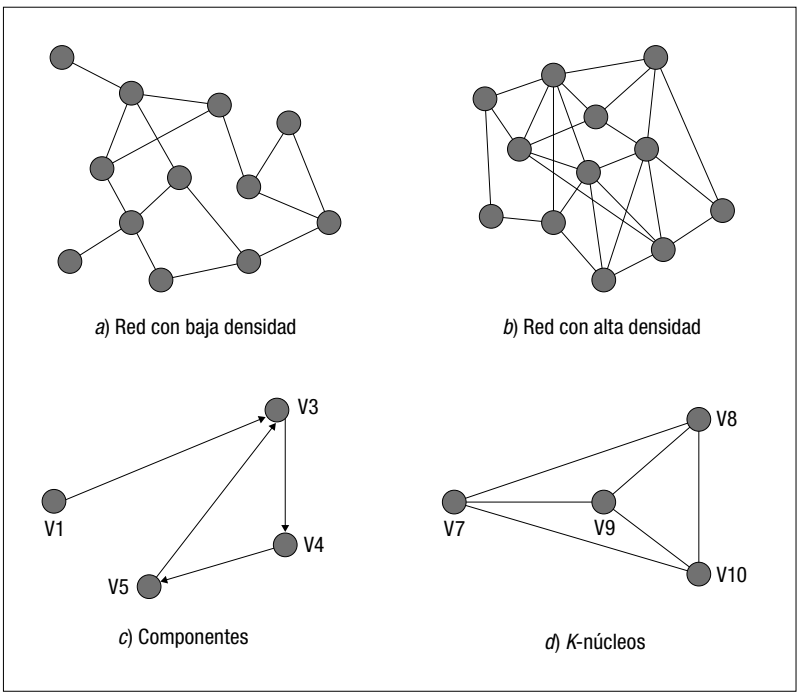
tructuras conectadas dentro del portal web. La figura 1.*c*) detalla el componente formado por los vértices v3, v4 y v5.

- *K*-Núcleos (*k-cores*): un *k*-núcleo es una subred en la que cada nodo tiene *k* grados dentro de esa subred. El núcleo con mayor grado representa el núcleo central de la red. Los *k*-núcleos han sido utilizados en trabajos previos para detectar subredes entre portales web académicos de países nórdicos (Ortega y Aguillo, 2008). La figura 1.*d*) detalla un k-núcleo, con $k = 3$.
- Distancia: se define como el número de pasos en el camino más corto entre dos nodos de la red. En el caso de los portales web, existe un claro nodo principal definido por el dominio raíz. Consecuentemente tiene sentido medir la distancia del resto de páginas respecto a ese nodo.
- Centralidad cercana *(Closeness centralization):* es un índice de centralidad basado en el concepto de distancia. La centralidad cercana de un nodo se calcula considerando el total de distancias entre un nodo y todos los demás nodos, donde la distancia más larga ofrece una menor puntuación de centralidad cercana. La centralidad cercana es un índice definido para toda la red y se calcula como la variación en la centralidad cercana de los vértices dividida por la variación máxima posible en la puntuación de centralidad cercana en una red del mismo tamaño (Toral y otros, 2009b).
- Grado de Intermediación *(Betweenness):* es una medida de la centralidad que reside en la idea de que un nodo es más central en la medida en que actúe como intermediario en una red de comunicación (Nooy y otros, 2005). Es decir, la centralidad de un nodo depende de la medida en la que es necesario como enlace para facilitar la conexión de otros nodos dentro de la red. Si se define una geodésica como el camino más corto entre dos nodos, la centralidad de intermediación de un vértice es la proporción de todas las geodésicas entre pares de nodos que incluyen este nodo, y la centralidad en la intermediación de una red es la variación en la centralidad de intermediación de los nodos dividida por la máxima variación posible en la centralidad de intermediación en una red del mismo tamaño. Desde la perspectiva del análisis de enlaces esta medida permite detectar pasarelas que conectan a redes separadas (Faba-Pérez y otros, 2005).
- Correlación entre particiones: una partición de una red es una clasificación o *clustering* de los nodos en una red, de modo que cada nodo se asigna únicamente a una clase o *cluster* (Toral y otros, 2010). Existen dos particiones significativas que pueden extraerse a partir de la red de páginas. La primera es la partición de los k-vecinos a partir del dominio raíz, en la cual los nodos son clasificados usando la distancia al nodo raíz. La segunda es la partición del grado de salida en la cual los nodos son clasificados atendiendo a su valor de grado de salida. La correlación entre ambas particiones es definida como la medida en la cual el sitio web sigue una estructura en forma de árbol desde el dominio raíz. Se evaluarán dos tipos de índices de asociación referenciados en la literatura: la V de Cramer y el índice de información de Rajski (Nooy y otros, 2005). La V de Cramer mide la depen-

dencia estadística entre dos clasificaciones. El índice de Rajski mide el grado por el cual la información de una clasificación se preserva en la otra clasificación. Sólo se considerará la versión simétrica del índice de Rajski.

FIGURA 1

Representación gráfica de algunas características medibles en redes sociales



2.2. Análisis factorial

El análisis factorial es una manera de ajustarse a un modelo de datos multivariados, estimando su interdependencia. Esto aborda el problema de analizar la estructura de interrelaciones entre un número de variables usando un conjunto de dimensiones subyacentes comunes, los factores, los cuales no son directamente observables, segmentando una muestra en segmentos relativamente homogéneos (Rencher, 2002). Ya que cada factor puede afectar a varias variables en común, estos son conocidos como «factores comunes». Se asume que cada variable es dependiente en una combinación lineal de factores comunes y los coeficientes son conocidos como «*loadings*» o cargas factoriales (Martínez-Torres y Toral, 2010a).

