Revista Española de Documentación Científica
40(3), julio-septiembre 2017, e181
ISSN-L:0210-0614. doi: http://dx.doi.org/10.3989/redc.2017.3.1420

ESTUDIOS / *RESEARCH STUDIES*

Determinación de grupos de usuarios de bibliotecas digitales mediante el análisis de ficheros log

Juan-Antonio Martínez-Comeche

Universidad Complutense de Madrid, Facultad de Ciencias de la Documentación
Correo-e: juaamart@ucm.es | ORCID iD: http://orcid.org/0000-0001-9074-8116

Recibido: 25-09-2016; 2ª versión: 10-12-2016; Aceptado: 22-12-2016.

Cómo citar este artículo/Citation: Martínez-Comeche, J. A. (2017). Determinación de grupos de usuarios de bibliotecas digitales mediante el análisis de ficheros log. *Revista Española de Documentación Científica*, 40(3): e181. doi: http://dx.doi.org/10.3989/redc.2017.3.1420

Resumen: En este estudio se analiza el modo en que los usuarios realizan tareas de búsqueda y recuperación de información mediante consulta en la Biblioteca Digital Hispánica, distinguiendo grupos de usuarios en función de su distinto comportamiento informacional. Para ello se emplean los ficheros log recopilados por el servidor durante un año y se cotejan distintos algoritmos de agrupamiento. Se observa que el algoritmo k-means es un procedimiento de agrupamiento adecuado al análisis de extensos ficheros log de consultas en bibliotecas digitales. En el caso de la Biblioteca Digital Hispánica se distinguen tres grupos de usuarios cuyo comportamiento informacional distintivo se describe.

Palabras clave: Agrupamiento; algoritmo k-means; bibliotecas digitales; ficheros log; análisis de ficheros de transacciones; Biblioteca Digital Hispánica.

Clustering of users of digital libraries through log file analysis

Abstract: This study analyzes how users perform information retrieval tasks when introducing queries to the Hispanic Digital Library. Clusters of users are differentiated based on their distinct information behavior. The study used the log files collected by the server over a year and different possible clustering algorithms are compared. The k-means algorithm is found to be a suitable clustering method for the analysis of large log files from digital libraries. In the case of the Hispanic Digital Library the results show three clusters of users and the characteristic information behavior of each group is described.

Keywords: Clustering; k-means algorithm; digital libraries; log files; Transaction Log Analysis (TLA); Biblioteca Digital Hispánica.

1. INTRODUCCIÓN

La presencia de las bibliotecas en la Web es cada vez más frecuente, pues permite el acceso remoto a un número creciente de colecciones digitalizadas y a los metadatos incorporados a sus documentos, favoreciendo en gran medida la difusión de sus fondos. Sin embargo, la puesta a disposición de recursos en línea plantea el reto de la adecuación entre el proceso técnico llevado a cabo internamente en las bibliotecas y las necesidades de los usuarios externos a la misma. El tratamiento documental comporta una organización y disposición de los fondos que suele diferir de la estructura del conocimiento que posee el usuario, pudiendo dificultar en ocasiones la tarea de localización y acceso a la documentación deseada.

Para superar este desajuste, además de poner a disposición del usuario diversas herramientas que le ayuden en su tarea, como los buscadores, es muy importante comprender el modo de actuar de los usuarios en su interacción con las bibliotecas digitales. De este modo podremos no solo eliminar o minimizar los obstáculos que encuentren al tratar de acceder a la documentación deseada, sino que podremos mejorar el funcionamiento y eficiencia de los sistemas de recuperación. Con interfaces mejor adaptadas al modo de proceder de los usuarios, los sistemas pueden recuperar la información buscada de manera más rápida y precisa.

El análisis que presentamos aquí se enmarca, pues, dentro del campo del comportamiento informacional, entendido como el estudio de cualquier experiencia de un individuo o grupo de individuos relacionada con la necesidad, búsqueda, gestión, difusión y uso de la información en diferentes contextos (González Teruel y Barrios Cerrejón, 2012; Fisher y otros, 2005).

Dentro de este marco general, en este estudio abordaremos específicamente el modo en que los usuarios realizan búsquedas en una biblioteca digital concreta, la Biblioteca Digital Hispánica (http://bibliotecadigitalhispanica.bne.es), portal desarrollado por la Biblioteca Nacional de España que actualmente facilita el acceso a miles de documentos digitalizados de una treintena de colecciones de contenido muy diverso: material cartográfico, dibujos, discos, libros, manuscritos o música, entre otros.

Emplearemos para ello los ficheros log de uno de los servidores utilizados por la Biblioteca Digital Hispánica (BDH en adelante). Existe una amplia literatura que avala la utilidad del análisis de ficheros log como procedimiento para comprender el comportamiento informacional de los usuarios.

En cuanto que registran las interacciones producidas entre las personas y los sistemas de información accesibles en Internet (Peters, 1993), el estudio de los ficheros log es un procedimiento útil para observar cómo actúan los usuarios en la Web: análisis de redes sociales (Rechavi y Rafaeli, 2014), marcadores sociales (Borrego y Fry, 2012), gobierno electrónico (Wang y otros, 2014), weblogs (Priya y Vadivel, 2012), organismos públicos y privados (Ortega Priego, 2005) o juegos online (Wang y otros, 2011a) pueden ser objeto de investigación por este medio.

Cuatro ámbitos pueden destacarse en la literatura sobre este tema: el sanitario, el académico y el empresarial, además de las unidades de información o sistemas de recuperación de información –que abordaremos después-. En cuanto al primero, son objeto de interés reciente el comportamiento de enfermos y pacientes (Yom-Tov y otros, 2014; Van Gemert-Pijnen y otros, 2014) o el empleo de directorios de temática sanitaria (Zhang y An, 2010).

En relación al segundo, se han realizado estudios sobre sedes electrónicas de universidades (Wang y otros, 2003), revistas electrónicas (Ortega Priego, 2004), repositorios digitales (Asunka y otros, 2011), uso de materiales docentes (Hershkovitz y otros, 2014), redes de investigación (Kahlon y otros, 2014) o la influencia de conflictos bélicos en el comportamiento del personal académico (Gul y otros, 2013). Dentro del marco empresarial, por su parte, interesa desde el comercio electrónico o las transacciones web (Ma, 2013) hasta las comunicaciones entre los participantes en un determinado negocio (Stuit y Wortmann, 2012).

En estos estudios se hace hincapié en los patrones de navegación (Guerbas y otros, 2013) y en el tráfico web (Dick y otros, 2014), en ocasiones con datos recopilados con herramientas como Google Analytics (Ozen y otros, 2014) o integrando su uso con datos del ámbito concreto analizado, recopilados mediante otros medios (Iyer y Raman, 2011).

En relación al empleo de ficheros log para la observación del comportamiento de los usuarios en unidades de información o en sistemas de recuperación de información, este tipo de datos ha sido importante fuente de información en un amplio abanico de estudios que van desde los buscadores genéricos (Jansen y Spink, 2006) hasta los catálogos de bibliotecas (Villén-Rueda y otros, 2007).

En el ámbito bibliotecario, de especial interés para este trabajo, destaca el análisis de los servicios ofrecidos en las unidades correspondientes (Leeder y Lonn, 2014), como el préstamo interbibliotecario (Munson y Otto, 2013), el servicio de referencia

(Rozaklis y MacDonald, 2011) o el servicio de atención al usuario (Arnason y Reimer, 2012).

La implantación de nuevos servicios aprovechando las posibilidades tecnológicas ocupa también un lugar destacado en la literatura. Entre estos estudios, pueden apuntarse el desarrollo de un chat (Berndt-Morris y Minnis, 2014), de un sistema de recomendación personalizada (Lai y Zeng, 2013), de un catálogo interactivo (Spiteri y Tarulli, 2012) o del préstamo bibliotecario mediante el teléfono móvil (Wang y otros, 2012).

Las colecciones y los recursos disponibles son igualmente objeto de interés, como las bases de datos accesibles (Mbabu y otros, 2013), la sección de libros digitales (Ahmad y otros, 2014) o la diferencia de empleo entre las colecciones digitales y las colecciones impresas (Kapoor, 2010).

Los objetivos finales de estos artículos inciden en la mejora de los servicios y de la experiencia del usuario (Tobias y Blair, 2015) o la mejora de la usabilidad y el diseño del sitio web (Shieh, 2012).

Referente al proceso de búsqueda y recuperación de información por parte de los usuarios, objeto específico de este estudio, sobresale el análisis de las consultas (Dempsey y Valenti, 2016) con fines muy diversos: mejorar la interfaz de búsqueda (Brett y otros, 2015), rediseñar sistemas de información (Waller, 2010) o servicios (Malliari y otros, 2010), mitigar las búsquedas fallidas (Moulaison y Stanley, 2013), descubrir la clase de información más consultada (Lambert, 2013) o las estrategias de búsqueda seguidas (Shiri, 2011).

También a raíz de los datos proporcionados por los ficheros log de las consultas realizadas por los usuarios se han tratado de inferir aspectos ajenos estrictamente al proceso de recuperación, como son los objetivos o las tareas en los que se enmarcan las consultas (Strohmaier y Kroll, 2012), ciclos económicos (Chen y Tsai, 2012) o incluso la posible orientación del voto de los ciudadanos (Borra y Weber, 2012).

Para alcanzar los objetivos señalados, estas investigaciones analizan muy diversos aspectos del uso: términos de búsqueda (Zhang y Zhao, 2013), clasificación de las consultas (Maabreh y otros, 2012), duración de las sesiones de búsqueda (Park y Lee, 2016), datos demográficos de los visitantes, tráfico, detección de problemas en las páginas web o transacciones comerciales primordialmente, aspectos que los programas de análisis web (Google Analytics, por ejemplo) han ido incorporando a los informes que generan.

Sin embargo, pocos han sido los esfuerzos por descubrir grupos de usuarios en base a su comportamiento con los sistemas en línea (Chapman, 1981; Hunt y otros, 2013), más allá de los factores geográficos o del tipo de dispositivo o programas de acceso que suelen incluir las herramientas de analítica web (Clifton, 2012).

