

Análisis de sesiones de la web del Cindoc: una aproximación a la minería de uso web

Artículo

Por José Luis Ortega Priego

Resumen: Se pretende realizar un estudio de usabilidad y navegabilidad de la web del Cindoc a través de los archivos de transacciones web de su servidor central durante el mes de octubre de 2003. Para ello se aplican técnicas de minería web y, en concreto, minería de uso web para la detección de sesiones que permitan determinar pautas de navegación y fallos en el diseño. Se detectan distintos problemas en los menús de navegación, en la disposición de los contenidos y en la estructura de la web. Se discuten las distintas pautas de navegación identificadas y se realizan recomendaciones sobre su diseño en futuras modificaciones.



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 Laboratorio de Internet del Centro de Información y Documentación Científica (CSIC) a cargo del proyecto Wiser (Web Indicators for Science, Technology & Innovation Research) trabajando en el ámbito de la cibermetría, minería web, visualización de información y usabilidad.

Palabras clave: Archivos de transacciones web, Análisis de sesiones web, Cindoc, Minería de uso web, Usabilidad, Navegabilidad, World wide web.

Title: Session analysis of Cindoc's web: an approach to web usage mining

Abstract: This paper try an usability and navigability study of the Cindoc web site through web log files of the main server for october 2003. For this, web mining are used, concretly, web usage mining techniques to the detection of sessions with the aim of determine navigation patterns and design faults. Several design problems are detected in the navigation menu, in the layout of the contents and in the web structure. Different navigation identificated patterns are discussed and many advices are made about its design in forward changes.

Keywords: Web log files, Web session analysis, Cindoc, Web usage mining, Usability, Navigability, World wide web.

Ortega Priego, José Luis. "Análisis de sesiones de la web del Cindoc: una aproximación a la minería de uso web". En: El profesional de la información, 2005, mayo-junio, v. 14, n. 3, pp. 190-198.

1. Introducción


HOY EN DÍA los estudios centrados en el diseño y en la arquitectura web están cobrando gran importancia ante los problemas de uso y acceso que muchas páginas presentan. El gran crecimiento que están experimentando las sedes web y la gran cantidad de información que ofrecen están creando problemas en la navegación de los usuarios.

La usabilidad, definida por Nielsen (1994) como aquello que responde a si un sistema es suficientemente bueno para satisfacer todas las necesidades y requerimientos de los usuarios, pretende evaluar los diseños para una mejor calidad en la navegación y una mayor comodidad de uso por parte del usuario. Para ello, se nutre de distintas metodologías que le permiten esta evaluación (Nielsen; Mack, 1994). El recorrido cognitivo (Lewis, et al., 1990; Polson, et al., 1992) es un

método de inspección que se centra en la evaluación de un diseño a través de la exploración. En él, un grupo de pares evalúan el diseño a través de la consecución de una o más tareas. Por otro lado, la evaluación heurística (Nielsen; Molich, 1990) es un método para la detección de problemas de usabilidad en el diseño de una interfaz e implica tener un pequeño grupo de evaluadores que examinan la interfaz y juzgan su adecuación con principios reconocidos de usabilidad.

Sin embargo, los archivos de transacciones web (*web log files*) también pueden ser analizados con la intención de conocer las acciones que realizan los usuarios en un entorno web, y con ello detectar los errores y anomalías en el diseño que puedan interferir en la navegación. Se pueden definir como archivos creados por un servidor web o proxy que recoge de forma estructurada todos las acciones o peticiones de información realizadas a una sede web en un periodo

Artículo recibido el 16-12-04
Aceptación definitiva: 08-03-05



Años de experiencia

Esto es lo que EBSCO ofrece.
Nuestro personal gestiona sus suscripciones a revistas electrónicas individuales o incluidas en paquetes de revistas, suscripciones en papel y bases de datos.

La Lista A-Z (A-to-Z) agrupa todos sus recursos electrónicos en una misma lista, para que sus usuarios localicen de manera rápida los títulos disponibles y accedan fácilmente a los contenidos. Además, la lista A-to-Z alimenta al servidor de enlaces LinkSource™, que utilizando la norma OpenURL enlaza todos sus recursos de forma compacta e inteligente.

Experiencia, servicio,
contenido, soluciones.
Hablemos hoy de sus necesidades.

