2da. Jornada sobre la Biblioteca Digital Universitaria JBD 2004
"Oportunidades y Desafíos: Formación de Usuarios, Metadatos y Arquitectura de la Web"
17 y 18 de junio de 2004. Universidad de Buenos Aires. Argentina

Integrando la información digital y no digital a través de una interfaz de consulta única potenciada con acceso temático jerárquico generado por *clustering* y con exploración entre registros

Cristian Merlino Santesteban
Centro de Documentación. Facultad de Cs. Económicas y Sociales
Universidad Nacional de Mar del Plata. Argentina
csantest@mdp.edu.ar

Resumen: Se plantea la necesidad de integrar la información digital y no digital a través de una interfaz de consulta única que reduzca la carga cognitiva del usuario al momento de interactuar con ella, y permita visualizar y explorar el contenido de la colección mediante el uso de una estructura temática jerárquica generada por *clustering,* y de la red semántica generada por las conexiones entre registros.

Palabras clave: interfaz de usuario; exploración; jerarquía temática; *clustering*

1. Introducción.

En los últimos años, la producción de información en formato digital ha experimentado un crecimiento acelerado nunca antes visto [Lyman y Varian, 2000]. Este incremento, favorecido por la evolución de Internet -el canal de información binaria por excelencia-, propició que las organizaciones documentales proveyeran acceso a una vasta cantidad de fuentes de información digitales, usualmente vía URLs (*Uniform Resources Locator*).

La tendencia general para dar acceso a la información digital, por alguna razón sin justificación cierta, fue implementar mecanismos de consulta (listas estáticas, bases de datos dinámicas) no complementarios a los sistemas de recuperación tradicionales, los catálogos en línea. De este modo, el acceso al fondo documental quedó segmentado por su formato físico, y el usuario se vio forzado a interactuar con interfaces disímiles y a incrementar esfuerzos en la difícil tarea de encontrar información relevante.

El presente trabajo plantea la necesidad de brindar una interfaz de consulta única que minimice la carga cognitiva del usuario en la instancia de formular la búsqueda y permita recuperar la información apropiada centrándose en su cualidad principal, el contenido.

2. Integración de las fuentes de información. ¿Importa el formato?.

Cuando se establece una distinción entre dos o más elementos, objetos o cosas, se presupone que se realiza para destacar de éstos una característica diferenciadora por encima del cualquier otra. En el ámbito de las organizaciones documentales, al encontrar conjuntamente guías de recursos digitales o las llamadas "bibliotecas digitales/electrónicas/virtuales"[1] y catálogos bibliográficos en línea, puede apreciarse que se ha realizado una diferenciación de los materiales de la colección de acuerdo a su registro físico: lo digital y lo analógico pero, ¿es el formato de la información esa característica?, ¿qué "valor" puede

[1] Para obtener una definición o interpretación adecuada de los conceptos biblioteca digital, biblioteca electrónica y biblioteca virtual véase el trabajo de Tramullas Saz [2002] y de Sánchez Díaz y Vega Valdés [2002].

tener el formato para hacer esa diferencia?, ¿qué "beneficio" real obtiene el usuario con la separación de los recursos?. Principalmente, desde el punto de vista del usuario, eje central de todo servicio, la distinción planteada carece de sentido puesto que lo que él valora no es el tipo de formato sino la accesibilidad al *corpus* documental y la calidad de las fuentes (contenido y valor), ya que uno de los grandes problemas que debe enfrentar cotidianamente en la mayoría de las unidades informativas en línea u *off-line* es la desintegración de los recursos documentales. Por un lado, el acervo bibliográfico no digital brindado a través de un OPAC (*Online Public Access Catalog*) y por el otro, los recursos digitales ofrecidos como listados estáticos, combinados o no, de artículos, libros y revistas, entre otras fuentes, o por medio de bases de datos referenciales (por ejemplo, de URLs) o de texto completo. La separación de las fuentes de información potencialmente útiles y el empleo de herramientas de consulta poco amigables sólo sirve para maximizar, aún más, la carga cognitiva del usuario. No hay que olvidar que el servicio informativo que se ofrece al usuario final tiene que ayudarlo a satisfacer su requerimiento de información de la mejor manera posible y no a desorientarlo en su tarea.