Así pues, el objetivo principal de este estudio consiste en descubrir grupos de usuarios en función del comportamiento mostrado al afrontar tareas de búsqueda y recuperación de información en la Biblioteca Digital Hispánica mediante formulación de consultas, a partir de los ficheros log de tales actividades correspondientes al año 2013. La interfaz no ha sufrido cambios desde entonces y el sistema apenas ha incorporado algunas colecciones nuevas (rollos de pianola, por ejemplo), por lo que los resultados son totalmente aplicables en la actualidad. El hecho de considerar información relativa a un año en su integridad evita la incorporación de efectos estacionales indeseados en el análisis.

Los ficheros log recopilados incluyen, tras el proceso de limpieza, un total de 195.497 sesiones de consulta, cantidad equiparable a la empleada en otros estudios sobre comportamiento de búsqueda en bibliotecas digitales: en la New Zealand Digital Library se analizaron 251.878 consultas, considerando que una sesión puede abarcar varias consultas (Mahoui y Cunningham, 2000); en el caso de la Biblioteca de la Universidad de Granada se analizaron 200.000 transacciones (Villén-Rueda y otros, 2007). En otras investigaciones el número de sesiones es muy inferior, con 2509 sesiones de búsqueda (Brett y otros, 2015), aunque puede alcanzar también cifras muy superiores cuando se recopilan datos de varios años (Park y Lee, 2016). El número de sesiones consideradas en este estudio es, pues, cantidad suficiente para avalar la representatividad de los resultados alcanzados.

Como se ha indicado anteriormente, el empleo de ficheros log referentes a episodios de recuperación de información comporta tres posibles niveles de análisis (Jansen, 2006): términos, consultas y sesiones. Dado que el objetivo del estudio consiste en delimitar grupos de usuarios con un comportamiento semejante en las tareas de búsqueda de información mediante introducción de consultas, nos centraremos inicialmente en las sesiones que incluyan consultas.

El concepto de sesión y la delimitación de su duración puede ser diverso en función del contexto y del objetivo del análisis (Mahoui y Cunningham, 2000). En nuestro caso, conforme al objetivo expuesto, concebimos una sesión como el conjunto de actividades de consulta realizadas en el sistema por un cierto usuario con una duración máxima de

24 horas. Estos límites relativos al tiempo y al tipo de actividad vienen impuestos por el método de recopilación de la información seguido en el servidor de la BDH, puesto que la visualización de documentos es gestionada por un servidor distinto al empleado en las consultas y porque el sistema estructura los ficheros log por periodos de 24 horas.

Aunque no podamos analizar las acciones realizadas por los usuarios durante la visualización, debido a la estructuración seguida en la BDH para la recopilación de datos log, el tiempo ocupado en estas tareas se añade a la sesión de consulta correspondiente, pues esta acción no cierra la sesión en curso. Este hecho implica que la sesión de consulta considerada aquí aporta información añadida sobre el proceso de búsqueda y recuperación en su totalidad, más allá de la formulación de la consulta y la consulta de la lista de resultados.

Por su parte, partiremos de un concepto de consulta (Jansen y Pooch, 2001) que incluye las dos modalidades que permite el sistema Solr utilizado en la BDH: la consulta sencilla (mediante la introducción de términos en un área de texto) y la consulta avanzada (mediante la cumplimentación de un formulario con diversos campos relativos a los principales puntos de acceso: autor, título, edición, fechas o tipo de documento, entre otros). Estas dos modalidades de consulta fueron ya descritas entre los estudios iniciales de los catálogos en línea (Hancock-Beaulieu, 1989).

Una particularidad del sistema de recuperación de información de la BDH consiste en la posibilidad de filtrar los resultados obtenidos por cualquier modalidad de consulta, gracias a la inclusión de seis filtros en el sistema: acceso temático, tipo de material, colecciones destacadas, autor, lengua y año. Ello permite al usuario seleccionar características añadidas a la documentación aportada inicialmente por el sistema en respuesta a la consulta formulada.

El análisis del empleo de filtros por parte de los usuarios es de especial interés para los responsables de la BDH con el fin de complementar los datos que sobre ellos aporta el software de análisis web utilizado. En consecuencia, será objetivo prioritario de este estudio el análisis del empleo de los distintos filtros disponibles, incluyendo los diversos filtros como los factores principales para la constitución de los diversos grupos de usuarios.

En función de los objetivos expuestos anteriormente y de la definición de sesión adoptada, será precisamente la sesión el eje nuclear que nos permitirá delimitar comportamientos distintos en los usuarios. En consecuencia, nos detendremos inicialmente en aquellos aspectos cuantitativos relativos a las sesiones detectadas que configurarán, a su vez, los factores que permitan caracterizar los distintos grupos de usuarios: duración, modalidad de consulta y filtros empleados.

Con el objeto de conocer al mismo tiempo las características generales del sistema analizado, mostraremos también el número de sesiones y su distribución por día y mes del año, aunque estos datos no se hayan empleado en el estudio de grupos de usuarios, objetivo primero de este estudio.

Una vez determinados los grupos de usuarios de la BDH, analizaremos los términos presentes en las consultas formuladas por los miembros de cada grupo, con el fin de mostrar algunas opciones de mejora del sistema de recuperación en función de su modo de proceder.

2. METODOLOGÍA

La metodología empleada para desvelar grupos de usuarios con características semejantes en función del comportamiento mostrado al realizar consultas, tal como el sistema los ha registrado en los ficheros log, es de carácter cuantitativo, entendiendo bajo esta denominación la combinación de procedimientos esencialmente matemáticos y estadísticos que permiten alcanzar respuestas, entre otros ámbitos, en la investigación social y del comportamiento (Vogt, 2011). El motivo que justifica esta elección es la conveniencia de aplicar algoritmos matemáticos comúnmente utilizados para afrontar el problema del agrupamiento o clustering de datos en relación a un conjunto de factores concreto y conforme a la noción de similitud o distancia adoptada (Kaufman y Rousseeuw, 2005).

Pueden distinguirse cuatro fases en el desarrollo de este tipo de estudios (Adèr y otros, 2008):

1. Recopilación de datos. Como se ha señalado previamente, los responsables de la BDH han facilitado los 365 ficheros de texto plano recogidos día a día por el servidor Solr que gestiona el sistema de recuperación de información de este organismo durante el año 2013. Dado que no es el único servidor involucrado durante una visita de un usuario al sitio web de la BDH, no es posible analizar el proceso íntegro seguido por los usuarios, sino tan solo las actividades de consulta en el buscador.
2. Procesamiento de datos. Con el fin de preservar el anonimato en la identidad de los usuarios, los responsables de la BDH sometieron los ficheros a un proceso de modificación de las direcciones IP en cada uno de los archivos que recoge la actividad producida en el servidor durante un día, al tiempo que se han reunido en una única línea todos los valores adoptados

por las variables durante la consulta. Así, cada sesión de consulta queda reflejada en un único estado final. Ello impide abordar las sesiones de consulta como un proceso compuesto de diversas acciones y, por otra parte, implica que puede aparecer una misma IP en dos días distintos sin que ello necesariamente suponga un mismo usuario, equipo o lugar de conexión con la BDH. De ahí que el análisis sea de carácter estático, sin entrar en la dinámica del proceso de consulta, y que nuestro concepto de sesión incorpore un límite temporal de 24 horas. De esta manera, cada dirección IP durante un cierto día permite identificar al mismo usuario que ha accedido al sistema.

3. Limpieza de datos. Una vez procesados los datos, cada uno de los ficheros con la actividad diaria se incorporó, mediante el software libre de tratamiento estadístico R (https://www.r-project.org), a tablas donde las filas recogen cada una de las sesiones de consulta realizadas al servidor por parte de los usuarios y las columnas las distintas variables incorporadas a las peticiones con sus valores respectivos. Se han desarrollado diversos scripts en el propio lenguaje R, con el objeto de estructurar cada sesión en las filas y de discernir en las columnas los valores que presenta cada variable (correspondientes básicamente a cada uno de los seis filtros, tipo de consulta –sencilla o avanzada-, duración de la sesión, dirección IP y día del año) en las respectivas consultas. Una vez incorporados los datos a tablas legibles en R, durante esta tercera fase se procedió a limpiar los datos, eliminando los casos de sesiones con datos incorrectos (duraciones negativas o nulas), las peticiones duplicadas (las consultas avanzadas recogidas inicialmente como "AdvancedSearch.do" y posteriormente como "Search.do") o las peticiones no consideradas en este estudio (como las solicitudes de visualización de datos, recogidas mediante la acción "CompleteSearch.do"). No se consideraron datos atípicos, dado el número significativo de sesiones con cualquier duración, incluyendo las muy cortas y las extremadamente prolongadas. El conjunto final empleado en este estudio incluye un total de 195.497 sesiones de consulta producidas durante el año.

4. Análisis de datos. Esta cuarta fase abarca tanto un análisis inicial del número de sesiones producidas por día y mes del año, con el objeto de tener una visión global del sistema, como el estudio de los ocho factores o parámetros cuantitativos disponibles para establecer posteriormente los distintos grupos de usuarios:

   a. Duración de la sesión.
   b. Tipología de consulta empleada en la sesión: sencilla o avanzada.
   c. Filtros empleados en la sesión, de los seis posibles: acceso temático, tipo de material, colecciones destacadas, autor, lengua y año.

Para ello se elaboraron diversos scripts en lenguaje R. Por último, se sometieron los valores de estos ocho parámetros para cada una de las 195.497 sesiones de consulta al algoritmo k-means de agrupamiento, analizándose los resultados y características de cada uno de los grupos obtenidos. Posteriormente se cotejó este algoritmo de agrupamiento con otros posibles procedimientos de clustering para observar sus diferencias cuando manejamos volúmenes grandes de datos, como es el caso de las consultas realizadas a una biblioteca digital de carácter nacional. Para el procesamiento de los distintos algoritmos de agrupamiento se emplearon los scripts disponibles en diversos paquetes del lenguaje R.

