JBDU 2008: 6º Jornada sobre la Biblioteca Digital Universitaria

Representación semántica de un catálogo de tesis por medio de una interfaz de visualización gráfica basada en la metodología Topic Maps

Gustavo Liberatore
Leticia Lizondo
gliberat@mdp.edu.ar
leticia_lizondo@yahoo.com.ar
Departamento de Documentación.
Universidad Nacional de Mar del Plata

Resumen

Se presentan los resultados preliminares del desarrollo de una interface de visualización gráfica aplicada a un catálogo de tesis académicas del área de la nutrición. El objetivo de esta investigación se basa en la generación de una aplicación que permita la recuperación de este tipo de documentos a partir de las temáticas de investigación y las relaciones existentes entre ellas brindando al usuario información valiosa sobre la actividad académica realizada en un campo disciplinar. Para generar este producto se utilizó una metodología mixta, empleando técnicas de análisis de co-ocurrencia de términos, representación por medio de redes sociales y generación de interfaces semánticas por medio de Topic maps.

Palabras clave: Visualización de información – Interfaz gráfica – Topic Map – Nutrición

Abstract

Presents the preliminary results of the development of a graphical interface for viewing applied to a catalog of academic theses in the field of nutrition. The objective of this research is based on the generation of an application that enables the retrieval of such documents from the subjects of research, and relations between them providing the user with valuable information on the academic discipline in a field. To generate this product was used a mixed methodology, using analytical techniques of co-occurrence of terms, represented by social networks and semantic generation of interfaces by Topic maps.

Keywords: Information display - Graphical Interface - Topic Map - Nutrition

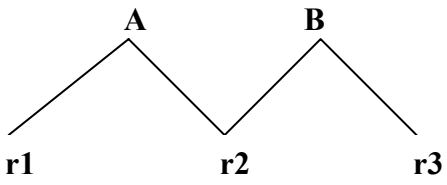
Introducción

La representación gráfica de la información para su posterior visualización es una actividad común en la mayoría de las disciplinas científicas en los últimos tiempos (Klovdhal, A. S., 1981), (Crosby, A. W., 1997). Pero el uso de las representaciones gráficas, en combinación con la tecnología informática, para conseguir una adecuada visualización de la información, es una tarea relativamente nueva, que se ha convertido en uno de los principales objetos de estudio de los últimos años (Vargas-Quesada, 2005).

La visualización de la información no es el resultado implícito del acto de ver, ni tampoco es un producto espontáneo del individuo que recibe la información ya visualizada. La visualización de la información es una tarea del comunicador visual, que transforma datos abstractos y fenómenos complejos de la realidad en mensajes visibles, haciendo posible que los individuos vean con sus propios ojos, datos y fenómenos que son directamente inaprensibles, y que por tanto comprendan la información que yace oculta (Costa, J., 1998).

En términos de recuperación de información el uso de interfaces gráficas para la representación de bases de datos bibliográficas se ha ido acentuando como fruto de líneas de investigación iniciadas hace varios años atrás. Uno de los principales objetivos es lograr un modelo de representación que supere al tradicional bibliográfico referencial que, aunque útil, posee serias limitaciones al momento de mostrar al potencial usuario la verdadera estructura y dimensión del campo de conocimiento al cual se está enfrentando.

Detrás de estás interfaces se encuentran las diferentes metodologías que permiten niveles de representación más complejos. En este sentido, el método bibliométrico ha aportado tempranamente una de las fórmulas más eficaces para los modelos de representación de información a través de los análisis de co-citación (Small, 1973). En un sentido general, la co-citación es una relación de co-ocurrencia que se manifiesta cuando dos ítems de la literatura existente son citados juntos por un tercero (Miguel et. al., 2007). Gráficamente, una relación de co-citación puede presentarse de la siguiente manera:



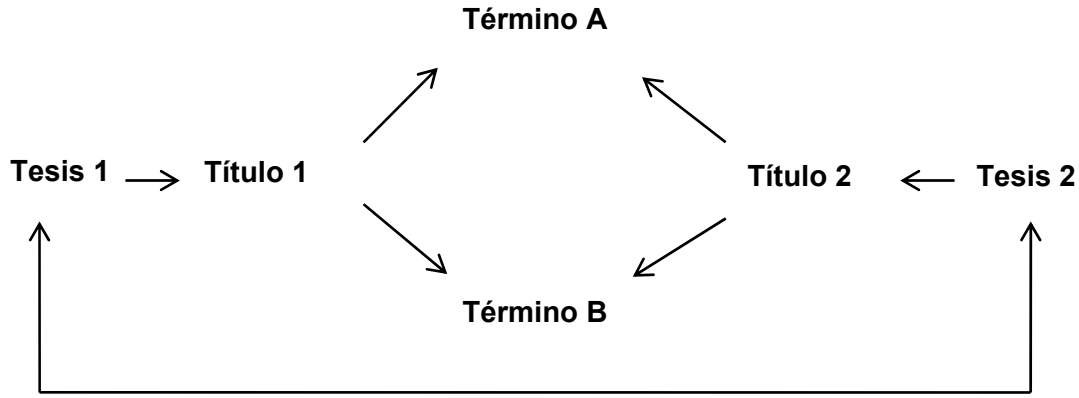
En la representación, **A** y **B** son documentos que han citado las referencias **r1**, **r2**, y **r3**. Ambos documentos han co-citado (citado conjuntamente) las referencias **r1** y **r2** en el caso de **A** y **r2** y **r3** en el caso de **B** (Spinak, 1996). Siguiendo el ejemplo, el principio básico que plantea la co-citación es que si un tercer documento **C** repitiese entre sus referencias alguna de las co-citaciones anteriores, por caso **r1** y **r2**, el documento **A** y **C** estarían relacionados temáticamente desde la perspectiva de los autores citantes. Finalmente, en la medida que una o más co-citaciones (parejas de referencias) se repitan en un conjunto de documentos las relaciones temáticas entre ellos será más fuerte.