CUSTOMERFOCUSEDCONTENTDRIVEN
www.ebsco.com

EBSCO
INFORMATION SERVICES

EBSCO
15
Years & Counting
1990-2005

de tiempo. Por el contrario, su análisis requiere un tratamiento complejo de los datos y la aplicación de diversas técnicas de minería de datos. Las actuales herramientas comerciales (*Funnel Web Analyzer*, *Web-Trends*, etc.) aportan limitados mecanismos para conocer la actividad de los usuarios ya que nos suministran poca información sobre las sesiones realizadas, aspecto esencial en el comportamiento del usuario dentro de una web.

De cualquier forma, los ficheros de transacciones no han sido considerados una herramienta viable para el estudio de la usabilidad web debido a dos problemas básicos:

—Información incompleta: no recogen todas las acciones de un usuario debido a las copias caché, tanto locales como proxies. Esta laguna impide conocer todas las peticiones realizadas por el usuario y qué secuencia se ha seguido para llevarlas a cabo.

—Dificultad en la identificación del usuario y de las sesiones: si no existe una identificación explícita del usuario, la de las sesiones se dificulta ya que un usuario puede utilizar diferentes máquinas y diferentes navegadores en cada momento. Por otro lado, las direcciones ip dinámicas dificultan conocer desde qué máquina se está produciendo la sesión, ya que éstas se asignan de forma aleatoria por el proveedor de servicios de internet (ISPs) a cada usuario al iniciar su sesión.

Pese a ello, la minería de datos aplicada al entorno web (*web mining*) ha aportado algoritmos y técnicas que nos permiten en cierta medida solucionar estos problemas, en concreto la minería de uso web (*web usage mining*) permite analizar de forma más concreta los archivos de transacciones con el fin de conocer el comportamiento de los usuarios en un entorno web (Tec-Ed, 1999). La aplicación de estas técnicas a los archivos de transacciones surgió en 1996 de manos de diversos autores. Mannila y Toivonen (1996) usaron las páginas de acceso como medio de descubrir rutinas. Chen et al. (1996), aportaron la identificación de diferentes sesiones de un usuario a través de los referentes de máximo avance (*maximal forward references*), esto es, el máximo grado de profundidad en la navegación antes de salir de la web o volver por el camino inverso. Yan et al. (1996), a través del sistema *Analog*, agrupan los usuarios en función de las páginas visitadas para detectar qué sesiones son más frecuentes. Recientemente, se han presentado distintos métodos para la identificación de sesiones (Huang et al., 2004; Abraham; Ramos, 2003; Spiliopoulou et al., 2003).

Con la intención de aplicar estas técnicas se ha tomado como ejemplo la web del *Centro de Información y Documentación Científica (Cindoc)*, centro depen-

diente del *Consejo Superior de Investigaciones Científicas (CSIC)* que actúa de unidad de servicios de información y documentación (construcción de tesauros, creación de bases de datos, elaboración de directorios de revistas, suministro de documentos, portales de información, etc.) para la actividad científica tanto del CSIC como del resto de la comunidad académica. Aparte, desarrolla distintas actividades de investigación en el campo de la *cienciometría*, especialmente en el desarrollo de indicadores de la actividad científica nacional.

Así pues, el objetivo del presente trabajo es estudiar la viabilidad de la minería de uso web y el análisis de sesiones como metodologías válidas para el estudio de la usabilidad y del diseño de la arquitectura web del *Cindoc*. Para ello, se analizan 2.748 sesiones obtenidas del archivo de transacciones del servidor central del centro durante el mes de octubre de 2003.

2. Metodología

La web del *Cindoc* cuenta con varios servidores, pero sólo hemos trabajado con las transacciones del principal. También hemos obviado los accesos a una de las secciones más relevantes y de mayor autonomía, la revista electrónica *Cybermetrics*, ya que anteriormente fue objeto de un estudio más pormenorizado sobre su consumo (Ortega Priego, 2004).

La web del se estructura de forma jerárquica a través de un menú de navegación donde se recogen las categorías principales en las que está estructurada. A partir de cada categoría se despliegan diversos menús que organizan el resto de los contenidos; además, varios enlaces directos (*hot links*) apuntan a diferentes secciones de la web. Por último, existen diversas sedes web anidadas (Aguillo, 1998) dentro de la principal como son las publicaciones periódicas *Cybermetrics* y la *Revista Española de Documentación Científica (REDC)*, o el *Manual de Microisís*.

2.1. Minería de uso web (web usage mining)

Se ha utilizado la minería de uso web como metodología, al igual que fue utilizada en el sistema *Web-miner* (Cooley et al., 1997, 1999). Se puede definir como el proceso de aplicar técnicas de minería de datos al descubrimiento de patrones de uso a partir de datos web (Srivastava et al., 2000). Los pasos establecidos son: limpieza de datos, identificación del usuario, identificación de la sesión y construcción de la sesión.