El análisis del factor puede ser usado tanto para exploración como para propósitos confirmatorios: A diferencia de los análisis confirmatorios, los análisis exploratorios no establecen ninguna constante a priori en la estimación de fac-

tores o del número de factores a ser extraídos (Toral y otros, 2009c). La naturaleza exploratoria de este estudio tiene varias implicaciones:

- El elevado número de indicadores relacionados con SNA que pueden extraerse de las dos redes consideradas. Los antecedentes teóricos existentes no permiten descartar previamente indicadores antes de comenzar el análisis factorial.
- El número de factores latentes es desconocido. De nuevo, la falta de antecedentes teóricos suficientes significa que los factores deberían ser seleccionados atendiendo a la homogeneidad de sus indicadores.

En la siguiente sección se propone el uso de algoritmos genéticos para buscar una solución óptima y resolver esos problemas.

3. Metodología de búsqueda genética de las dimensiones latentes en portales web

Un Algoritmo Genético (AG) es una abstracción computacional de una evolución biológica que puede utilizarse para resolver algunos problemas de optimización. La técnica fue primeramente introducida por Holland (1975) para su uso en sistemas adaptativos, y se basan en los principios de la evolución natural y la supervivencia de los más fuertes. Por imitación de este proceso, los Algoritmos Genéticos son capaces de ir creando soluciones para problemas del mundo real. Trabajan con una población de individuos, cada uno de los cuales representa una solución factible a un problema dado. A cada individuo se le asigna un valor o puntuación, relacionado con la bondad de dicha solución (es el valor de *fitness,* que cuantifica su valor como solución al problema). En la naturaleza esto equivaldría al grado de efectividad de un organismo para competir por unos determinados recursos. El algoritmo comienza con una población inicial que se selecciona al azar desde el espacio de posibles soluciones. A partir de ella, las siguientes operaciones combinan la información genética de los elementos que la componen para formar nuevas generaciones.

- En la operación de reproducción, los individuos compiten por reproducirse basándose en sus valores de fitness, de modo que aquellos individuos que representen mejores soluciones tienen mayores probabilidades de supervivencia.
- La operación de recombinación implica a dos individuos que intercambian parte de su información genética. La selección de los individuos padre también se realiza acorde a sus valores de fitness y, como resultado, proporcionan dos individuos hijos con parte de su información genética intercambiada. La operación de recombinación permite que trozos de la información genética que contribuyan a buenas soluciones pervivan a lo largo de la evolución.

En general, cuanto mayor sea la adaptación de un individuo al problema, mayor será la probabilidad de que el mismo sea seleccionado para reproducirse y recombinarse, cruzando su material genético con otro individuo seleccionado de igual forma. Este cruce producirá nuevos individuos descendientes de los anteriores, los cuales comparten algunas de las características de sus padres. Cuanto menor sea la adaptación de un individuo, menor será la probabilidad de que dicho individuo sea seleccionado para la reproducción y, por tanto, de que su material genético se propague en sucesivas generaciones. De esta manera se produce una nueva población de posibles soluciones, la cual reemplaza a la anterior y verifica la interesante propiedad de que contiene una mayor proporción de buenas características en comparación con la población anterior. Así a lo largo de las generaciones las buenas características se propagan a través de la población. Favoreciendo el cruce de los individuos mejor adaptados, van siendo exploradas las áreas más prometedoras del espacio de búsqueda. Si el Algoritmo Genético ha sido bien diseñado, la población convergerá hacia una solución óptima del problema.

El AG usa una estrategia elitista que significa que el mejor individuo es siempre reproducido a la generación siguiente, de modo que siempre se conserva la mejor solución obtenida a lo largo de la evolución. El algoritmo se detiene cuando se satisface algún criterio de parada de su ejecución (Martínez-Torres y Toral, 2010b).

Para una correcta aplicación de los algoritmos genéticos, es preciso tener en cuenta varias cuestiones:

- Codificación de los individuos, es decir, cómo se van a codificar los individuos de una población de modo que esta codificación permita recoger el conjunto de soluciones posibles al problema.
- Selección de la función de *fitness,* de modo que represente la bondad de las soluciones según el problema planteado.
- Selección de los valores de los parámetros (tamaño de la población, número de iteraciones, probabilidades, etc.).

En este estudio, el uso del AG está justificado debido a su naturaleza exploratoria. De acuerdo a las características medibles detalladas en el apartado 2.1 se han obtenido un total de 64 indicadores. La elección de un subconjunto de indicadores para la realización del estudio exploratorio resultaría prohibitiva si se tratase de explorar la totalidad de soluciones posibles. El espacio de posibles soluciones está formado por 2^{64} = 1,8447*e* + 019 posibilidades, que significan que deberíamos ejecutar 2^{64} análisis factoriales diferentes para explorar completamente el espacio de posibles soluciones. A diferencia de esta alternativa, AG permite llevar a cabo una búsqueda guiada de la solución óptima con un menor coste computacional.