3. RESULTADOS

Los resultados del estudio se organizan en los siguientes apartados: en primer lugar, se presentan las características generales de las sesiones de consulta producidas durante el año, a fin de tener una visión general del sistema; a continuación, se muestran los valores obtenidos –con carácter global- en los distintos factores o parámetros empleados posteriormente en el cálculo de los grupos de usuarios (duración de las sesiones, tipología de consulta empleada en las mismas y porcentajes de utilización de los filtros por parte de los usuarios). Por último, se aborda la determinación de los grupos de usuarios y sus características más destacadas.

3.1. Número de sesiones por día y por mes

De las 195.497 sesiones de consulta efectuadas en la BDH durante el año, se han obtenido los siguientes valores medios:

- Media de sesiones por día: 535,61
- Media de sesiones por mes: 16291,42

Los usuarios mantienen un nivel de actividad parecido durante todo el año, con un cierto aumento durante los últimos cuatro meses del año. La máxima actividad se produce en el mes de septiembre (un 11,03%) y el mes de julio presenta la mínima actividad del año con un 6,47%.

Considerando las sesiones durante la semana, se produce un nivel de actividad semejante durante todos los días, con un ligero descenso durante el

fin de semana. Los lunes registran el nivel máximo de actividad (un 16,12%), mientras que los sábados presentan el nivel mínimo con un 10,83%.

3.2. Duración de las sesiones

Las sesiones presentan, en cuanto a su duración, valores en todo el intervalo comprendido entre un mínimo de 1 segundo y un valor máximo de 24 horas. La mitad de las sesiones presentan una duración menor o igual a 7,12 minutos, con una duración media de 102,20 minutos. Este valor medio tan elevado se justifica por el número significativo de sesiones con duraciones muy elevadas (1057, en concreto, duran entre 23 y 24 horas). En la Tabla I se resumen los tiempos de las sesiones distribuidos en intervalos, junto con su porcentaje.

3.3. Tipología de consulta empleada en las sesiones

Son posibles dos modalidades de consulta en las sesiones consideradas: consulta sencilla y consulta avanzada. Predominan en el corpus, con un 64,90%, las consultas sencillas en las que el usuario se limita a introducir una cadena de caracteres en el área de texto desarrollada al efecto, frente al 35,10% que ha optado por las consultas avanzadas durante la sesión, en las que el usuario señala valores en relación a algunos de los puntos de acceso del catálogo (autor, título o tipo de documento, por ejemplo).

Como dato complementario, en la tabla II se recogen los campos más frecuentemente utilizados por los usuarios cuando cumplimentan el formulario de la búsqueda avanzada (sumando los valores de las tres variables "field" que las acogen).

3.4. Empleo de filtros en las consultas

El sistema de recuperación de información empleado en la BDH permite restringir los resultados iniciales mostrados ante una consulta del usuario (ya sea sencilla o avanzada), mostrando en el lateral izquierdo de la pantalla seis filtros desplegables: "acceso temático", "tipo de material", "colecciones destacadas", "autor", "lengua" y "año". De este modo, el usuario puede –sin necesidad de reformular la consulta- discernir estas características añadidas en los documentos hallados por el sistema en respuesta a su pregunta.

Los datos muestran que, de las 195.497 sesiones consideradas, en 116.216 sesiones (un 59,45% de los casos) los usuarios no utilizaron esta he-

Tabla I. Duración de las sesiones

Duración de la sesión	Frecuencia	Porcentaje
<=1 min.	45.961	23,51
>1 y <=2 min.	16.435	8,41
>2 y <=3 min.	11.020	5,64
>3 y <=4 min.	8.072	4,13
>4 y <=5 min.	6.333	3,24
>5 y <=6 min.	5.138	2,63
>6 y <=7 min.	4.342	2,22
>7 y <=8 min.	3.703	1,89
>8 y <=9 min.	3.357	1,72
>9 y <=10 min.	2.882	1,47
>10 y <=20 min.	18.093	9,26
>20 y <=30 min.	9.692	4,96
>30 y <=40 min.	6.575	3,36
>40 y <=50 min.	4.898	2,51
>50 y <=60 min.	3.782	1,93
>60 y <=70 min.	3.053	1,56
>70 y <=80 min.	2.468	1,26
>80 y <=90 min.	2.100	1,07
>90 y <=100 min.	1.760	0,90
>100 y <=1440 min.	35.833	18,33

Tabla II. Frecuencia de selección de campos en búsqueda avanzada

Campo seleccionado	Frecuencia	Porcentaje de empleo
Todos los campos	55.441	49,95
Autor	36.107	32,53
Materia	8.526	7,68
Lugar de publicación	4.214	3,80
Colección	3.971	3,58
Título	1.393	1,26
Otros	1.340	1,21

rramienta, mientras que en las 79281 sesiones restantes –lo que supone un 40,55% del total- los usuarios se sirvieron de esta herramienta para precisar los resultados deseados. En la Tabla III se puede consultar el número de sesiones en que se utilizó cada uno de los filtros, junto con el porcentaje del total de sesiones que representa.

De la tabla se concluye que el filtro más empleado por los usuarios fue el de "tipo de material" (en un 19,71% de las sesiones), seguido a bastante distancia por el filtro de "colecciones destacadas" (un 10,46%) y el filtro de "acceso temático" (un 7,93%). De igual forma, resalta la muy escasa utilización de los filtros de "año", "autor" y "lengua", que en ningún caso alcanzan un 2,5% de las sesiones.

3.5. Grupos de usuarios

La partición de las 195.497 sesiones en grupos se realizará mediante la aplicación del algoritmo k-means, debido a que no existen factores o características irrelevantes o que introduzcan ruido (Amorim y Mirkin, 2012), a que se trata de un algoritmo de poca complejidad, por lo que es adecuado para conjuntos de datos grandes (Celebi y otros, 2013), y a que es una técnica no supervisada, por lo que es adecuado para colecciones de datos no etiquetados o clasificados a priori como sucede en la BDH (Verma y otros, 2012).

Sus ventajas lo han convertido en uno de los algoritmos más frecuentemente empleados (Steinbach y otros, 2000). Entre los inconvenientes señalados en relación a este procedimiento destaca la decisión previa sobre el número de grupos que desean formarse (Amorim y Hennig, 2015; Huang, 1998). Para resolver este problema, se ha aplicado uno de los criterios más conocidos disponibles estadísticamente, denominado Sum of Squared Errors (SSE) (Celebi y otros, 2013; Schalkoff, 2001). En consecuencia, se establecen los siguientes apartados en esta sección, correspondientes a las sucesivas fases llevadas a cabo: Determinación del número de grupos, validación del agrupamiento y análisis de los grupos.

Determinación del número de grupos

SSE calcula la suma de los cuadrados de las distancias entre cada elemento de un grupo y su centroide, de manera que un número apropiado de grupos sería aquel en el que la disminución del SSE es más drástica (Peeples, 2011). La fórmula empleada es la siguiente, donde 'n' es el número de elementos de cada grupo:

$$SSE = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Tabla III. Frecuencia de empleo de los filtros en las consultas

Filtro	Frecuencia	Porcentaje de sesiones
Acceso temático	15.498	7,93
Tipo de material	38.536	19,71
Colecciones destacadas	20.449	10,46
Autor	3.531	1,81
Lengua	1.162	0,59
Año	4.334	2,22

En nuestro caso, como puede observarse en la Figura 1, el codo –esto es, el punto donde la reducción del valor de SSE es mayor- corresponde a N= 3 grupos.

En efecto, la diferencia entre N=2 y N=3 (en relación a la suma de cuadrados de las distancias intra-grupos) es mayor que la diferencia entre N= 3 y N= 4 (las cifras exactas pueden consultarse en la Tabla IV: la diferencia de SSE entre N=2 y N=3 es de 1,54 e+09, mientras que la diferencia de SSE entre N=3 y N=4 es de 4,74 e+08).

Una vez determinado que tres es el número de grupos inicialmente más adecuado para nuestro corpus de datos, sometemos los valores binarios (aplicación -1- o no -0-) de los 7 factores (6 filtros y la consulta avanzada) más el tiempo ocupado por cada una de las 195.497 sesiones al algoritmo k-means conforme al modelo clásico (MacQueen, 1967).

Validación del agrupamiento

Emplearemos dos criterios internos (intra-grupos) y dos criterios externos (inter-grupos) para comparar diversas posibilidades de particiones próximas entre sí (dos, tres, cuatro y cinco grupos mediante k-means):

- Criterios internos: índice Calinski-Harabasz e índice Silhouette
- Criterios externos: índice Pearson Hubert's Gamma e índice Single Link Average.