Al integrar, desde la perspectiva del acceso y del usuario, las fuentes digitales y no digitales se consigue:

- Proveer un acceso único a la información independientemente de la forma física que tome.
- Maximizar la consulta potencial del fondo documental.
- Brindar un acceso ágil, rápido y conveniente.
- Prevenir la desorientación del usuario.
- Evitar el incremento de esfuerzos tanto del usuario (acceso y utilización) como del proveedor del acceso (desarrollo y mantenimiento).
- Centrar todas las mejoras en una sola interfaz de búsqueda integral.
- Acceder simultáneamente a toda la colección (interna y externa).
- Concentrar todas las consultas a través de un medio unificado.

3. Interfaz de consulta mejorada con acceso temático jerárquico y exploración entre registros

La interfaz de usuario es el canal de comunicación entre los hombres y las computadoras[2], es decir, el medio por el cual los hombres y las computadoras interactúan de manera precisa y concreta [Marcos, 2001]. La interacción del usuario con este medio es un paso breve pero decisivo, ya que si se encuentra con muchas dificultades (físicas y/o mentales) quedará frustrado o abandonará la tarea. Por ende, el diseño conveniente de la interfaz es un elemento clave para que las organizaciones documentales puedan ofrecer acceso efectivo a sus colecciones a través de redes de telecomunicaciones.

La implementación de una interfaz de consulta única pretende lograr que el usuario interactúe con un sólo intermediario artificial del sistema de recuperación de información con la menor carga cognitiva posible, o sea, con el menor esfuerzo mental necesario, tanto al momento de plantear su necesidad de información como al momento de interpretar los resultados obtenidos[3]. Para ello se propone potenciar su usabilidad complementando, para

[2] La disciplina que explora las teorías que explican las interacciones entre los hombres y las computadoras y las interfaces que apoyan estas interacciones se denomina Interacción Hombre-Computadora (*Human Computer Interaction*).

[3] Según Hearst [1999], la interfaz de consulta (ideal) debería ayudar al usuario en la clarificación, compresión y definición de sus necesidades de información. También debería auxiliarlo en formular sus búsquedas, escoger entre diferentes fuentes de información disponibles, interpretar los resultados y mantener un historial del proceso de búsqueda.

la enunciación de la búsqueda, el tradicional sistema de consulta por interrogación (*querying*) con técnicas de exploración (*browsing*), específicamente la exploración temática arbórea (a macronivel) y la exploración entre registros (a micronivel). Y proveyendo, para la interpretación de los resultados recuperados, los elementos del *cluster* activo y su relación semántica con *clusters* vecinos.

3.1. Por qué la exploración

En la consulta por interrogación, se obliga al usuario a formular la búsqueda utilizando los atributos (campos) con que se caracterizan documentos o los términos que piensa que se han podido utilizar para describir documentos que respondan a su demanda de información[4]. Estos atributos o términos pueden ser unitérminos o expresiones, combinables en ambos casos por medio de operadores lógicos (AND, OR y NOT) o sintácticos (adyacencia y distancia).

En la consulta por exploración, se adapta el sistema a un usuario que tiene una idea imprecisa o no puede expresar con precisión el objeto de la búsqueda. Lin [1997, p. 41] la define como "*un proceso interactivo en el que uno puede visualizar grandes cantidades de información, percibir o encontrar estructuras o relaciones, y seleccionar ítems centrando su atención visual en ellos*". Por su parte, Oddy y Balakrishnan [1991] sostienen que la exploración está íntimamente relacionada con el reconocimiento visual y el razonamiento espacial, de forma opuesta a la especificación lingüística y razonamiento lógico de la búsqueda por interrogación. Asimismo, una estructura de exploración reduce el desbordamiento cognoscitivo del usuario por requerirle a éste solamente la detección de conceptos relevantes, evitándole así plantear la consulta con palabras propias que posiblemente no aparezcan expresadas como tales en los campos de descripción bibliográfica.