2.1.1. Limpieza de datos

Su objetivo es excluir todo tipo de accesos que entorpezcan el análisis. Debido a que nuestro interés está en los accesos a contenidos y páginas web, hemos eliminado los accesos redundantes de los elementos

gráficos que acompañan al contenido de un documento html. De esta forma todos los accesos a ficheros .gif, .jpg, .bmp, etc. no se tuvieron en cuenta. También se eliminaron scripts y archivos de programación (.js, .cgi, .pl, etc.) que permiten ejecutar funciones pero no aportan contenido en sí. De esta forma contamos finalmente con 185.800 accesos constituidos por documentos html y pdf.

2.1.2. Identificación del usuario

En este proceso se asocian las páginas de referencia con la misma dirección ip. Esta tarea es complicada por la existencia de cachés locales, cortafuegos corporativos y servidores proxy (Pitkow, 1997). La siguiente heurística para la identificación del usuario es usar los accesos junto con las páginas de referencia y la topología del sitio para construir sesiones de navegación. Los servidores proxy hacen que distintos usuarios tengan la misma ip. De esta forma, también identificamos el navegador, su versión y su sistema operativo. Así reducimos los accesos coincidentes y podemos detectar a usuarios diferentes. Como último caso para separar usuarios idénticos utilizamos las propias sesiones de navegación. Esto es, de una página A no podemos pasar a otra B porque no existe ningún enlace entre ellas, por lo tanto el acceso a la página A es un usuario distinto del que accede a la B.

2.1.3. Identificación de la sesión

El objetivo es dividir los accesos de un mismo usuario en distintas sesiones y la forma más simple para ello es establecer un tiempo límite. Catledge y Pitkow (1995) propusieron 25 minutos como tiempo máximo, aunque generalmente se estima que media hora entre un acceso y otro es la medida adecuada. En nuestro caso hemos contado con 30 minutos de límite.

2.1.4. Construcción de la sesión

Otro problema con el que nos podemos encontrar es que se han producido accesos que no se han recogido en el fichero de transacciones, mermando así la construcción de la sesión. Esto se produce por las copias caché que utilizan nuestros ordenadores o servidores proxy y por el uso de los botones de avance y retroceso del navegador. Como solución, podemos contrastar la página de referencia con la solicitada y comprobar si existe un enlace entre ellas y así reconstruir aquellos accesos que no han sido computados por el fichero de transacciones. Existen otros procedimientos más complejos, pero en nuestro caso sólo hemos considerado como sesiones a los accesos con las páginas de referencia identificadas.

Después de todo este proceso se han detectado 2.748 sesiones, para ello hemos establecido algunas restricciones. Sólo hemos considerado las sesiones con

tres o más accesos consecutivos, ya que un número menor nos parece insuficiente para determinar el comportamiento de los usuarios.

Para el análisis de estos datos hemos contado con una herramienta especializada en la minería de datos: el software *3dv8 Enterprise* (2004). Esta aplicación nos permite presentar gráficamente distintas características de una población a través de una representación tridimensional. Para ello todas las variables deben ser numéricas, por lo que hemos codificado los accesos en función de una estructura numérica jerárquica. De esta forma hemos considerado a la raíz con el valor 1 y a las siguientes páginas en la jerarquía se le ha añadido un dígito en función de su posición en ella. Para una mejor comprensión consideremos un ejemplo: la página principal de la *REDC* ha sido codificada con la clave 1161 porque para acceder a ella debemos seguir los siguientes pasos en la jerarquía: 1 (Raíz); 11 (“Productos y servicios”); 116 (“Revistas”); 1161 (*REDC*). De esta manera, los valores más altos son los más profundos en la jerarquía web y nos puede orientar a la hora de conocer su accesibilidad y de qué modo se accede. Por último, las páginas de referencia han sido consideradas y codificadas por simple orden alfabético.

3. Resultados

3.1. Accesos

Podemos considerar que existen tres formas de acceder a los contenidos de una página web:

- Directamente a través de su url escrita en el navegador,
- a través de un enlace situado en otra página, o
- mediante de una consulta en un buscador.