Estos criterios, a los que se añade el índice SSE empleado anteriormente, nos permitirán validar la elección adoptada. Los criterios internos proporcionan información sobre el grado de cohesión, relación o semejanza entre los elementos dentro de cada grupo. Los criterios externos, por su parte, indican en qué medida los grupos poseen límites cla-

Figura 1. Determinación del número de grupos mediante SSE

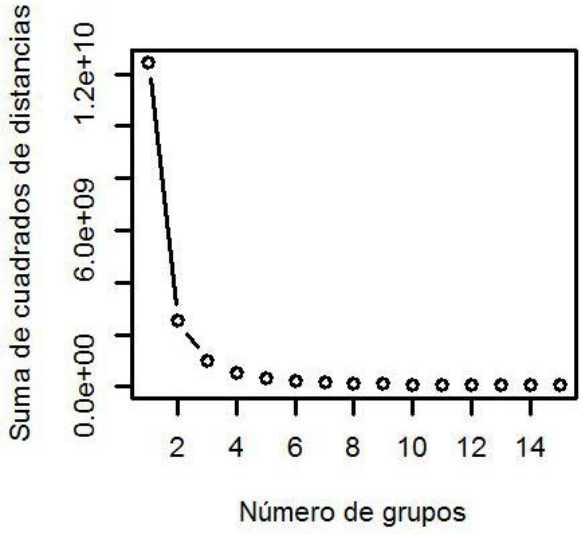

Tabla IV. Criterios intra-grupos e inter-grupos de validación del agrupamiento

K-medias	Criterios intra-grupos			Criterios inter-grupos	
	Calinski-Harabasz	Silhouette	SSE	Pearson Hubert Gamma	Single Link (Media)
2 grupos	757603	0,74	2,55 e+09	0,88	0,1
3 grupos	1107739	0,68	1,01 e+09	0,82	187,35
4 grupos	1443773	0,64	5,36 e+08	0,76	269,04
5 grupos	1714638	0,61	3,44 e+08	0,76	309,76

ros con respecto a los demás, cuantificando el nivel de desemejanza o dispersión entre los grupos. Un determinado agrupamiento será mejor cuanta mayor cohesión o semejanza exista entre los elementos de cada grupo y, a la vez, presente una mayor desemejanza entre los grupos. En consecuencia, cuanto mayor valor posea un cierto agrupamiento, en relación a cualquiera de los criterios internos o externos, mejor partición se considera en relación a dicho índice (Maulik y Bandyopadhyay, 2002).

El índice Calinski-Harabasz se basa en la relación entre la varianza entre los grupos y la varianza dentro de los grupos, de manera que un valor mayor del cociente indica una mejor partición (Calinski y Harabasz, 1974). El índice Silhouette compara las distancias de cada elemento con el resto de puntos de su grupo y con los demás puntos pertenecientes a los grupos restantes. Su valor oscila entre 1 (los elementos están apropiadamente situados en sus grupos) y -1 (los elementos deberían adscribirse a otros grupos) (Rousseeuw, 1987; Liu y otros, 2010).

En cuanto a los criterios externos, la versión de Pearson del coeficiente Gamma de Hubert halla la correlación entre las distancias entre todos los pares de elementos y el vector 0-1, donde 0 significa mismo grupo y 1 significa grupos o clusters diferentes (Halkidi y otros, 2001). El índice Single Linkage señala la distancia más corta entre dos elementos pertenecientes a dos grupos o clusters diferentes (Zhu y otros, 2015). En la Tabla IV se incluye la media de dichas distancias inter-grupos.

Como puede observarse en la tabla, las particiones en 2, 4 y 5 grupos presentan un mínimo en al menos un criterio: el agrupamiento en 2 conjuntos posee el valor mínimo en Calinski-Harabasz y en Single Link; el agrupamiento en 4 conjuntos o clusters posee el valor mínimo en el índice Pearson Hubert Gamma; la partición en 5 grupos posee el valor mínimo en el índice Silhouette y en el índice Pearson Hubert Gamma (junto al agrupamiento en 4 clusters).

De estos resultados se concluye que la partición en tres grupos no solo posee el máximo valor en el índice SSE, sino que no presenta tampoco ningún mínimo en los cuatro índices restantes. En consecuencia, se confirma la partición en tres grupos como una solución apropiada al corpus de datos.

Análisis de los grupos

El análisis cuantitativo de los tres grupos en relación a los ocho parámetros considerados en el algoritmo de agrupamiento se estructura en los siguientes puntos esenciales, pudiéndose consultar un resumen en la Tabla V.

Grupo 1

El primer grupo (que podemos denominar de usuarios experimentados) se caracteriza por los siguientes aspectos:

- Es el grupo más numeroso (172.402 sesiones) al acoger el 88,19% del total de 195.497 sesiones y, en consecuencia, agrupa la mayoría de los usuarios de la BDH.
- Este grupo engloba los usuarios cuyas sesiones de consulta ocupan el menor tiempo de conexión. En relación a este factor, los usuarios se conectan una media de 24 minutos con el buscador de la BDH, con un máximo de 237 minutos.
- Necesidad informativa centrada en la localización de documentos.
- Aproximadamente un 35% de este grupo utiliza la búsqueda avanzada, lo que supone un porcentaje igual a la media general.
- El filtro más frecuentemente utilizado por los usuarios de este grupo es el filtro "tipo de material" (un 20% de las sesiones), seguido del filtro de colecciones (11%) y del filtro temático (8%), en porcentajes semejantes a la media general.

Grupo 2

En el segundo grupo (que podemos denominar de usuarios poco experimentados) pueden destacarse los siguientes aspectos:

- Es el grupo menos numeroso (7.781 sesiones) con un 3,98% del total de 195.497 sesiones, por lo que agrupa a una minoría de los usuarios de la BDH.
- Este grupo incluye las sesiones más prolongadas; los usuarios se conectaron un mínimo de 799 minutos (13 horas) y un máximo de 1440 minutos (24 horas), con una media de 1147 minutos (19 horas).
- Necesidad informativa centrada en la visualización de documentos (para su análisis, copia y/o impresión).
- Aproximadamente en un 40% de estas sesiones se utilizó la búsqueda avanzada, en un porcentaje superior al resto de los grupos y a la media general.
- El filtro más frecuentemente utilizado por los usuarios de este segundo grupo es el filtro "tipo de material", aunque con un porcentaje inferior al grupo 1 (un 18% de las sesiones frente al 20% de las sesiones del grupo 1),

Tabla V. Factores y aspectos distintivos en los grupos de usuarios de la BDH

	Nº de sesiones	Duración media de sesiones (min)	Porcentaje de consultas avanzadas	Filtros más empleados (porcentaje de sesiones)	Aspectos distintivos
Valores globales	195497 (100%)	102,2	35,10	*Tipo de material (19,71) *Colecciones (10,46) *Temático (7,93)	
Grupo 1 (usuarios experimentados)	172402 (88,2%)	24	35,17	*Tipo de material (19,88) *Colecciones (10,77) *Temático (8,02)	*Grupo más numeroso (88,2%) *Menor tiempo de conexión * Necesidad informativa centrada en la localización de documentos *Valores semejantes a la media global en el empleo de filtros y consultas avanzadas
Grupo 2 (usuarios poco experimentados)	7781 (3,98%)	1147	39,35	*Tipo de material (18,35) *Autor (10,67) *Temático (10,53)	*Grupo menos numeroso (3,98%) *Mayor tiempo de conexión *Necesidad informativa centrada en la consulta o visualización de documentos *Mayor empleo de filtros de autor y temático
Grupo 3 (usuarios muy experimentados)	15314 (7,83%)	451,3	32,17	*Tipo de material (18,51) *Colecciones (8,15) *Temático (5,59)	*Menor empleo de consulta avanzada *Menor empleo de filtro de colecciones *Sesiones prolongadas *Necesidad informativa centrada en la consulta o visualización de documentos *Menor empleo de filtro temático

seguido de los filtros temático y de autor (un 11%) y del filtro de colecciones (8%). Hay, en consecuencia, un aspecto distintivo en este grupo consistente en un aumento en el empleo de los filtros de autor (que pasa del 2% general al 11%) y temático (que pasa del 8% general al 11%) y en una disminución del filtro de colecciones con respecto al primer grupo y a la media general (pasando del 10% al 8%). También se observa un aumento del empleo del filtro de lengua (0,9%, frente a un 0,6% en el total del corpus).

Grupo 3

El tercer grupo (que podemos denominar de usuarios muy experimentados) se caracteriza por los siguientes aspectos distintivos:

- Es un grupo poco numeroso (15314 sesiones), con un 7,83% del total de sesiones, lo que muestra la uniformidad general en el comportamiento de los usuarios de la BDH en tareas de consulta, agrupados mayoritariamente en el primer grupo.
- Los usuarios de este tercer grupo se caracterizan por sesiones prolongadas, entre 238

minutos (casi 4 horas) y 798 minutos (aproximadamente 13 horas), con una media de 451,3 minutos (7,5 horas).

- Necesidad informativa centrada en la visualización de documentos (para su análisis, copia y/o impresión).
- Los usuarios de este grupo son los que menos utilizaron la búsqueda avanzada durante sus consultas (un 32% de las sesiones, frente a un 40% en el grupo 2 y a un 35% en el grupo 1).
- El filtro más frecuentemente utilizado por los usuarios de este grupo es también el filtro "tipo de material", en un porcentaje (18,5%) parecido al resto de los grupos y a la media general (19,7%). A continuación figuran el filtro de colecciones (8%) y el filtro temático (un 5,6%). Este grupo se caracteriza, pues, por una disminución en el empleo de los filtros temático y de colecciones frente a la media general (8% y 10,5% respectivamente).

Una explicación para las sesiones prolongadas de los grupos segundo y tercero consiste en que las necesidades informativas de estos usuarios no se limitan a la búsqueda y localización de determinados textos o ejemplares de documentos conservados en la Biblioteca Nacional de España (como en el primer grupo), sino que se centra en la posterior visualización o consulta de las copias digitalizadas de tales documentos, bien para su análisis o estudio, bien para su copia y/o impresión, o para ambas tareas. Como se señaló en la introducción, la visualización o acceso a los documentos es gestionada por un servidor distinto al empleado en las consultas, pero siempre que el usuario regresa a la página de resultados tras la visualización (para ver otros documentos del listado o para acceder al contenido de otro u otros documentos relevantes recuperados), dicha acción quedará reflejada en el servidor empleado en la consulta como parte de su sesión, englobando así el tiempo empleado en la visualización. Por otra parte, Park y Lee (Park y Lee, 2013) -en el análisis de las consultas al sistema de recuperación del Instituto de Ciencia y Tecnología de Corea- han detectado sesiones de 10 horas, concluyendo que las sesiones en un sistema de recuperación de información son más prolongadas que en un buscador genérico.

En la Tabla VI se resumen algunas características de los tres grupos de usuarios de la BDH relativas a consultas, términos, colecciones y filtros. En el caso de las colecciones más consultadas, debe tenerse en cuenta que la solicitud de la colección "Cartas náuticas" queda reflejada en los ficheros log dentro del área de texto, motivo por el cual figura como una consulta más. Tampoco las solicitudes de las colecciones "Fotografía" y "Grabados de Rembrandt" se registran como el resto, motivo por el que no se incluyen en los datos mostrados.