La primera condición para aplicar AG adecuadamente es una buena selección de la codificación de individuos, la cual debería ser válida y completa. Nuestra codificación de los individuos está constituida por una secuencia binaria de 64 valores, en las que los «unos» representan las variables que van a ser usadas en el análisis factorial y los «ceros» representan variables que van a ser excluidas de

este análisis. La figura 2 muestra un ejemplo de codificación para el caso concreto de considerar las variables impares. Claramente, la representación de codificación es completa, pues las 2^{64} posibilidades pueden ser representadas, y válidas, ya que todas ellas pueden ser evaluadas. La figura 3 detalla la operación genética de recombinación para la codificación utilizada. El punto de cruce se elige aleatoriamente y, mediante un cruce, se generan los individuos hijos a partir de la codificación de los individuos padres.

FIGURA 2

Codificación binaria de las posibles soluciones

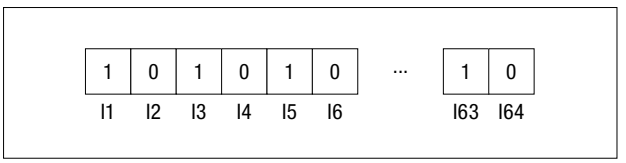
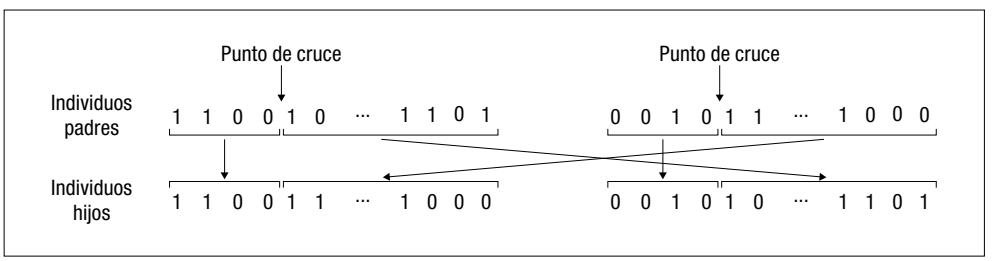


FIGURA 3

Operación genética de recombinación



El siguiente paso es la selección de la función de *fitness,* que cuantifica la idoneidad de cada individuo como solución al problema. Los individuos con un alto valor de *fitness* tienen más posibilidad de ser seleccionados, pasando su material genético (vía reproducción o recombinación) a la siguiente generación. La función de *fitness* es quien impulsa la evolución de la población hacia una nueva generación de individuos con una mayor idoneidad que los de la generación anterior. El individuo que represente la solución óptima debería tener el máximo valor de *fitness* dentro del espacio de las soluciones y las soluciones óptimas cercanas deberían tener valores de *fitness* cercanos. En el contexto del análisis factorial no es posible construir una función de *fitness* simple. Por el contrario, es necesario construirla como una función multi-objetivo considerando varios parámetros, como la varianza explicada, correlaciones e interpretación de los factores latentes.

$$F = c_1\ Var + c_2 \frac{1}{n} \sum_{i=1}^{k} r_i^2 + c_3 Interp \qquad (7)$$

- Varianza explicada (*Var*). Los resultados del análisis factorial muestran la varianza explicada por los factores considerados (normalmente, el número de factores viene dado por el número de autovalores de la matriz de correlación mayores que 1). Aunque la varianza explicada por el número seleccionado de indicadores debería ser maximizado, no es sin embargo el único parámetro a tener en cuenta. Una función de *fitness* que únicamente considere la varianza explicada tenderá a la solución trivial de considerar sólo un indicador. Esto se debe al hecho de que es más fácil explicar la varianza de un conjunto de datos cuando está integrado por un número pequeños de ellos.
- Correlaciones entre variables $\left(\frac{1}{n}\sum_{i=1}^{k} r_i^2\right)$. La media de la suma de los coeficientes de correlación al cuadrado de los indicadores se usará como la segunda parte de la función de *fitness*. Este término considerado por sí sólo también tendería a la solución trivial de considerar el conjunto completo de los datos de partida. Es la fuerza inversa a la parte previa de la función de *fitness*.
- Interpretación de factores. La tercera parte de la función de *fitness* se utiliza para penalizar a los factores con menos de tres indicadores. La razón de elegir el valor de tres es porque los factores explicados con menos de tres indicadores no se consideran bien definidos en la literatura (Rencher, 2002). Esta parte de la función de *fitness* es la más importante ya que promueve un número reducido de factores con más indicadores, mejorando la interpretación final de los factores latentes.

Los coeficientes C1, C2, y C3 se usan para ajustar la importancia relativa de las tres partes de la función de *fitness*. Obviamente, su rango es [0,1] con la restricción de C1 + C2 + C3 = 1. La decisión final para la aplicación del AG se refiere a determinados parámetros previos a la ejecución del algoritmo. La representación del AG puede ser sensible a ciertos valores de estos parámetros, particularmente al tamaño de la población, la frecuencia de selección de operador y el criterio de finalización. Por lo general, un alto valor para el tamaño de población ayuda a reducir esta sensibilidad a los parámetros del AG. En este trabajo, el tamaño de la población es igual a 10.000, y durante la ejecución de mantiene un 20% de tasa de reproducción y un 80% de tasa de recombinación. El valor de 10.000 se considera un valor adecuado para obtener suficiente riqueza de información. Estos valores son típicos en la literatura sobre AG (Goldberg, 1989: Martínez-Torres y Toral, 2010b).