Vía la exploración temática jerárquica, el usuario ve el contenido de la colección bibliográfica de forma que va percibiendo el conocimiento organizado y seleccionando gradualmente los temas de lo general a lo específico (controla el camino y la profundidad). A su vez, con la exploración entre registros puede navegar una red semántica usando las numerosas conexiones establecidas entre los atributos (autor, revista, tema, etc,) de los documentos.

Entre los beneficios de la exploración a macronivel y micronivel se pueden mencionar:

► son tipos de exploración conocidas por los usuarios;
El auge de la World Wide Web propició el uso masivo de directorios web para buscar información e hiperenlaces para navegar la red interconectada de nodos.

► son fáciles de usar por usuarios no expertos;
Los usuarios son guiados por la categorización predefinida o los atributos activos. El reconocimiento visual les evita conocer las instrucciones de búsqueda.

► son útiles como punto de inicio de una búsqueda;
Cuando no hay una idea clara o definición precisa del objeto de la búsqueda, moverse fácilmente entre términos o conceptos permite al usuario reconocer, delimitar o redefinir su necesidad de información.

► son independientes de la lengua de los documentos;
La exploración, al no requerir la definición de una expresión de búsqueda, recupera simultáneamente ítems interrelacionados en cualquier lengua.

[4] El acceso a los sistemas de recuperación puede ser complejo si los usuarios no tienen un conocimiento amplio sobre sus contenidos.

► son ajenas a cualquier error tipográfico u ortográfico.
Al evitar ingresar una enunciación de búsqueda en una casilla de interrogación no existe posibilidad alguna de incurrir en errores tipográficos u ortográficos.

3.2. *Clustering* temático de documentos

El *clustering* de documentos es la operación de agrupar documentos similares o relacionados entre sí en clases comunes. Las agrupaciones generadas, al basarse enteramente en las propiedades internas de la colección, pueden revelar la estructura intrínseca de ésta[5] (por ejemplo, temas y subtemas).

El análisis de clases, en recuperación de información, se basa en la "hipótesis *cluster*" [van Rijsbergen, 1979], documentos fuertemente asociados entre sí tienden a satisfacer las mismas consultas. De acuerdo a esto, aquellos documentos que poseen más temas comunes son potencial respuesta de las mismas necesidades informativas, por lo que si estuvieran agrupados antes de su recuperación se facilitarían las búsquedas [Moya Anegón, 1995].

Cualquier sistema de generación de *cluster* en un catálogo bibliográfico o herramienta similar, debería basarse en la información clasificatoria que se incluye en las referencias bibliográficas (código de clasificación, tema ...), puesto que el *clustering* potencia a esta información en proporcionar agrupaciones útiles de documentos que contienen un alto grado de similitud en su contenido temático que en determinar y exhibir las relaciones fundamentales o taxonomías de conceptos [Fernández Molina y Moya Anegón, 1998].

La información temática enriquecida con palabras del título [Larson, 1992] sería el patrón para crear la estructura de *cluster* (clasificación semiautomática) porque [Moya Anegón, 1995]:

1) La información es muy precisa cuando es general y no tanto cuando es específica.
2) La información es suficiente para organizar la estructura de *cluster* que se quiera. Por ejemplo, con la Clasificación Decimal Universal (CDU) se pueden crear estructuras de *clusters* jerárquicos. Basta con comparar cada notación con los valores de una tabla definida de antemano para que el sistema sepa el *cluster* en el que deben ser incluidas las materias asociadas a esa notación. La sencillez del proceso garantiza que la respuesta sea rápida.
3) La existencia de una o varias notaciones asociadas a cada registro hace que la estructura de datos resultante sea muy flexible desde el punto de vista lógico, dado que las materias podrán formar parte de varios *cluster*s siendo recuperadas desde todos ellos (el usuario puede llegar así al mismo documento por distintos caminos sin ser consciente de ello).
4) Al estar el criterio de generación de *cluster* determinado por el sistema de clasificación, el propio bibliotecario puede establecer la forma concreta de distribución de las materias confeccionando la tabla de comienzos de notaciones correspondiente.
5) Se relaciona la información de las materias y de la clasificación lo que supone una enorme cantidad de información.