En la tabla I se puede apreciar el número y porcentajes de estas formas de acceso. En el caso de la web del *Cindoc*, la entrada a través de la url constituye el 26,67%; el de buscadores el 42,78% y el de enlaces un 30,55%. Como podemos apreciar, la mayoría de las visitas se producen por buscadores, lo que nos hace pensar que es una url difícil de recordar, que no

Formas de acceso	Sesiones %
Buscadores	1.060 42,78
Enlaces	757 30,55
Url	661 26,67
Total	2.478 100,00

Tabla I. Formas de acceso



Gráfico 1. Web del Cindoc en octubre de 2003

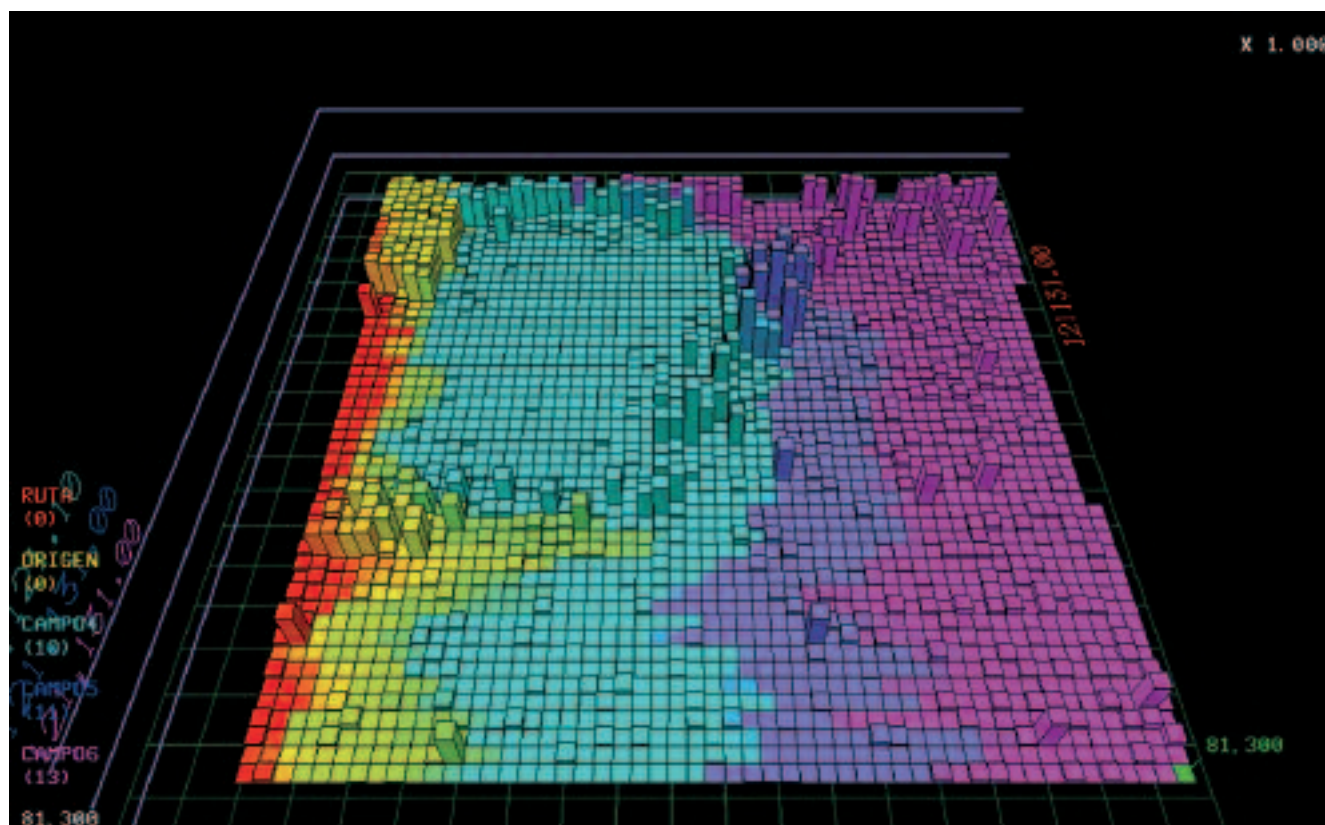


Gráfico 2. Representación tridimensional de las sesiones

Requisitos técnicos mínimos:
 Procesador Pentium o superior.
 32 Mb RAM (recomendado
 64 Mb). Windows 95, 98, Me,
 2000, NT 4, XP, Linux. Unidad
 de CD-ROM.

Sistema Integrado de Gestión Bibliotecaria

basado en ISBD/GARR, IBERMARC/MARC21 en un entorno XML diseñado específicamente para el intercambio de información en Internet.

Módulos de DIGIBIB®

- Adquisiciones.
- Catalogación.
- Autoridades.
- Circulación.
- Importación/Exportación.

Otras prestaciones de DIGIBIB®

- Gestión de objetos digitales.
- Gestión por radiofrecuencia RFID.
- Pasarela web para búsqueda, recuperación y presentación de registros y objetos digitales.
- Migración de registros.