Por otra parte, el análisis mediante clustering jerárquico de los filtros ha permitido obtener el nivel de correlación entre los filtros en cada uno de los tres grupos de usuarios, indicador de la utilización simultánea de los mismos. De dicho análisis se desprende también que la búsqueda avanzada no está relacionada ni con la utilización de ciertos filtros ni con la consulta de determinadas colecciones, por lo que su empleo puede considerarse independiente de ellas.

4. DISCUSIÓN

Se ha destacado en diversas ocasiones que los usuarios, cuando se enfrentan a un sistema de recuperación de información, suelen realizar consultas sencillas con un número de términos pequeño (Markey, 2007). En el caso de los motores de búsqueda genéricos, Spink y Jansen (Spink y Jansen, 2004) confirman que el empleo de características avanzadas en los buscadores genéricos es muy baja (en torno al 5%), resaltando el caso de Altavista, que en 2002 alcanzó un 27,6% de utilización de operadores de búsqueda en las consultas efectuadas.

En el caso del OPAC de una biblioteca universitaria, por el contrario, se ha señalado una utilización similar de ambas modalidades de consulta, fenómeno explicado porque los usuarios, cuanto mayor conocimiento de la información que están buscando y más familiaridad con el OPAC poseen, optan en mayor medida por las búsquedas sencillas (Villén-Rueda y otros, 2007).

El hecho de que las colecciones de la BDH sean especializadas y posean un alto valor histórico restringe en buena medida los usuarios potenciales, que en mayor proporción saben con precisión qué documentación buscan y qué biblioteca la conserva. En consecuencia, el carácter más especializado de la documentación conservada y la mayor proporción de usuarios con un conocimiento previo de sus necesidades informativas y del sistema de recuperación entre los que acceden a esos fondos permite justificar la preponderancia de las consultas sencillas (casi un 65%) sobre las consultas avanzadas (un 35%), tanto con carácter general como específicamente en el grupo 1 que mayoritariamente los acoge.

Tabla VI. Características de los grupos relativas a consultas, términos, colecciones y filtros

	Consultas más frecuentes	Longitud media de consultas (nº de términos)	Términos más frecuentes (eliminadas las palabras vacías)	Co-términos más frecuentes (eliminadas las palabras vacías)	Colecciones más frecuentes	Filtros de mayor utilización simultánea
Valores globales	"Avicena" "guitarra" "partituras" "portulano" "Filipinas"	1,8	"Juan" "historia" "España" "Antonio" "san"	"carta"+"náutica" "Puerto"+"Rico" "Johann"+"Strauss" "Josef"+"Strauss" "Carlos"+"Gardel"	*Obras maestras *Carteles *Grabados de Durero *Ephemera *Viajes	* Tipo de material * Lengua
Grupo 1	"Avicena" "guitarra" "partituras" "portulano" "carta náutica"	1,8	"Juan" "historia" "España" "Francisco" "Antonio"	"carta"+"náutica" "Puerto"+"Rico" "Johann"+"Strauss" "Josef"+"Strauss" "Carlos"+"Gardel"	*Obras maestras *Carteles *Grabados de Durero *Ephemera *Libros de caballería	* Tipo de material * Lengua
Grupo 2	"calzada romana" "Puente la Reina" "Nasarre" "guitarra" "Puerto Rico"	1'9	"calzada" "romana" "reina" "España" "puente"	"calzada"+"romana" "Puente"+"Reina" "Puerto"+"Rico" "Manuel"+"Navarro" "ejército"+"español"	*Dibujos de los niños de la guerra *Grabado contemporáneo *Hispanoamérica *Obras maestras *Carteles	* Autor * Temático
Grupo 3	"Thesaurus Sacrarum Historiarum" "el toreo" "partituras" "guitarra" "óperas"	1'8	"Juan" "Antonio" "historia" "san" "España"	[Combinaciones de dos términos entre "Thesaurus", "Sacrarum" e "Historiarum"] "Puerto"+"Rico" "Edvard"+"Grieg" "Guerra"+"Independencia" "Don"+"Quijote"	*Obras maestras *Carteles *Grabados de Durero *Ephemera *Dibujos de los niños de la guerra	* Temático * Autor

Es por ello que hemos empleado para designar al grupo 1 de usuarios, el más numeroso y con iguales valores a los generales, la denominación de usuarios experimentados. Los grupos restantes confirman esta argumentación: el segundo grupo de usuarios, menos experimentado, debería presentar, en consecuencia, un porcentaje mayor de consultas avanzadas, como así es (con un 39,35%); al tercer grupo de usuarios, muy experimentados, le correspondería un menor empleo de consultas avanzadas, como en efecto sucede (con un 32,17%).

El empleo de filtros por parte de los distintos grupos confirma el razonamiento basado en el grado de conocimiento de los documentos buscados y la experiencia previa en el empleo del sistema de recuperación de la BDH, aunque con menor intensidad. Los valores globales dan un 40,55% de utilización de algún filtro, frente a un 59,45% de sesiones en que no se utilizó ningún filtro.

Tanto el primer grupo (la mayoría de los usuarios, con experiencia y conocimientos previos) como el segundo grupo (una minoría de usuarios, menos experimentados) presentan un grado de utilización parejo a estos valores generales: un 41,05% de las sesiones del grupo 1 y un 40,03% de las sesiones del grupo 2 utilizaron alguno de los filtros, no percibiéndose una diferencia significativa. El segundo grupo, sin embargo, se distingue por un mayor empleo de los filtros temático y de autor que el grupo 1 y que los usuarios de la BDH en general, lo que reflejaría unas necesidades informativas menos delimitadas por parte de este grupo de usuarios.

Por su parte, el tercer grupo (de usuarios más experimentados que la media) presenta un grado significativamente menor de empleo de algún filtro (un 35,24%) y específicamente de los filtros temático y de colecciones, acorde con una concreción mayor del documento deseado y una menor necesidad de herramientas que permitan restringir la búsqueda.

Estos datos confirman los resultados obtenidos por Ferl y Millsap, quienes destacan que una proporción mucho mayor de estudiantes (usuarios menos experimentados) realizaron búsque-

das por temas que el profesorado y el personal de la biblioteca de la Universidad de California (usuarios más experimentados) (Ferl y Millsap, 1996). Larson había identificado anteriormente los problemas del acceso por encabezamientos temáticos en los OPACs, entre los que enunciaba precisamente la falta de conocimiento por parte de los usuarios de dicha herramienta de acceso (Library of Congress Subject Headings o LCSH) (Larson, 1991).

En relación a la duración de las sesiones, muy diversos factores influyen en una estancia más o menos prolongada por parte de los usuarios, pudiéndose destacar el contenido del sitio web, la tarea que afronta el usuario (en la que se enmarca la búsqueda de información) y las características del usuario (Lalmas y otros, 2014).

El tiempo transcurrido en un cierto sitio web depende en buena medida de la clase de contenidos que oferte. Se ha constatado que los sitios web a los que se acude a consultar noticias presentan sesiones mucho más cortas que los sitios que involucran la realización de procesos o actividades durante la visita, como el comercio electrónico o las redes sociales (Benevenuto y otros, 2009). Por este motivo las sesiones detectadas en los buscadores generalistas son mucho más cortas que las sesiones en sistemas de recuperación de bibliotecas o repositorios de documentos digitales (Park y Lee, 2013).

La tarea que se desarrolle en el momento de la visita repercute igualmente en el tiempo de permanencia en el sitio web. Un usuario con una necesidad informativa específica (como la comprobación de recepción de un correo) ocupa menos tiempo que si el usuario navega por mera curiosidad en un sitio web sobre su afición favorita (Wang y otros, 2011b). De igual forma, si el sitio web enlaza con otros sitios o páginas relacionadas, añadiendo tareas alternativas o complementarias, las sesiones pueden prolongase en buena medida (Lalmas y otros, 2014). Es el caso de la BDH, pues las acciones de visualización, lectura, copia o impresión de uno o varios de los resultados de la búsqueda se añaden al tiempo de la sesión de consulta. Cabe esperar, pues, que los usuarios cuya necesidad informativa se limite a la localización de cierta documentación tengan sesiones de duración más cortas que los usuarios que deban o deseen consultar la documentación previamente recuperada, bien por enmarcarse en tareas académicas o profesionales, ya sea por simple curiosidad.

Por este motivo caracterizamos los grupos 2 y 3 (con sesiones más prolongadas) con necesidades informativas más complejas (incluyendo la visualización de documentos), mientras que las sesiones más cortas del grupo 1 se explican bien si las necesidades informativas de sus usuarios son más sencillas o específicas, limitándose a la búsqueda y localización de documentos.

Otro gran grupo de factores que afectan a la duración de las sesiones tienen que ver con las aptitudes y actitudes del usuario que acude a un sitio web. Si un usuario se topa, por ejemplo, con dificultades en la interacción con el sistema, por falta de comprensión de la estructura y organización de la información que presenta el sitio web, la sesión será más prolongada que si el usuario conoce bien cómo proceder para resolver su necesidad. De igual forma, un usuario distraído o que está involucrado simultáneamente en diversas tareas tendrá sesiones más largas que un usuario concentrado o interesado en la información disponible en el sitio web (Huang y White, 2010).

Así pues, este grupo de factores explicaría bien las sesiones más prolongadas en el grupo 2 (caracterizándolo como poco experimentados) frente a las sesiones más cortas del grupo 3 (muy experimentados, aunque involucrados en tareas más complejas).

Como puede observarse en la Tabla VI, la longitud media de las consultas efectuadas al sistema (esto es, el número de términos que incluyen) está cerca de los dos términos por consulta, valor constante en todos los grupos de usuarios de la BDH. Estos valores confirman los datos hallados en estudios anteriores centrados en bibliotecas digitales, donde se concluye que el 80% de las consultas incluyen hasta 3 términos como máximo (Agosti y otros, 2012).