4. Resultados

La búsqueda genética de las dimensiones latentes de portales web se ha aplicado a 80 portales web corporativos pertenecientes a Universidades españolas. Todos ellos están incluidos en el Ranking mundial de Universidades en la Web (www.webometrics.info), donde más de 6.000 Universidades de todo el mundo están ordenadas según el tamaño y la visibilidad. En la lista de la tabla I se enu-

meran los dominios raíz de los portales web considerados. Prácticamente cubren casi todo el rango del ranking mundial de Universidades en la web y muestran diversidad de tamaños en términos de dominios y páginas web. La tabla II resume algunas estadísticas descriptivas. La primera columna muestra que en la extracción de datos han sido considerados más de 718.000 páginas web y más de cuatro millones de enlaces de salida. A título ilustrativo, la figura 4 y la figura 5 muestran, respectivamente, la red de dominios y la red de páginas correspondientes al caso particular de la Universidad de Sevilla. La red de dominios es una red en forma de estrella, en cuyo centro se sitúa el dominio raíz (www.us.es), y el resto de nodos está formado por todos los subdominios y dominios externos referenciados desde cualquier página que cuelgue del dominio raíz. La red de páginas está formada por todas las páginas accesibles desde el nodo raíz de la Universidad de Sevilla, que suman un total de 11.455 páginas. La figura 5 ilustra la dificultad de representar de una manera legible una red con un número de nodos tan elevado, por lo que es preciso caracterizarlas mediante las medidas definidas en la sección 2.1. Por cada portal web de la tabla I se han sido extraído estas dos mismas redes: la red de dominios y la red de páginas.

TABLA I

Lista de los portales web considerados

www.ucm.es	portal.uned.es	www.ual.es	www.cef.es
www.upc.edu	www.uva.es	www.udl.es	www.uch.ceu.es
www.upm.es	www.upf.edu	www.ujaen.es	www.nebrija.com
www.uab.es	www.unav.es	www.umh.es	www.uic.es
www.ehu.es	www.uc3m.es	www.deusto.es	www.url.es
www.ub.edu	www.uniovi.es	www.unavarra.es	www.esdi.es
www.us.es	www.uma.es	www.upct.es	www.uax.es
www.upv.es	www.uco.es	www.upo.es	www.vives.org
www.um.es	www.ull.es	www.ie.edu	www.uimp.es
www.ugr.es	www.udc.es	www.upcomillas.es	www.ucjc.edu
www.ua.es	www.unex.es	www.ceu.es	www.ucv.es
www.uvigo.es	www.uah.es	www.iese.edu	www.uspceu.com
www.uv.es	www.uoc.edu	www.ubu.es	www.cesdonbosco.com
www.uam.es	www.udg.edu	www.urv.net	www.ufv.es
www.usal.es	www.ulpgc.es	www.unirioja.es	www.esic.es
www.uji.es	www.unican.es	www.uem.es	www.cepade.es
www.unizar.es	www.unileon.es	www.esade.edu	www.eoi.es/portal
www.usc.es	www.urjc.es	www.ucam.edu	www.esmuc.net
www.uib.es/ca	www.uca.es	www.mondragon.edu	www.udima.es
www.uclm.es	www.uhu.es	www.uvic.es	www.eupmt.es

TABLA II

Estadísticas descriptivas de los 80 portales web considerados

	Total	Media	SD	Mínimo	Máximo
Subdominios	2.438	30,47	38,10	0	180
Dominios ext.	30.500	381,25	580,32	0	3.623
Páginas	718.272	8.978,40	15.334,01	110	122.930
Enlaces de salida	4.429.231	55.365,38	73.290,17	435	445.368

Las características de la red social de la sección 2.1 se han medido considerando en algunos casos la totalidad de la red y en otros casos las subredes excluyendo nodos con grado de salida *(Outdegree)* cero o subredes con $k > 1$ núcleos *(cores)*. Como resultado, un total de 64 indicadores han sido obtenidos.

FIGURA 4

Red dominante de la Universidad de Sevilla

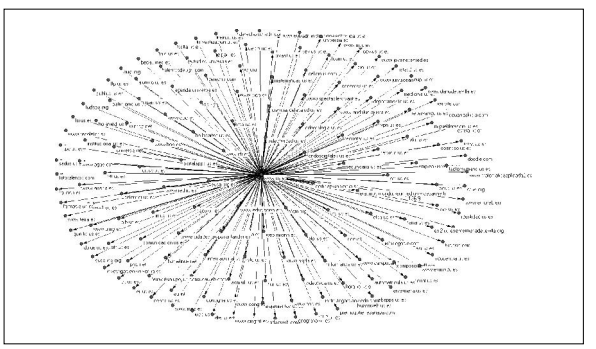
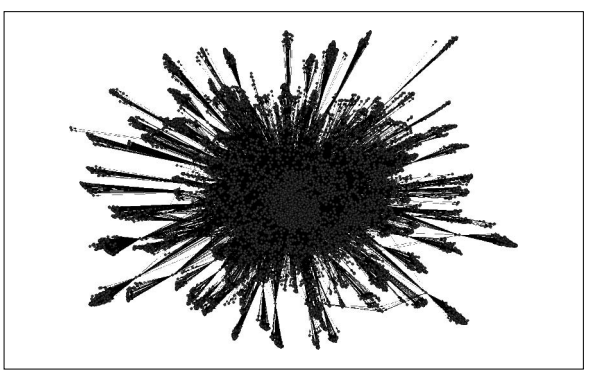


FIGURA 5

Red de páginas de la Universidad de Sevilla



4.1. Análisis de los datos

El AG se ha aplicado para obtener un subconjunto óptimo de indicadores capaces de identificar los perfiles de los portales web de acuerdo a su estructura de organización interna. La función de coste sigue la estructura general definida en el apartado 3 pero considerando los valores $c1 = 0{,}15$, $c2 = 0{,}1$ y $c3 = 0{,}75$. Obsérvese que la parte dedicada a la interpretación de los factores ha sido claramente sobreponderada. Esta estrategia parece razonable, ya que los factores con menos de tres indicadores no son admisibles en el análisis de los factores. Además, este término hace que el AG evolucione hacia un número reducido de factores, lo que también resulta razonable para que los factores obtenidos tengan un significado claro y separado.