Ú

ltima tecnología de
 creación, consulta
 e intercambio de
 información bibliográfica
 al alcance
 de todas las Bibliotecas.

DIGIBIB® es un producto de



Leada, empresa especializada en el desarrollo de Bibliotecas Virtuales en colaboración con diversas instituciones, lidera el campo de la distribución de recursos electrónicos. En nuestro catálogo de publicaciones se pueden encontrar más de 2.000 obras digitalizadas.

DIGIBIS: Producciones digitales. • Claudia Coello, 123, 4ª Planta • 28006 Madrid
 Tel.: (34) 91 581 20 01 • Fax: (34) 91 581 47 36 • digibis@digibis.com • www.digibis.com



es fácil de localizar recursos dentro de la web, etc.

3.1.1. Buscadores

En la tabla II se puede ver que el 42,7 % de los accesos se realiza a través de buscadores, de los cuales, *Google* representa el 85,5% a mucha distancia del directorio *Yahoo!* con un 5,7% y *MSN* con un 4,6%.

De estos accesos (tabla III) se ha detectado que un 43% son realizados al *Manual de Microsis*, el 26% se produce a la página principal, el 12% a las publicaciones y un 5% a la *Revista Española de Documentación Científica (REDC)*. A partir de estos

resultados podemos deducir que el uso de buscadores para acceder a la web del *Cindoc* se debe a que la información relevante está situada en los niveles más profundos de la jerarquía. Este es el caso del *Manual de Microsis* que, al tratarse de un documento de esta naturaleza, se tiende a consultar en él información puntual, lo cual parece ser más preferible hacerlo vía buscador que navegando a través del propio manual.

Buscadores	Sesiones	%
<i>Google</i>	907	85,57
<i>Yahoo!</i>	61	5,75
<i>MSN</i>	49	4,62
<i>Altavista</i>	21	1,98
<i>Terra</i>	7	0,66
<i>Otros</i>	15	1,42
Total	1.060	100,00

Tabla II. Buscadores más usados

Esto nos lleva a pensar que sus sistemas de guía o índices pueden ser deficitarios o que debería acompañarse de alguna herramienta que permitiera localizar información puntual, como puede ser un buscador interno. En este caso también está la *REDC*, cuyos contenidos están a dos clics desde la página principal.

Por otro lado, existen secciones que han sido actualizadas con nuevas páginas y urls, aunque las antiguas no han sido eliminadas del servidor, lo que provoca que sigan estando operativas. A pesar de que no son accesibles a través de la navegación por enlaces, sí lo son gracias a la navegación por directorios o llevando a cabo consultas en buscadores.

Todo ello está provocando una "navegación subterránea" de la web del *Cindoc*. En esta situación se encuentra la sección "Publicaciones" que actualmente posee el catálogo de publicaciones en el ordenador *pci204.cindoc.csic.es*, pero en su directorio (*/webpublic/*) sigue albergando información sobre publicaciones, novedades, catálogo, peticiones, etc. Es

lógico pensar que esta información se encuentra desfasada y muchos de los productos no están a la venta, pero el usuario no es consciente de ello, más aún si entra a través de un buscador.

El 12% de los accesos producidos a la página principal puede deberse a la dificultad de recordar la url del *Cindoc*, incluso olvidar o desconocer el significado del acrónimo.

3.1.2. Enlaces

El resto de los accesos se realiza a través de enlaces (26,6 %), de los cuales el más significativo es el proveniente de la web principal del *CSIC* con un 41,61% (tabla IV), a mayor distancia destacan la web de la *Sociedad Española de Documentación e Información Científica (Sedic)* con un 7% y el portal de acuicultura *Mispecies.com* con un 1,85%.

Los accesos por enlaces nos permiten saber qué partes están actuando como “puertas traseras” de entrada a una web, pero además su mayor importancia está en permitirnos conocer qué secciones son relevantes a otros recursos y qué porcentajes de visitas se cuelan por estas puertas traseras. Un problema que puede representar este grupo de accesos es que la web no esté diseñada para ello y el usuario no tenga mecanismos de navegación y de “saber dónde está” cuando entra a través de ellos. En nuestro caso hemos podido comprobar que en las entradas provenientes del *CSIC* (el 87,9%) se realizan a la página principal de la web del *Cindoc*; esto es debido a que aparece enlazado en la web principal del *CSIC* como “Búsquedas bibliográficas”. Otro caso lo encontramos en *Sedic*, donde un 86,7% de accesos desde su web se hace a la *REDC*, algo que se explica gracias a que regala un año de suscripción de la publicación al asociarse. Por último, cabe destacar el caso del portal *Mispecies.com*, ya que el 85,7% de las entradas provenientes desde él se hacen al *Centro de Documentación de Acuicultura* del *Cindoc*, único centro de documentación especializado en esta materia de España.