Los términos más frecuentes en dichas consultas (eliminadas las palabras vacías) muestran el carácter de los fondos predominantes de la BDH, directamente relacionados con los objetivos prioritarios de la Biblioteca Nacional, de manera que los usuarios solicitan primordialmente documentación de carácter histórico relativo a España (los términos 'historia' y 'España' destacan entre los más habituales en todos los grupos).

Las parejas de términos que en mayor número aparecen simultáneamente en las consultas (eliminadas las palabras vacías) permiten precisar mejor las necesidades informativas de los usuarios. Considerando los co-términos más frecuentes en los tres grupos, predominan claramente las consultas sobre personas, lugares geográficos y obras.

Esta tipología contrasta con la reducida utilización de la búsqueda avanzada (un tercio de consultas aproximadamente), modalidad que permi-

tiría al usuario precisar mejor su necesidad informativa. En parte este hecho puede estar motivado por una falta de adecuación entre los campos que se ofrecen al usuario en la consulta avanzada y la naturaleza de la necesidad informativa. Por ejemplo, si el usuario busca información sobre una determinada persona, no necesariamente está interesado en las obras de las que sea autor. Este hecho, unido a que el usuario raramente modifica la configuración inicial de las herramientas puestas a su disposición (Jones y otros, 2000), hace que la opción por defecto ('Todos los campos' en la BDH), sea la más empleada –incluyendo los casos en que el usuario no encuentra una opción adecuada a su necesidad concreta-.

Una recomendación, pues, que permitiría mejorar la experiencia del usuario de la BDH (Tobias y Blair, 2015) consistiría en modificar la interfaz de búsqueda sencilla de manera que, conforme el usuario vaya introduciendo los términos de la consulta, el sistema despliegue sugerencias de nombres basadas en el catálogo de autoridades de persona, geográficos y de título que posee la Biblioteca Nacional, y que está actualmente a disposición de los usuarios pero en una página web independiente (http://catalogo.bne.es/uhtbin/authoritybrowse.cgi). De esta manera no solo facilitamos la tarea de búsqueda al usuario, sino que mitigaremos las búsquedas fallidas (Moulaison y Stanley, 2013) al evitar nombres mal escritos o ambigüedades (por ejemplo, 'Manuel Navarro', consulta muy frecuente en el grupo 2, puede referirse a un político argentino, un poeta cubano o un compositor español).

De igual forma, aunque las colecciones más consultadas son distintas en cada uno de los grupos de usuarios, también es cierto que existen tres colecciones que son solicitadas por todos los grupos: Obras maestras, Carteles y las de Grabados (especialmente los de Durero y los contemporáneos). Una segunda recomendación a este respecto sería añadir, en el menú desplegable situado justo a la derecha del área de texto en la búsqueda sencilla, al menos las tres colecciones más consultadas en todos los grupos de usuarios: Obras maestras, Carteles y las de Grabados. Con ello se pone el énfasis en los aspectos más empleados y, al tiempo, optimizamos las estrategias de búsqueda a disposición de los usuarios al añadir una posibilidad más en la modalidad de consulta sencilla.

En relación al procedimiento seguido para hallar los grupos de usuarios, debe destacarse la dimensión de los datos como un factor determinante a favor del algoritmo k-means, en detrimento de otras posibilidades matemáticas como el clustering jerárquico, cuya complejidad hace que sea un procedimiento poco adecuado para grandes volúmenes de datos, como es nuestro caso (Steinbach y otros, 2000).

Otro gran grupo de procedimientos de agrupamiento se basan en modelos de distribución, entre los que destacan las distribuciones gaussianas, que se basan en la probabilidad de adscribir cada elemento a una determinada distribución gaussiana. A su vez, dentro de este tipo de distribución, sobresalen los modelos mixtura de gaussianas, que emplean algoritmos esperanza-maximización (Hastie y otros, 2009) con el objetivo de hallar iterativamente los parámetros de las distribuciones gaussianas que mejor se adapten a los datos (Bouveyron y otros, 2007).

Sometiendo el corpus de datos a este algoritmo entre 1 y 11 grupos, se aconseja formar 10 grupos con las sesiones. Este número tan elevado de grupos puede deberse al problema de sobreajuste que afecta a esta clase de procedimientos (Tu, 2005).

Los modelos de agrupamiento basados en densidad descubren grupos en áreas donde se localiza una mayor concentración o densidad de elementos. Los elementos dispersos que no pertenezcan a zonas de mayor aglomeración de observaciones se denominan ruido (Ester y otros, 1996).

Realizadas las pruebas correspondientes, pueden obtenerse entre 8 y 14 grupos, imponiendo un valor de 3 para el parámetro 'eps' y variando el parámetro minPts (con valor minPts=100 se obtienen 8 grupos y con valor minPts=50 se obtienen 14 grupos). Conforme disminuye el número de grupos, más elementos son considerados como ruido o puntos límite (para 14 grupos obtenemos 187 elementos no adscritos a ninguno de dichos grupos, hasta un máximo de 634 sesiones consideradas ruido o puntos límite para 8 grupos). A su vez, si disminuimos el valor del parámetro 'eps' (entre 0,40 y 1,5), obtenemos un número muy abultado de grupos (con eps= 0,43 se obtienen 69 grupos, mientras que con eps= 1,5 se obtienen 56 grupos). En consecuencia, este tipo de modelos no ofrece una solución aplicable a nuestro caso, debido al elevado número de grupos propuesto.

Por último, se han realizado pruebas con un modelo de clustering semejante en sus principios al algoritmo k-means denominado k-medoids. La diferencia entre ambos algoritmos estriba en que, mientras los grupos en k-means están representados por un punto central que no tiene por qué ser necesariamente un elemento del grupo, en k-medoids los grupos están representados por un elemento del grupo cuya disimilaridad media con todos los objetos o elementos del grupo es mínima (Velmurugan y Santhanam, 2010; Kaufman y Rousseeuw, 2005).

Los resultados obtenidos en las pruebas proporcionan 3 grupos con un número semejante de elementos al algoritmo k-means (un grupo 1 con 168.427 elementos; el grupo 2 posee 9.789 elementos; y el grupo 3 reúne 17.281 elementos), aunque con peores valores en los índices Calinski_Harabasz (1013477 frente a 1107739 del algoritmo k-means) y Silhouette (0,635 frente a 0,676 del algoritmo k-means).

Por último, comentar que uno de los inconvenientes que suele señalarse al analizar el algoritmo k-means es la tendencia a formar grupos de tamaño similar (Kaufman y Rousseeuw, 2005). Como hemos podido observar, en nuestro caso los grupos son muy dispares entre sí en cuanto al número de elementos, por lo que esta tendencia no ha tenido una influencia negativa en el caso analizado.

5. CONCLUSIONES

El algoritmo k-means es un procedimiento de agrupamiento muy conocido que se adecua bien al análisis de extensos ficheros log de sesiones de consulta, debido a su poca complejidad, su eficiencia en tiempos de ejecución y a su aplicabilidad a conjuntos de datos de grandes dimensiones, además de los altos valores que obtiene en los diversos índices intra-grupos e inter-grupos empleados para su validación.

El análisis cuantitativo de los ficheros log de la BDH correspondientes a un año ha permitido determinar tres grupos de usuarios en función de su comportamiento en tareas de búsqueda de información mediante consulta. Distintos valores de la duración de las sesiones, de la utilización de la consulta avanzada y del empleo de filtros permite caracterizar cada uno de esos grupos, cuyo resumen se puede consultar en las Tablas V y VI. A raíz de estos datos se proponen recomendaciones para mejorar la experiencia de búsqueda del usuario.

La relación inversa entre la utilización de herramientas para la formulación de la consulta y el nivel de conocimiento de la información que busca el usuario o la experiencia previa en el funcionamiento del buscador, relación señalada ya en estudios anteriores, permite calificar los grupos con las denominaciones 'experimentados', 'poco experimentados' y 'muy experimentados' en función de los niveles de utilización de la consulta avanzada y de la herramienta de filtración de resultados.

El grupo 2 (usuarios poco experimentados) emplea en mayor medida la ayuda que brinda la consulta avanzada, distinguiéndose al tiempo por una mayor utilización de los filtros temático y de autor. Ello se explica por unas necesidades informativas menos perfiladas y por un menor conocimiento de la BDH en general y de su sistema de recuperación en particular. El grupo 3 (usuarios muy experimentados), por el contrario, emplea menos la consulta avanzada y los filtros temático y de colecciones, debido a una concreción mayor de la documentación buscada y de un mejor conocimiento de los fondos de la BDH y de su buscador.

De cara al futuro, sería de gran utilidad poder completar estas conclusiones con información sobre el proceso de consulta llevado a cabo por los usuarios y las diversas acciones que la componen, de manera que podamos analizar en profundidad el modo de proceder de los usuarios durante las tareas de búsqueda y recuperación de información. A este respecto sería de interés simplificar la estructura de servidores que recopilan los datos log y conservarlos para mejorar el servicio mediante su análisis periódico.

De igual forma, sería conveniente completar la información meramente cuantitativa sobre los grupos hallados con un análisis posterior, de carácter cualitativo, que permita sacar a la luz características de los grupos que el tratamiento cuantitativo no desvela. Al mismo tiempo, este segundo análisis permitiría corroborar los resultados numéricos obtenidos previamente.

6. AGRADECIMIENTOS

Quisiera agradecer la colaboración de los responsables de la Biblioteca Digital Hispánica de la Biblioteca Nacional de España, sin cuya ayuda este análisis no habría sido posible. En especial quiero mostrar mi agradecimiento a Isabel Bordes Cabrera, jefa de Área de Biblioteca Digital, por su interés y permanente ayuda al facilitarme todos los datos utilizados en este estudio.