Comenzando con una población inicial generada aleatoriamente, el AG converge después de 30 generaciones, con una varianza explicada de 77,10% y 25 indicadores agrupados en 6 factores. Todos ellos incluyen al menos tres indicadores y su significado, usando una rotación Varimax, son interpretables. El tiempo de ejecución del algoritmo genético es de 4.822,49 segundos (80,37 minutos). Este valor es mucho más pequeño que la opción alternativa de explorar al completo el espacio de soluciones. Teniendo en cuenta que cada análisis de factores requiere 12,9 ms (milisegundos) para ser procesado, las $2^{64} = 1{,}8447e + 0{,}19$ posibilidades del espacio de soluciones requeriría millones de años. El subconjunto de indicadores seleccionados se muestra en la tabla III. En particular, se detalla la descripción de los indicadores y de la red sobre la que se calculan.

TABLA III

Subconjunto de indicadores seleccionados

	Indicador	Red
I1	Dominios externos.	Red de dominios.
I2	Grado medio.	Red de dominios.
I3	Densidad.	Red de dominios.
I4	Número de páginas.	Red de páginas.
I5	Número de páginas en el ultimo nivel (profundidad 7).	Red de páginas.
I6	Número de páginas sin retorno (excluyendo el último nivel).	Red de páginas.
I7	Desviación típica del grado de salida..	Red de páginas.
I8	Número de componentes Fuertes (Strong Components).	Red de páginas.
I9	% de páginas incluidas en los componentes fuertes.	Red de páginas.
I10	*K*-cores que incluye el máximo número de páginas.	Red de páginas.
I11	Valor Medio de centralidad cercana.	Red de páginas.
I12	Desviación típica de centralidad cercana.	Red de páginas.
I13	Número de páginas.	Red de páginas excluyendo grado de salida = 0.

TABLA III *(continuación)*

	Indicador	Red
I14	Centralización de intermediación.	Red de páginas.
I15	Desviación típica de la densidad egocéntrica.	Red de páginas.
I16	Valor medio de la centralización de intermediación de los nodos.	Red de páginas, *k*-core, $k > 0$.
I17	Desviación típica de la centralización de intermediación de los nodos.	Red de páginas, *k*-core, $k > 0$.
I18	Valor medio de densidad egocéntrica.	Red de páginas, *k*-core, $k > 0$.
I19	Valor medio de la centralización de intermediación de los nodos.	Red de páginas excluyendo grado de salida = 0.
I20	Valor medio de densidad egocéntrica.	Red de páginas excluyendo grado de salida = 0.
I21	Numero de nodos que desarrollan un rol de intermediación.	Red de páginas excluyendo grado de salida = 0.
I22	Desviación típica de los roles de intermediación.	Red de páginas excluyendo grado de salida = 0.
I23	Índice V de Cramer de la correlación de particiones (grado de salida, *k*-vecinos).	Red de páginas.
I24	Índice de correlación de Rajski de la correlación de particiones (grado de salida, *k*-vecinos).	Red de páginas.
I25	Índice de correlación de Rajski de la correlación de particiones (grado de salida, *k*-vecinos).	Red de páginas excluyendo grado de salida = 0.

Los resultados del análisis factorial usando el conjunto de variables seleccionadas por el algoritmo genético se detallan en la tabla IV. Por lo general, se selecciona un número de factores igual al número de autovalores superiores a 1 (Rencher, 2002). En este caso, son 6 factores latentes obtenidos como resultado del análisis factorial.

TABLA IV

Varianza explicada como resultado de un análisis factorial

Factor	Valores propios		
	Valor	% varianza	% acumulativo
1	7,990	31,962	31,962
2	3,852	15,407	47,369
3	2,911	11,646	59,015
4	1,857	7,427	66,442
5	1,656	6,624	73,065
6	1,010	4,039	77,104
7	0,833	3,333	80,437
...	...	...	...
25	0,007	0,029	100,000

Los indicadores asociados a cada factor se obtienen a partir de las cargas factoriales usando una rotación Varimax. Todos los indicadores asociados de esta manera con el mismo factor están bajo la hipótesis de que comparten un sentido común que el analista debe descubrir.

Por otra parte, las puntuaciones de los factores se usan para categorizar la muestra original de Universidades, cada una de las cuales puede aproximarse a uno de los factores latentes identificados. Para comprobar la hipótesis nula de igualdad de medias entre los grupos de Universidades se ha llevado a cabo un análisis de la varianza (ANOVA). La hipótesis nula ha sido rechazada para todos los indicadores con un significativo valor por debajo de 0,05. Usando la información de las cargas factoriales, así como los valores medios de las categorizaciones de Universidades, se pueden destacar las siguientes pautas de estructura en portales web (tabla V):

El factor 1 representa una estructura distribuida del portal web, con una gran cantidad de nodos desarrollando un papel de intermediación. El alto valor de las correlaciones de las particiones (indicadores I23 e I25) significa que el grado de salida crece a medida que nos alejamos del nodo raíz, lo que sugiere que las páginas de nivel inferior o intermedio (cerca del dominio raíz) actúan como directorios de información mientras que las páginas de nivel superior (lejos del dominio raíz) proporcionan información más detallada. Por otro lado, los altos valores medios y de desviación típica de la centralidad de intermediación (indicadores I16, I17 e I19) indican que el portal sigue una estructura tipo árbol, con vértices cada vez más conectados a medida que descendemos en niveles de profundidad. La figura 6.*a*) detalla una representación simbólica del portal.