3.2. Análisis de sesiones (session analysis)

Nos permite conocer el camino que utilizan los usuarios para obtener la información o servicios relevantes que necesitan y que se les ofrece en un sitio web. El análisis de transac-

Recursos	Sesiones % buscadores % total		
<i>Microisis</i>	447	42,17	83,24
/*	274	25,85	22,22
Publicaciones	132	12,45	99,25
<i>REDC</i>	53	5,00	31,36
Actividad científica	29	2,74	61,70
Directorio de C. de la Tierra	21	1,98	50,00
Otros	22	2,08	0,89
Total	1.060	100,00	42,78

* directorio raíz (<http://www.cindoc.csic.es>)

Tabla III. Recursos accedidos desde buscadores

ciones es una tarea necesaria para determinar este tipo de sesiones y su frecuencia.

De las sesiones identificadas (2.748) hemos seleccionado las que arrancan desde la raíz (/) del dominio *www.cindoc.csic.es*, que son 1.233, un 49,7% del total. En el menú inicial con ventanas desplegadas que aparece en la página de inicio hemos detectado que el 39,4% de los accesos se realizan hacia el menú “Productos y servicios” y el 21,9% hacia “Acceso gratuito”. Juntos constituyen el 61,4% de los llevados a cabo desde la raíz, lo que nos informa que los servicios y fuentes de información son los contenidos de más interés para las visitas de la web. Por el contrario, la actividad científica es menos considerada con apenas 6,7% de los accesos.

Enlaces	Sesiones %	
<i>CSIC</i>	315	41,61
<i>BBDD*</i>	66	8,72
<i>pci204*</i>	57	7,53
<i>Sedic</i>	53	7,00
<i>Mispecies.com</i>	14	1,85
Otros	252	33,29
Total	757	100,00

* otros servidores de la web del *Cindoc*

Tabla IV. Enlaces que más sesiones generan

Prosiguiendo con las sesiones, hemos continuado descendiendo en la estructura de la web. A partir del menú principal hemos seguido los accesos desde cada una de las secciones del menú, con la intención de saber hacia qué elementos del submenú se dirigen. En la tabla V se es posible ver todos estos comportamientos y podemos apreciar que los dos últimos sub-menús tienen un proceder muy distinto a los tres anteriores, ya que sólo poseen una página y sus enlaces se dirigen hacia los tres primeros sub-menús.

Más interesante es el caso de los tres primeros sub-menús, los que verte-

bran el contenido de la web del *Cindoc*. En ellos podemos ver porcentajes parecidos de accesos internos y externos: el 15% de los que se realizan fuera de las opciones desplegadas en el submenú correspondiente. Pero en el caso de “Acceso gratuito”, esta proporción es el doble de los anteriores, lo que nos lleva a pensar que muchas de las opciones desplegadas en el submenú no son elegidas y por tanto los usuarios optan por otras categorías en otros submenús.

En el gráfico 2, generado por el software *3dv8 Enterprise Edition*, podemos ver los accesos más importantes por el punto de inicio en la entrada a la web del *Cindoc*. En el eje horizontal encontramos el origen de las sesiones. Así, en violeta están los accesos desde la raíz, en morado desde el *CSIC*, en azul claro desde el buscador *Google* y el resto se diluye en accesos de menor importancia. En el eje vertical la página de inicio de la sesión y la altura representa la profundidad de la sesión en la jerarquía de la web. A primera vista podemos ver que los accesos desde buscadores, y en concreto desde *Google*, son mayoritarios (85,57%).

La meseta de azul claro en la mitad superior representa los accesos al *Manual de Microsis*, como podemos ver casi la totalidad de accesos a esta parte de la web se producen desde los buscadores (42,17% desde buscadores, lo que supone un 83,24% del total de los accesos a esta sección).

La mayoría de los accesos desde la raíz no llegan a niveles muy profundos, sin embargo estos niveles profundos de la web son accedidos desde enlaces y desde buscadores.

Se puede identificar cómo se concentran los enlaces en determinados recursos como la pequeña meseta en amarillo del margen superior izquierdo, que representa los de *Sedic* a *REDC* y las alturas en morado al enlace de la web principal del *CSIC* a la biblioteca del *Cindoc*.