ACKNOWLEDGEMENTS

This work was carried out with the collaboration of the manager staff of the Hispanic Digital Library of the National Library of Spain, without whose help this analysis would not have been possible. In particular I want to show my gratitude to Isabel Bordes Cabrera, head of Digital Library area, for her interest and permanent help to provide me with all data used in this study.

7. REFERENCIAS

Adèr, H. J.; Mellenberg, G. J.; Hand, D. J. (2008). *Advising on research methods: a consultant's companion*. Johannes van Kessel Publishing; Huizen, the Netherlands.

Agosti, M.; Crivellari, F.; Di Nunzio, G. M. (2012). Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery*, vol. 24(3), 663-696. https://doi.org/10.1007/s10618-011-0228-8

Ahmad, P.; Brogan, M.; Johnstone, M. N. (2014). The e-book power user in academic and research libraries: Deep log analysis and user customization. *Australian Academic & Research Libraries,* vol. 45(1), 35-47. https://doi.org/10.1080/00048623.2014.885374

Amorim, R.; Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, vol. 324, 126-145. https://doi.org/10.1016/j.ins.2015.06.039

Amorim, R.; Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *Pattern Recognition*, vol. 45, 1061-1075. https://doi.org/10.1016/j.patcog.2011.08.012

Arnason, H.; Reimer, L. (2012). Analyzing public library service interactions to improve public library customer service and technology systems. *Evidence Based Library and Information Practice,* vol. 7(1), 22-40. https://doi.org/10.18438/B8NP6T

Asunka, S.; Chae, H. S.; Natriello, G. (2011). Towards an understanding of the use of an institutional repository with integrated social networking tools: A case study of PocketKnowledge. *Library & Information Science Research,* vol. 33(1): 80-88. https://doi.org/10.1016/j.lisr.2010.04.006

Benevenuto, F.; Rodrigues, T.; Cha, M.; Almeida, V. (2009). Characterizing user behavior in online social networks. *Proceedings of the 9th ACM SIG-COMM Conference on Internet Measurement Conference*, pp. 49-62. ACM; New York. https://doi.org/10.1145/1644893.1644900

Berndt-Morris, E.; Minnis, S. M. (2014). The chat is coming from inside the house: An analysis of perceived chat behavior and reality. *Journal of Library & Information Services in Distance Learning,* vol. 8(3-4), 168-180. https://doi.org/10.1080/1533290X.2014.945833

Borra, E.; Weber, I. (2012). Political insights: Exploring partisanship in web search queries. *First Monday*, vol. 17(7). https://doi.org/10.5210/fm.v17i7.4070

Borrego, A.; Fry, J. (2012). Measuring researchers' use of scholarly information through social bookmarking data: A case study of BibSonomy. *Journal of Information Science*, vol. 38(3), 297-308. https://doi.org/10.1177/0165551512438353

Bouveyron, C.; Girard, S.; Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics and Data Analysis,* vol. 52(1), 502-519. https://doi.org/10.1016/j.csda.2007.02.009

Brett, K.; German, E.; Young, F. (2015). Tabs and tabulations: Results of a transaction log analysis of a tabbed-search interface. *Journal of Web Librarianship,* vol. 9(1), 22-41. https://doi.org/10.1080/19322909.2015.1004502

Calinski, T.; Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics,* vol. 3(1), 1-27. https://doi.org/10.1080/03610927408827101

Celebi, M. E.; Kingravi, H. A.; Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, vol. 40(1), 200-210. https://doi.org/10.1016/j.eswa.2012.07.021

Chapman, J. L. (1981). A state-transition analysis of online information seeking behavior. *Journal of the American Society for Information Science,* vol. 32(5), 325-333. https://doi.org/10.1002/asi.4630320503

Chen, C. C.; Tsai, Y. (2012). A novel business cycle surveillance system using the query logs of search engines. *Knowledge-Based Systems,* vol. 30, 104-114. https://doi.org/10.1016/j.knosys.2011.12.012

Clifton, B. (2012). *Advanced web metrics with Google Analytics*. John Wiley & Sons; Indianapolis, Indiana.

Dempsey, M.; Valenti, A. M. (2016). Student use of keywords and limiters in web-scale discovery searching. *Journal of Academic Librarianship,* vol. 42(3), 200-206. https://doi.org/10.1016/j.acalib.2016.03.002

Dick, S.; Yazdanbaksh, O.; Tang, X.; Huynh, T.; Miller, J. (2014). An empirical investigation of web session workloads: Can self-similarity be explained by deterministic chaos? *Information Processing and Management,* vol. 50(1), 41-53.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226-231. AAAI Press; Menlo Park, California.

Ferl, T. E.; Millsap, L. (1996). The knuckle-cracker's dilemma: a transaction log study of OPAC subject searching. *Information Technology and Libraries,* vol. 15(2), 81-98.

Fisher, K. E.; Erdelez, S.; McKechnie, L. (editores) (2005). *Theories of information behavior*. Information Today; Medford, NJ,EE.UU.

González-Teruel, A.; Barrios Cerrejón, M. (2012). *Métodos y técnicas para la investigación del comportamiento informacional: fundamentos y nuevos desarrollos*. Editorial Trea; Gijón.

Guerbas, A.; Addam, O.; Zaarour, O.; Nagi, M.; Elhajj, A.; Ridley, M.; Alhajj, R. (2013). Effective web log mining and online navigational pattern prediction. *Knowledge-Based Systems,* vol. 49, 50-62. https://doi.org/10.1016/j.knosys.2013.04.014

Gul, S.; Nabi, S.; Mushtaq, S.; Shah, T. A.; Ahmad, S. (2013). Political unrest and educational electronic resource usage in a conflict zone, Kashmir (indian administered Kashmir): Log analysis as politico analytical tool. *Information World,* vol. 14(2): 388-399.

Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems,* vol. 17, 107-145. https://doi.org/10.1023/A:1012801612483

Hancock-Beaulieu, M. (1989). Online catalogues: a case for the user. En: Hildreth, C. R. (editor) *The online catalogue: developments and directions*. The Library Association; London.

Hastie, T.; Tibshirani, R.; Friedman, J. (2009). The EM algorithm. En: Hastie, T; Tibshirani, R.; Friedman, J. (autores) *The elements of statistical learning: data mining, inference, and prediction* . Springer; New York.

Hershkovitz, A.; Hardof-Jaffe, S.; Nachmias, R. (2014). Content consumption and hierarchical structures of web-supported courses. *Journal of Interactive Learning Research,* vol. 25(3), 353-371.

Huang, J.; White, R. W. (2010). Parallel browsing behavior on the web. *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT '10)*, pp. 13-18. ACM; New York. https://doi.org/10.1145/1810617.1810622

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery,* vol. 2(3), 283-304. https://doi.org/10.1023/A:1009769707641

Hunt, S.; Cimino, J. J.; Koziol, D. E. (2013). A comparison of clinicians' access to online knowledge resources using two types of information retrieval applications in an academic hospital setting. *Journal of the Medical Library Association,* vol. 101(1), 26-31. https://doi.org/10.3163/1536-5050.101.1.005

Iyer, L. S.; Raman, R. M. (2011). Intelligent analytics: Integrating business intelligence and web analytics. *International Journal of Business Intelligence Research,* vol. 2(1), 31-45. https://doi.org/10.4018/jbir.2011010103

Jansen, B. J. (2006). Search log analysis: what it is, what's been done, how to do it. *Library & Information Science Research,* vol. 28, 407-432. https://doi.org/10.1016/j.lisr.2006.06.005

Jansen, B. J.; Pooch, U. (2001). A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology,* vol. 52(3), 235-246. https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1607>3.0.CO;2-F

Jansen, B. J.; Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management,* vol. 42(1), 248-263. https://doi.org/10.1016/j.ipm.2004.10.007

Jones, S.; Cunningham, S. J.; McNab, R.; Boddie, S. (2000). A transaction log analysis of a digital library. *International Journal on Digital Libraries*, vol. 3(2), 152-169. https://doi.org/10.1007/s007999900022

Kahlon, M.; Yuan, L.; Daigre, J.; Meeks, E.; Nelson, K.; Piontkowski, C.; Reuter, K.; Sak, R.; Turner, B.; Weber, G. M.; Chatterjee, A. (2014). The use and significance of a research networking system. *Journal of Medical Internet Research,* vol. 16(2). https://doi.org/10.2196/jmir.3137

Kapoor, K. (2010). Print and electronic resources: Usage statistics at Guru Gobind Singh Indraprastha University library. *Program: Electronic Library and Information Systems,* vol. 44(1), 59-68. https://doi.org/10.1108/00330331011019690

Kaufman, L.; Rousseeuw, P. J. (2005). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons; Hoboken, New Jersey.

Lalmas, M.; O'Brien, H.; Yom-Tov, E. (2014). Measuring User Engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 6(4), 1-132. https://doi.org/10.2200/S00605ED1V01Y201410ICR038

Lambert, F. (2013). Seeking electronic information from government resources: A comparative analysis of two communities' web searching of municipal government websites. *Government Information Quarterly,* vol. 30(1), 99-109. https://doi.org/10.1016/j.giq.2012.07.007

Larson, R. R. (1991). Classification clustering, probabilistic information retrieval, and the online catalog. *The Library Quarterly,* vol. 61(2), 133-173. https://doi.org/10.1086/602331

Lai, Y.; Zeng, J. (2013). A cross-language personalized recommendation model in digital libraries. *The Electronic Library,* vol. 31(3), 264-277. https://doi.org/10.1108/EL-08-2011-0126

Leeder, C.; Lonn, S. (2014). Faculty usage of library tools in a learning management system. *College & Research Libraries*, vol. 75(5), 641-663. https://doi.org/10.5860/crl.75.5.641

Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. (2010). Understanding of internal clustering validation measures. *Proceedings of the 10th IEEE International Conference on Data Mining***,** pp. 911-916. IEEE Computer Society; Los Alamitos, California.