El factor 2 representa una estructura más centralizada en el sentido de la distancia al dominio raíz. Hay un núcleo de las páginas altamente interconectado, pero la información también se extiende a medida que avanzamos hacia niveles más profundos en la estructura. Los altos valores medios y de desviación típica de la centralidad cercana sugieren una estructura más plana, con caminos cortos para encontrar la información deseada. La representación simbólica de la figura 6.*b*) muestra este caso, con un camino reducido para alcanzar un nodo terminal B desde el nodo raíz A.

El factor 3 se refiere a una estructura egocéntrica, donde la red global podría ser considerada como la suma de subredes más o menos independientes. Es el caso de portales web con una clara división en áreas independientes, tal y como se muestra en la figura 6.*c*). En el contexto de los portales web Universitarios, se trataría de una división en unidades funcionales básicas (docencia, investigación, transferencia tecnológica, etc.).

El factor 4 considera los sitios web de gran tamaño. El número de páginas crece geométricamente con el nivel de profundidad, por lo que es necesario un proceso de larga navegación para lograr la información deseada. Al contrario que en el factor 1, los bajos valores de los índices de correlación de Rajski y Cramer indican un portal poco estructurado, donde las páginas poseen enlaces a otras muchas páginas de niveles diferentes, figura 6.*d*). Aunque esta estructuración del portal

permite que los visitantes puedan navegar de una forma mucho más libres, también es a costa de una mayor complejidad para encontrar la información requerida.

El factor 5 representa los sitios web más pequeños, donde una gran cantidad de información se proporciona usando referencias externas a otros sitios web o a subdominios. Esta idea se sustenta por el alto valor de páginas de no retorno, excluyendo las páginas localizadas en el último nivel, así como por el elevado valor de dominios externos y subdominios. Las páginas que cuelgan del dominio raíz se encuentran altamente interconectadas, figura 6.*e*).

Finalmente, el factor 6 representa portales web con una estructura dominada por una subred, que contiene la información más relevante. El alto valor del indicador I10, relacionado con los k-núcleos, sugiere una estructura como la representada en la figura 6.*f*). Un *k*-núcleo se caracteriza por identificar una subred en la que todo los nodos poseen al menos un grado *k*. Esta subred constituye el núcleo base del portal.