4. Conclusiones y recomendaciones

En vista de los resultados obtenidos podemos decir que la minería de uso web y el análisis de sesiones nos ha reportado resultados satisfactorios a la hora de analizar la usabilidad y navegabilidad de la web del *Cindoc*. A través de ella hemos podido conocer los conte-

<i>Sub-menús</i>	<i>Interno</i>	<i>% Interno</i>	<i>Externo</i>	<i>% Externo</i>	<i>Total</i>
Acceso gratuito	126	55,50	70	30,83	227
Productos y servicios	149	79,25	28	14,89	188
<i>Cindoc</i>	64	77,10	11	13,25	83
Tarifas	0	0,00	38	100,00	38
Mapa	1	1,88	50	94,33	53

Tabla V. Frecuencia de las sesiones de acceso a los submenús

nidos más demandados, las formas en que los usuarios acceden a ellos y las deficiencias en el diseño mediante el comportamiento de los usuarios. Creemos que el uso de estas metodologías basadas en la minería web, donde se constata el comportamiento de un gran número de usuarios, puede complementar otro tipo de análisis como son los recorridos cognitivos o los análisis heurísticos a la hora de conocer los problemas y deficiencias de navegación en un recurso web.

Hemos podido ver que el submenú “Acceso gratuito” induce a confusión ya que el 30% de los usuarios que acceden a él deciden salir y entrar a otro. Recomendamos que su etiqueta sea cambiada por otro término más clarificador, ya que también se confunde con el otro submenú “Productos y servicios”.

También se aprecia que existen contenidos muy demandados que se sitúan en niveles profundos de la jerarquía como el *Manual de Microsis* o la *REDC*, lo que provoca que la mayoría de sus accesos se realicen desde buscadores o enlaces externos. Sería recomendable que estos recursos tuvieran un enlace directo (*hotlink*) desde la página principal con el objeto de una mayor visibilidad dentro de la web y evitar los accesos externos desde enlaces o buscadores.

El análisis de sesiones nos ha permitido detectar aquellas secciones más frecuentes en la navegación de esta web: “Acceso gratuito” y “Productos y servicios” soportan el mayor peso de la navegación con un 61,4% de las sesiones detectadas desde la raíz, lo que nos lleva a pensar que el público accede a la web del *Cindoc* básicamente como fuente de servicios de información. Los escasos accesos al apartado de investigación nos informa del desconocimiento que esta actividad representa para el público en general. Este hecho se confirma con la escasa información existente en la web sobre esta faceta (web de grupos de investigación, publicaciones científicas de los investigadores, etc.). A nuestro entender se debería ampliar este tipo de información para un mayor conocimiento de la actividad investigadora del centro.

Nos resulta llamativo que tan sólo el 26,6% de los accesos se produzca a través de la url escrita en el navegador; nos hace suponer que la dirección, al estar compuesta por dos siglas (*cindoc.csic.es*), puede resultar algo difícil de recordar; la suposición la reforzamos con el dato de que la página principal es el segundo recurso más accedido a través de los buscadores (26%), ya que escribiendo algunos de los términos que compone el nombre del centro aparece fácilmente en las primeras posiciones de los buscadores, por ejemplo ocurre así en las búsquedas 'información y documentación' o 'documentación científica' (sin necesidad de utilizar las comillas ni las tildes).

Se recomienda también que en las actualizaciones las páginas no operativas sean borradas del servidor o su acceso sea restringido ya que está provocando una navegación "subterránea" a través de páginas desestructuradas, desactualizadas y sin conexión con la estructura lógica de la web. Todo ello lo hemos podido constatar en la sección "Publicaciones", cuya información y servicios han pasado al servidor *pci204*, pero que en el servidor central aún están accesibles los contenidos antiguos de la sección, de manera que recoge más del 12% de los accesos desde buscadores.

A través de los enlaces se produce el 30,5% de los accesos. Existen recursos que casi en exclusividad se acceden a través de enlaces externos. El hecho de ser enlazados es por sí un signo de valoración por lo que sería recomendable que fueran tenidos más en cuenta en el diseño de la web. Es destacable que muchos de estos recursos se encuentran en los niveles más profundos de la jerarquía.

Por último, el *Cindoc* posee más servidores donde se estructuran los principales servicios. Por ejemplo, *martecindoc.csic.es* alberga las peticiones del servicio de suministro de documentos; por otra parte, *pci204.cindoc.csic.es* recoge el catálogo de publicaciones. Desgraciadamente, contar sólo con las transacciones del servidor central no nos ha permitido conocer el uso y funcionamiento de esta otra parte tan importante del *Cindoc*. Futuros estudios sobre estos servidores podrían aportarnos una información más global sobre el uso de los servicios del *Cindoc* en la www.