Así generado, el procesamiento de *clustering* temático permite que cada documento se conecte de manera explícita a otros documentos relacionados del sistema y que éste se presente no sólo en una estructura de carácter jerárquico sino en una red interconectada de nodos informativos, donde cada acceso a una subcategoría crea un *cluster* nuevo vinculado por similitud semántica con *clusters* vecinos.

[5] Idealmente, los grupos serán completamente separados y distanciados tan lejos como sea posible. Pero algunas veces, el solapamiento de *clusters* es inevitable [van Rijsbergen, 1979].

4. Consideraciones finales.

A modo de conclusión, se quiere señalar que cuando se está diseñando una interfaz de consulta no sólo se está dando acceso a una herramienta de consulta subyacente sino también se está creando/imponiento un mecanismo intermediario entre el usuario y la colección. La presentación y accesibilidad de la información digital y no digital en pantalla son tan importantes como el propio contenido de los datos. Si la interfaz no es lo suficientemente amigable[6] o no se adapta al comportamiento del usuario todo lo que esté del otro lado, el sistema de búsqueda en sí mismo y el fondo bibliográfico, pierde utilidad. La interfaz debe tratar de adaptarse al usuario y no a la inversa. Por eso, centrar su diseño en casillas de búsqueda y/o listas de enlaces de manera arbitraria o estética es un error que conducirá a un mal uso del sistema y un usuario insatisfecho.

En vez de limitar al usuario a un único tipo de consulta, interrogación o exploración, se recomienda diseñar interfaces que permitan ambas. Los usuarios podrán navegar la estructura arborescente hasta el nivel que deseen, realizar consultas por interrogación de toda la colección o sólo la porción de la jerarquía que se ha recuperado y navegar entre registros interrelacionados por sus atributos. Finalmente, la aplicación del análisis de *clustering* temático en la generación de una estructura arbórea permite flexibilizar su rigidez y fortalecer las agrupaciones semánticas entre los documentos.

Referencias bibliográficas

Hearst, M. A. [1999]. User interfaces and visualization. *In* Baeza-Yates, R y Ribeiro-Neto, B. *Modern information retrieval*. pp. 257-323. Essex, England: Addison Wesley.

Fernández Molina, J. C. y Moya Anegón, F. [1998]. *Los catálogos de acceso público en línea: el futuro de la recuperación de información bibliográfica*. Málaga: AAB.

Lin, X. [1997]. Map displays for information retrieval. *Journal of the American Society for Information Science*, 48(1): 40-54.

Lyman, P. y Varian, H. R. [2000]. *How much information*. Disponible en: <http://www.sims.berkeley.edu/how-much-info>. Acceso: 28 diciembre 2003.

Larson, R. R. [1992]. Experiments in automatic Library Congress Classification. *Journal of the American Society for Information Science*, 43(2): 130-148.

Marcos, M. C. [2001]. HCI (human computer interaction): concepto y desarrollo. *El Profesional de la Información,* 10(6): 4-16.
Moya Anegón, F. [1995]. *Sistemas integrados de gestión bibliotecaria*. Madrid: ANABAD.

Oddy, R. N. y Balakrishnan, B. [1991]. PTHOMAS: an adaptive information retrieval system on the connection machine. *Information Processing and Management*, 27(4):317-335.

Sánchez Díaz, M. y Vega Valdés, J. C. [2002]. Bibliotecas electrónicas, digitales y virtuales: tres entidades por definir. *Acimed,* 10(6). Disponible en: <http://www.infomed.sld.cu/revistas/aci/vol10_6_02/aci05602.htm>. Acceso: 6 enero 2004.

Tramullas Saz, J. [2002]. Propuestas de concepto y definición de la biblioteca digital. Disponible en: <http://mariachi.dsic.upv.es/jbidi/jbidi2002/Camera-ready/Sesion1/S1-1.pdf>. Acceso: 8 enero 2004.

[6] La amigabilidad se refiere a su facilidad de uso. Una interfaz es tanto más amigable cuanto más fácil de usar resulta para una mayor proporción de usuarios de una población dada.

van Rijsbergen, C. J. [1979]. *Information retrieval*. London: Butterworths. Disponible en: <http://www.dcs.gla.ac.uk/~keith/>. Acceso: 3 agosto 2000.