FIGURA 6

Representación simbólica de los portales web identificados

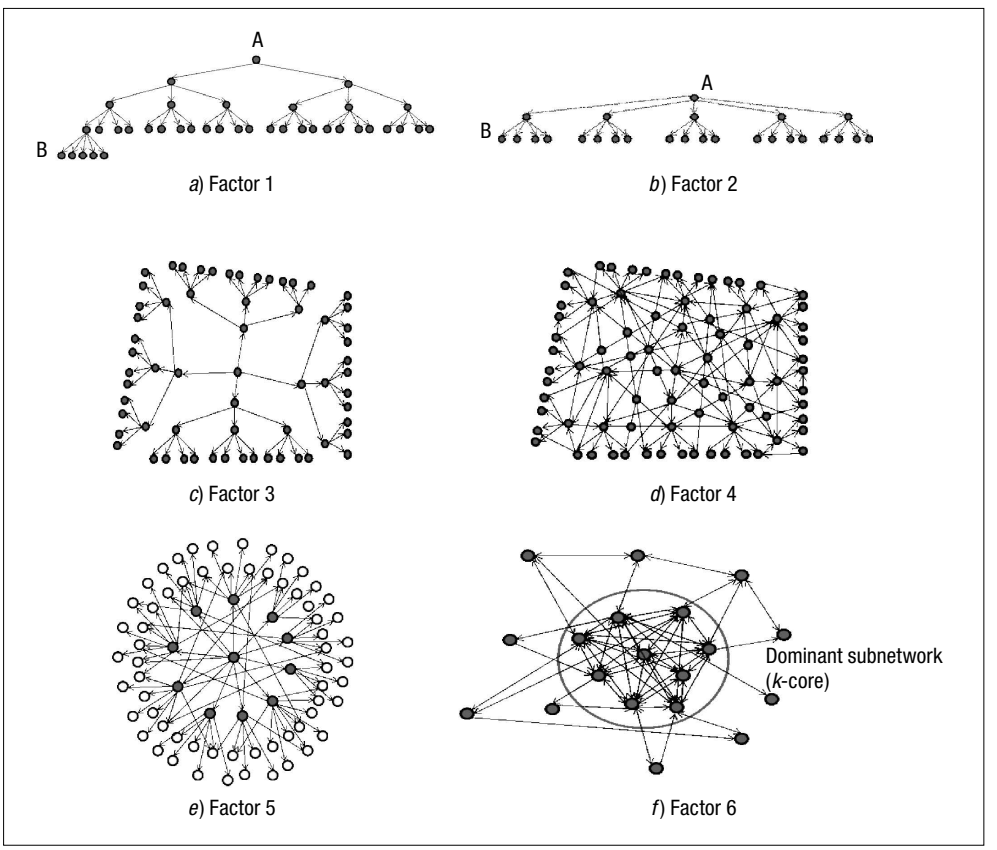


TABLA V

Factores identificados

		Descripción	Loading
F1	I2	Grado medio.	–0,724
	I16	Valor medio de la centralización de intermediación de los nodos.	0,903
	I17	Desviación típica de la centralización de intermediación de los nodos.	0,884
	I19	Valor medio de la centralización de intermediación de los nodos.	0,839
	I23	Índice V de Cramer de la correlación de particiones (grado de salida, *k*-vecinos).	0,703
	I25	Índice de correlación de Rajski de la correlación de particiones (grado de salida, *k*-vecinos).	0,746
F2	I9	% de páginas incluidas en los componentes fuertes.	0,722
	I11	Valor medio de centralidad cercana.	0,924
	I12	Desviación típica de centralidad cercana.	0,718
	I14	Centralización de intermediación.	0,826
	I24	Índice de correlación de Rajski de la correlación de particiones (grado de salida, *k*-vecinos).	0,578
F3	I15	Desviación típica de la densidad egocéntrica.	0,763
	I18	Valor medio de densidad egocéntrica (Red de páginas, *k*-core, $k > 0$).	0,895
	I20	Valor medio de densidad egocéntrica (Red de páginas excluyendo grado de salida = 0).	0,875
F4	I4	Número de páginas (Red de páginas).	0,900
	I5	Número de páginas en el ultimo nivel (profundidad de 7).	0,928
	I13	Número de páginas (Red de páginas excluyendo grado de salida = 0).	0,661
	I21	Numero de nodos que desarrollan un rol de intermediación.	0,510
F5	I1	Dominios externos.	0,852
	I6	Número de páginas sin retorno (excluyendo el último nivel).	0,647
	I8	Número de componentes fuertes.	0,831
F6	I7	Desviación típica del grado de salida.	0,786
	I10	*K*-cores que incluye el máximo número de página.	0,633
	I22	Desviación típica de los roles de intermediación.	0,635

Básicamente, los perfiles identificados en las estructuras de portales web responden a dos estrategias básicas a la hora de decidir su estructura final (Tan y Wei, 2006). La primera estrategia consiste en ofrecer una estructura que tenga sentido para el usuario final. En este sentido, los portales web sacrifican la accesibilidad de la información en busca de un esquema de navegación más es-

tructurado. La opción alternativa consiste en la reducción de grandes estructuras bajo el supuesto de que el desempeño del usuario es óptimo cuando la amplitud y profundidad de la página web se mantiene a un nivel moderado (Tan y Wei, 2006). En general, navegabilidad y accesibilidad son dos parámetros estrechamente relacionados con la estructura interna de los portales web. Existen estudios que los incluyen como características de diseño de portales web corporativos (Robbins y Stylianou, 2003), o dentro de los índices de evaluación de portales web (Miranda y Bañegil, 2004). En línea con estos trabajos, los factores identificados se pueden clasificar como portales fuertemente estructurados (factores 1 y 3), que sacrifican la accesibilidad por un esquema de navegación más comprensible, como portales estructurados que mejoran la accesibilidad mediante estructuras más planas (factor 2) y como portales poco estructurados que permiten una navegación más autónoma del usuario y mejoran la accesibilidad de la información a través de muchos caminos posibles (factores 4, 5 y 6).

Los perfiles identificados extienden, además, algunas estructuras previamente identificadas en la literatura. Por ejemplo, la estructura en árbol identificada por Huizingh (2000) se subdivide en una estructura en árbol profunda (factor 1), plana (factor 2) y estructurada en subredes (factor 3). Asimismo, las estructuras de portales web altamente conectados identificados en este mismo estudio se subdividen en portales web de gran tamaño (factor 4) y con subred dominante (factor 6).

5. Conclusión

Este trabajo ha desarrollado un sistema experto para la selección de indicadores en la realización de análisis factoriales exploratorios, que posteriormente se ha aplicado a la identificación de las estructuras de enlaces de portales web considerando dichos portales como redes sociales. El uso de técnicas de computación evolutiva como los algoritmos genéticos permite realizar una búsqueda guiada sobre el espacio total de soluciones, simplificando extraordinariamente el tiempo de computación respecto a la alternativa de evaluar el conjunto de todas las soluciones posibles (que en muchos casos, como en el descrito en el artículo, resultaría prohibitiva). El resultado de dicha búsqueda se caracteriza por proporcionar una solución interpretable y capaz de explicar un valor elevado de la varianza de los datos de partida. Su aplicación a la identificación de los patrones estructurales de portales web corporativos universitarios proporciona resultados no sólo acordes con lo descrito en la literatura sino que amplían y detallan patrones no considerados previamente. Asimismo, se relacionan con los conceptos de navegabilidad y accesibilidad, identificados como parámetros evaluables en portales web. Aunque el estudio se limita a los portales web de Universidades españolas, constituyen una muestra lo bastante rica dentro del ranking mundial de Universidades en la web. Este estudio podría extenderse a otros portales web institucionales para validar los resultados obtenido

6. Agradecimientos

Este trabajo ha sido apoyado por el Ministerio Español de Educación y Ciencia (Proyecto de investigación con referencia DPI2007-60128) y la Consejería de Innovación, Ciencia y Empresa (Proyecto de investigación con referencia P07-TIC-02621).

7. Bibliografía

Almind, T. C., y Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to Webometrics, *Journal of Documentation,* vol. 53 (4), pp. 404-426.

Almpanidis, G.; Kotropoulo, C., y Pitas, I. (2007). Combining text and link analysis for focused crawling. An application for vertical search engines, *Information Systems,* vol. 32, pp. 886-908.