5. Referencias

3DV8 Enterprise Edition, 2004. London: Advfn. Consultado en: 05-01-05.

<http://www.advfn.com/3dv8/>

Aguillo, I. "Hacia un concepto documental de sede web". En: *El profesional de la información*, 1998, enero-febrero, v. 7, n. 1-2, pp. 45-46.

Abraham, A.; Ramos, V. "Web usage mining using artificial ant colony clustering and genetic programming". En: *CEC03-Congress on evolutionary computation*, IEEE Press, 2003.

<http://143.129.203.3/eume/wp/eume.wp.2003-01.pdf>

Baeza-Yates, Ricardo. "Excavando la web". En: *El Profesional de la Información*, 2004, enero-febrero, v. 13, n. 1, pp. 4-10.

Catledge, L.; Pitkow, J. "Characterizing browsing behaviors on the world wide web". En: *Computer networks and ISDN systems*, 1995, v. 27, n. 6, pp. 1.065-1.073.

Cernuzzi, L.; Molas, M. L. "Integrando diferentes técnicas de data mining en procesos de web usage mining". En: *30ª Conferencia latinoamericana de informática (CLEI2004)*, 2004.
<http://clei2004.spc.org.pe/es/html/pdfs/53.pdf>

Chen, M. S.; Park, J. S.; Yu, P. S. "Data mining for path traversal patterns in a web environment". En: *Proc. 16th International conference on distributed computing systems*, 1996, pp. 385-392.

Cooley, R.; Mobasher, B.; Srivastava, J. "Data preparation for mining world wide web browsing pattern". En: *Knowledge and information systems*, 1999, v. 1, n. 1, pp. 5-32.

Cooley, R.; Mobasher, B.; Srivastava, J. "Web mining: information and pattern discovery on the world wide web". En: *Proceedings of the 9th IEEE international conference on tools with artificial intelligence (Ictai'97)*, 1997.

Huang, X.; Peng, F.; An, A.; Schuurmans, D. "Dynamic web log session identification with statistical language models". En: *Journal of american society for information science and technology*, 2004, v. 55, n. 14, pp. 1.290-1.303.

Lewis, C.; Polson, P.; Wharton, C.; Rieman, J. "Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces". En: *Proceedings of CHI 90*. N. Y.: ACM, 1990, pp. 235-242.

Mannila, H.; Toivonen, H. "Discovering generalized episodes using minimal occurrences". En: *Proc. Second international conference on knowledge discovery and data mining*, 1996, pp. 146-151.

Nielsen, J.; Molich, R. "Heuristic evaluation of user interfaces". En: *Proc. ACM CHI'90 Conf.*, 1990, pp. 249-256.

Nielsen, J. *Usability engineering*. San Francisco, CA: Morgan Kaufmann Publishers, 1994.

Nielsen, J.; Mack, R. (ed.). *Usability inspection methods*. New York: John Wiley & Sons, 1994.

Ortega Priego, J. L. "Análisis del consumo de información de una revista electrónica: análisis de ficheros log de Cybermetrics". En: *Revista española de documentación científica*, 2004, v. 27, n. 4.

Pitkow, J. "In search of reliable usage data on the www". En: *Sixth international world wide web conference*, 1997, pp. 451-463.

Polson, P.; Lewis, C.; Rieman, J.; Wharton, C. "Cognitive walkthroughs: a method for theory-based evaluation of user interfaces". En: *International journal of man-machine studies*, 1992, n. 36, pp. 741-773.

Spiliopoulou, M.; Mobasher, B.; Berendt, B.; Nakagawa, M. "A framework for the evaluation of session reconstruction heuristics in web usage analysis". En: *Inform journal on computing*, 2003, v. 15.
<http://maya.cs.depaul.edu/~mobasher/papers/informs02.ps>

Srivastava, J.; Cooley, R.; Deshpande, M.; Tan, P-T. "Web usage mining: discovery and applications of usage patterns from web data". En: *Sigkdd explorations*, 2000, v. 2, n. 1,
<http://citeseer.ist.psu.edu/srivastava00web.html>

Tec-Ed, Inc. Assessing web site usability from server log files. Ann Arbor, MI: Tec-Ed, 1999.

José Luis Ortega Priego, Laboratorio de Internet, Centro de Información y Documentación Científica (CSIC), Madrid.

Tlf: +34 915 635 482

Fax: +34 915 642 644

jortega@cindoc.csic.es