Ma, H. (2013). Tech services on the web: Google Analytics. *Technical Services Quarterly,* vol. 30(1), 119-200. https://doi.org/10.1080/07317131.2013.735978

Maabreh, M. A.; Al-Kabi, M.; Alsmadi, I. M. (2012). Query classification and study of university students' search trends. *Program: Electronic Library and Information Systems,* vol. 46(2), 220-241. https://doi.org/10.1108/00330331211221855

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations*. Proceedings of the 5th Berkeley Symposium on Mathematical*

Statistics and Probability, pp. 281-297. University of California Press.

Mahoui, M.; Jo Cunningham, S. (2000). A Comparative Transaction Log Analysis of Two Computing Collections. *Research and Advanced Technology for Digital Libraries: Proceedings of the 4th European Conference, ECDL* 2000, pp. 418-423. Springer; Berlin, Heidelberg. https://doi.org/10.1007/3-540-45268-0_53

Malliari, A.; Moreleli-Cacouris, M.; Kapsalis, K. (2010). Usage patterns in a greek academic library catalogue: A follow-up study. *Performance Measurement and Metrics*, vol. 11(1), 47-55. https://doi.org/10.1108/14678041011026865

Markey, K. (2007). Twenty-five years of end-user searching, Part 2: Future research directions. *Journal of the American Society for Information Science and Technology,* vol. 58(8), 1123-1130. https://doi.org/10.1002/asi.20601

Maulik, U.; Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Inteligence,* vol. 24(12), 1650-1654. https://doi.org/10.1109/TPAMI.2002.1114856

Mbabu, L. G.; Bertram, A.; Varnum, K. (2013). Patterns of undergraduates' use of scholarly databases in a large research university. *The Journal of Academic Librarianship,* vol. 39(2), 189-193. https://doi.org/10.1016/j.acalib.2012.10.004

Moulaison, H. L.; Stanley, S. N. (2013). Beyond failure: Potentially mitigating failed author searches in the online library catalog through the use of linked data. *Journal of Web Librarianship,* vol. 7(1), 37-57. https://doi.org/10.1080/19322909.2013.738562

Munson, D. M.; Otto, J. L. (2013). Have link resolvers helped or hurt? The relationship between ILL and OpenURL at a non-SFX library. *OCLC Systems & Services: International Digital Library Perspectives,* vol. 29(2), 78-86. https://doi.org/10.1108/10650751311319287

Ortega Priego, J. L. (2004). Análisis del consumo de información de una revista electrónica: análisis de ficheros log de Cybermetrics. *Revista Española de Documentación Científica,* vol. 27(4), 455-468.

Ortega Priego, J. L. (2005). Análisis de sesiones de la web del CINDOC: una aproximación a la minería de uso web. *El Profesional de la Información,* vol. 14(3), 190-198. http://www.elprofesionaldelainformacion.com/contenidos/2005/mayo/4.pdf

Ozen, Z.; Bakiolu, F.; Beden, S. (2014). The examination of user habits through the Google Analytic data of academic education platforms. *International Journal of E-Adoption,* vol. 6(2), 31-45. https://doi.org/10.4018/ijea.2014070103

Park, M.; Lee, T. S. (2016). A longitudinal study of information needs and search behavior in science and technology: a query analysis. *The Electronic Library,* vol. 34(1), 83-98. https://doi.org/10.1108/EL-04-2014-0058

Park, M.; Lee, T. S. (2013). Understanding science and technology information users through transaction log analysis. *Library Hi Tech*, vol. 31(1), 123-140. https://doi.org/10.1108/07378831311303976

Peeples, M. A. (2011). R script for k-means cluster analysis. http://www.mattpeeples.net/kmeans.html [Consulta: 20/08/2016]

Peters, T. A. (1993). The history and development of transaction log analysis. *Library Hi Tech,* vol. 11(2), 41-66. https://doi.org/10.1108/eb047884

Priya, R. V.; Vadivel, A. (2012). User behaviour pattern mining from weblog. *International Journal of Data Warehousing and Mining,* vol. 8(2), 1-22. https://doi.org/10.4018/jdwm.2012040101

Rechavi, A.; Rafaeli, S. (2014). Active players in a network tell the story: Parsimony in modeling huge networks. *First Monday,* vol. 19(8). https://doi.org/10.5210/fm.v19i8.5217

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* vol. 20, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Rozaklis, L.; MacDonald, C. M. (2011). A typology of collaborative communication in a digital reference environment. *Reference Librarian,* vol. 52(4), 308-319. https://doi.org/10.1080/02763877.2011.586907

Schalkoff, R. J. (2001). Pattern Recognition. En: *Wiley Encyclopedia of Electrical and Electronics Engineering*. John Wiley & Sons, Inc.; Indianapolis, Indiana.

Shieh, J. (2012). From website log to findability. *The Electronic Library,* vol. 30(5), 707-720. https://doi.org/10.1108/02640471211275747

Shiri, A. (2011). Revealing interdisciplinarity in nanoscience and technology queries: A transaction log analysis approach. *Knowledge Organization,* vol. 38(2), 135-153.

Spink, A.; Jansen, B. J. (2004). *Web search: Public searching of the Web*. Kluwer; New York.

Spiteri, L. F.; Tarulli, L. (2012). Social discovery systems in public libraries: If we build them, will they come? *Library Trends*, vol. 61(1), 132-147. https://doi.org/10.1353/lib.2012.0019

Steinbach, M.; Karypis, G.; Kumar, V. (2000). A comparison of document clustering techniques*. KDD-2000 workshop on text mining*, pp. 525-526. Boston.

Strohmaier, M.; Kroll, M. (2012). Acquiring knowledge about human goals from search query logs. *Information Processing and Management*, vol. 48(1), 63-82. https://doi.org/10.1016/j.ipm.2011.03.010

Stuit, M.; Wortmann, H. (2012). Discovery and analysis of e-mail-driven business processes. *Information Systems,* vol. 37(2): 142-168. https://doi.org/10.1016/j.is.2011.09.008

Tobias, C.; Blair, A. (2015). Listen to what you cannot hear, observe what you cannot see: An introduction to evidence-based methods for evaluating and enhancing the user experience in distance library services. *Journal of Library & Information Service in Distance Learning,* vol. 9(1-2), 148-156. https://doi.org/10.1080/1533290X.2014.946354

Tu, Z. (2005). Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. *Tenth IEEE International Conference on Computer Vision*, vol. 2, pp. 1589-1596. IEEE; Beijing.

Van Gemert-Pijnen, J.; Kelders, S. M.; Bohlmeijer, E. T. (2014). Understanding the usage of content in a mental health intervention for depression: An analysis of log data. *Journal of Medical Internet Research,* vol. 16(1). https://doi.org/10.2196/jmir.2991

Verma, M.; Srivastava, M.; Chack, N.; Kumar, A.; Gupta, N. (2012). A comparative study of various clustering algorithms in data mining. *International Journal of Engineering Research and Applications*, vol. 2(3), 1379-1384.

Velmurugan, T.; Santhanam, T. (2010). Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of Computer Science*, vol. 6(3), 363-368. https://doi.org/10.3844/jcssp.2010.363.368

Villén-Rueda, L.; Senso, J. A.; Moya-Anegón, F. de (2007). The use of OPAC in a large academic library: a transactional log analysis study of subject searching. *The Journal of Academic Librarianship,* vol. 33(3), 327-337. https://doi.org/10.1016/j.acalib.2007.01.018

Vogt, W. P. (editor) (2011). *Sage quantitative research methods*. SAGE; Los Angeles. https://doi.org/10.4135/9780857028228

Waller, V. (2010). Accessing the collection of a large public library: An Analysis of OPAC use. *LIBRES: Library and Information Science Research Electronic Journal,* vol. 20(1).

Wang, C.; Ke, H.; Lu, W. (2012). Design and performance evaluation of mobile web services in libraries: A case study of the Oriental Institute of Technology library. *The Electronic Library,* vol. 30(1), 33-50. https://doi.org/10.1108/02640471211204051

Wang, J.; Huffaker, D. A.; Treem, J. W.; Fullerton, L.; Ahmad, M. A.; Williams, D.; Poole, M. S.; Contractor, N. (2011a). Focused on the prize: Characteristics of experts in massive multiplayer online games. *First Monday,* vol. 16(8). https://doi.org/10.5210/fm.v16i8.3672

Wang, P.; Berry, M. W.; Yang, Y. (2003). Mining longitudinal web queries: trends and patterns. *Journal of the American Society for Information Science and Technology,* vol. 54(8), 743-758. https://doi.org/10.1002/asi.10262

Wang, S.; Zhang, J.; Yang, F.; Ye, J. (2014). Research on cluster analysis method of E-government public hotspot information based on web log analysis. *CIT – Journal of Computing and Information Technology,* vol. 22, 11-19. https://doi.org/10.2498/cit.1002281

Wang, X.; Shen, D.; Chen, H.; Wedman, L. (2011b). Applying web analytics in a K-12 resource inventory. *The Electronic Library*, vol. 29(1), 20-35. https://doi.org/10.1108/02640471111111415

Yom-Tov, E.; White, R. W.; Horvitz, E. (2014). Seeking insights about cycling mood disorders via anonymized search logs. *Journal of Medical Internet Research,* vol. 16(2). https://doi.org/10.2196/jmir.2664

Zhang, J.; An, L. (2010). Visual component plane analysis for the medical based on transaction log. *The Canadian Journal of Information and Library Science,* vol. 34(1), 83-111. https://doi.org/10.1353/ils.0.0006

Zhang, J.; Zhao, Y. (2013). A user term visualization analysis based on a social question and answer log. *Information Processing and Management,* vol. 49(5), 1019-1048. https://doi.org/10.1016/j.ipm.2013.04.003

Zhu, D.; Guralnik, D.; Wang, X.; Li, X.; Moran, B. (2015). Statistical estimation for Single Linkage Hierarchical Clustering. *Proceedings of the IEEE 5th Annual International Conference on Cyber Technology in Automation, Control and Intelligent Systems (CYBER 2015)*, pp. 745-750. IEEE Computer Society; Los Alamitos, California. https://doi.org/10.1109/CYBER.2015.7288035