Baeza-Yates, R., y Castillo, C. (2007). Characterization of national web domains, *ACM Transactions on Internet Technology*, vol. 7 (2), pp. 1-32.

Berlt, K.; Silva de Moura, E.; Carvalho, A.; Cristo, M.; Ziviani, N., y Couto, T. (2010). Modeling the web as a hypergraph to compute page reputation, *Information Systems,* vol. 35 (5), pp. 530-543.

Björneborn, L., y Ingwersen, P. (2004). Toward a basic framework for webometrics, *Journal of the American Society for Information Science and Technology,* vol. 55 (14), pp. 1216-27.

Faba-Pérez, C.; Zapico-Alonso, F.; Guerrero-Bote, V. P., y de Moya-Anegón, F. (2005). Comparative analysis of webometric measurements in thematic environments, *Journal of the American Society for Information Science and Technology,* vol. 56 (8), pp. 779-785.

Goldberg, D. A. (1989). *Genetic Algorithm-in Search, Optimization and Machine Learning*, Addison-Wesley Publishing Company, Inc.

Goldfarb, A. (2006). The (teaching) role of universities in the diffusion of the Internet, *International Journal of Industrial Organization,* vol. 24 (2), pp. 203-225.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems,* University of Michigan Press, Ann Arbor, MI.

Huizingh, E. K. (2000). The content and design of web sites: an empirical study, *Information & Management,* vol. 37 (3), pp. 123-134.

Iacobucci, D. (1994). *Graphs and matrices*. En: Wasserman, S. y Faust, K. (eds.), *Social network analysis-methods and applications*. New York, NY: Cambridge University Press, pp. 92-166.

Martínez Torres, M. R., y Toral, S. L. (2010a). International Comparison of R&D Investment By European, US and Japanese Companies, *International Journal of Technology Management,* vol. 49 (1-2-3), pp. 107-122.

Martínez-Torres, M. R., y Toral, S. L. (2010b). Strategic group identification using evolutionary computation, *Expert Systems with Applications,* vol. 37 (7), pp. 4.948-4.954.

Martínez-Torres, M. R.; Toral, S. L.; Barrero, F., y Cortés, F. (2010). The role of Internet in the development of Future Software Projects, *Internet Research,* vol. 20 (1), pp. 72-86.

Miranda González, F. J., y Bañegil, T. M. (2004). Quantitative evaluation of commercial web sites: an empirical study of Spanish firms, *International Journal of Information Management,* vol. 24, pp. 313-328.

Nooy, W.; Mrvar, A., y Batagelj, V. (2005). *Exploratory Network Analysis with Pajek,* Cambridge University Press, New York.

Ortega, J. L., y Aguillo, I. F. (2008). Visualization of the Nordic academic web: Link analysis using social network tools, *Information Processing and Management,* vol. 44, pp. 1.624-1.633.

Ortega, J. L., y Aguillo, I. F. (2009). Mapping world-class universities on the web, *Information Processing and Management,* vol. 45, pp. 272-279.

Park, H. W., y Thelwall, M. (2003). Hyperlink analysis: Between networks and indicators, *Journal of Computer-Mediated Communication,* vol. 8 (4). (http://www.ascusc.org/jcmc/vol8/issue4/park.html) [consulta: mayo de 2010].

Pinto-Molina, M.; Alonso-Berrocal, J. L.; Cordón-García, J. A.; Fernández-Marcial, V.; García-Figuerola, C.; García-Marco, J.; Gómez-Camarero, C.; Zazo, Á. F., y Doucet, A. V. (2004). Análisis cualitativo de la visibilidad de la investigación de las universidades españolas a través de sus páginas web. *Revista Española de Documentación Científica,* vol. 27 (3), pp. 345-370.

Rencher, A. C. (2002): *Methods of Multivariate Analysis*. 2nd ed. Wiley Series in Probability and Statistics, John Wiley & Sons.

Robbins, S. S., y Stylianou, A. C. (2003). Global corporate web sites: an empirical investigation of content and design, *Information & Management,* vol. 40 (3), pp. 205-212.

Tan, G. W. y Wei, K. K. (2006). An empirical study of Web browsing behaviour: Towards an effective Website design, *Electronic Commerce Research and Applications,* vol. 5, pp. 261-271.

Thelwall, M. (2004). *Link Analysis: An Information Science Approach,* Amsterdam, Elsevier 2004.

Thelwall, M. (2008). Bibliometrics to webometrics, *Journal of Information Science,* vol. 34 (4), pp. 605-621.

Toral, S. L.; Martínez Torres, M. R., y Barrero, F. (2010). Analysis of Virtual Communities supporting OSS Projects using Social Network Analysis, *Information and Software Technology,* vol. 52 (3), pp. 296-303.

Toral, S. L.; Martínez-Torres, M. R., y Barrero, F. (2009a). Virtual Communities as a resource for the development of OSS projects: the case of Linux ports to embedded processors, *Behavior and Information Technology,* vol. 28 (5), pp. 405-419.

Toral, S. L.; Martínez-Torres, M. R.; Barrero, F., y Cortés, F. (2009b). An empirical study of the driving forces behind online communities, *Internet Research,* vol. 19 (4), pp. 378-392.

Toral, S. L.; Martínez-Torres, M. R., y Barrero, F. (2009c). Modelling Mailing List Behaviour in Open Source Projects: the Case of ARM Embedded Linux, *Journal of Universal Computer Science,* vol. 15 (3), pp. 648-664.

Yang, B., y Qin, J. (2008). Data collection system for link analysis, Third International Conference on Digital Information Management, pp. 247-252.