Los análisis de co-citación no se realizan solamente sobre las referencias de un documento (autor o fuente citada) sino también sobre términos o expresiones (*co-word analysis)*. Este tipo de estudios permite reflejar la red semántica latente que toda colección de documentos afines guarda en su interior y que es casi imposible de observar a simple vista. Precisamente, este tipo de análisis pone en evidencia las temáticas y sus relaciones generando la posibilidad de establecer un análisis mucho más preciso y real de los contenidos de las fuentes consultadas en un dominio temático dado (Rip y Courtial, 1984).

Desde la perspectiva de la recuperación y visualización de la información la co-citación de términos ha sido ampliamente utilizada. Desde los campos tradicionales de descripción del contenido de las bases de datos bibliográficas (palabras clave/descriptores y resúmenes) e incluso de los títulos puede generarse un instrumento

que describa los vínculos existentes entre los contenidos almacenados facilitando la compresión del dominio a través de su estructura y alcance (Bhattacharya y Basu, 1998; Ding et.al., 2000; Fong y Hui, 2004).

Tomando en consideración el caso que nos ocupa, un listado de tesis del área de la nutrición, se ha seleccionado el título de las mismas para el análisis de co-citación ya que la fuente de datos para este estudio es un listado publicado en la web. Volviendo al ejemplo clásico de co-citación expuesto más arriba, el gráfico explicativo del proceso realizado puede expresarse de la siguiente manera:



Una de las cuestiones más importantes a destacar en base a este modelo presentado es que el resultado es una representación semántica a posteriori basada en el lenguaje natural, es decir, los términos o expresiones utilizadas no han sido controladas por un lenguaje documental, aspecto que se inscribe cada vez más como una tendencia en la representación de dominios científicos (Ibekwe-Sanjuan y Sanjuan, 2002).

El otro item a considerar en este tipo de procesamiento es el método que se va a utilizar para la representación gráfica del producto de la co-citación que, de acuerdo al tipo de análisis que se desee hacer, toman la forma de matrices de datos o índices de parejas.

Para este caso particular se ha optado por el análisis de redes sociales, método ampliamente difundido para este tipo de estudios. Una red social o grafo consiste en un conjunto finito de nodos o actores y de relaciones definidas entre ellos siendo su atributo más importante la capacidad de representación de las estructuras que conjugan los nodos y sus conexiones (Wasserman y Faust, 1998). La representación por medio de una red social sirvió de fase intermedia para la generación de la interfaz gráfica dinámica a través de la técnica de Topic Maps. La razón de esta combinación de técnicas es que el Topic Map no es una herramienta pensada para la representación de análisis de co-citación sino para la realización de redes semánticas asociadas a recursos de información. En resumen, la red social constituye aquí la fuente de datos a partir de la cual se genera el Topic Map.

Los Topic Maps surgieron a partir del propósito de desarrollar un estándar para la documentación técnica de software. El *Grupo de Davenport* – conformado por editores de libros electrónicos – se abocó a esta tarea y la dividió en dos partes: armar una *estructura común* que representara el contenido de los manuales, y desarrollar un *índice base* que sirviera para la generación y modelado de otros índices.

Esta última tarea requirió comprender la semántica esencial de los índices analíticos de libros y llevó a relacionar su consistencia con los modelos de estructura del conocimiento presentes en el contenido de los libros.

El Grupo de Davenport presentó su trabajo en un manuscrito ante la ISO, el cual fue aceptado como norma en 1999 y publicado como tal bajo el nombre de ***ISO/IEC 13250:2000 Topic Maps***. En el año 2003 dicha norma fue adaptada al lenguaje XML (eXtensible Markup Language) mediante la especificación ***XTM 1.0 (XML Topic Maps)*** que se añadió a la misma como enmienda.

Los Topic Maps proporcionan un esquema de representación de estructuras de conocimiento en forma de red semántica, y asociaciones con recursos de información heterogénea.

La conexión semántica que construye entre los conceptos asemeja a una *capa independiente* ubicada por *encima* de los recursos, permitiendo funcionalidad aún si éstos están presentes o no, o si se trata de grandes colecciones de información y en continuo crecimiento. Es decir, ofrecen una *navegación por conceptos* en primer lugar, y la recuperación de la información buscada – si es que existen recursos asociados a dichos conceptos – en segundo lugar.

El núcleo central de este modelo de datos está constituido por tres elementos básicos, bautizados por Steve Pepper (uno de los editores de la especificación XTM 1.0) como el **TAO** de los topic maps:

- ***Topics* (tópicos)**: un tópico representa un tema (o *subject*) en un determinado dominio de aplicación. Un tema es la percepción humana abstracta de una realidad, y el tópico es la representación material o concreta de ese tema.
- ***Associations* (asociaciones)**: una asociación permite describir las relaciones existentes entre dos o más tópicos. Formalmente, una asociación es un elemento de vínculo que define una relación entre dos o más tópicos.
- ***Ocurrences* (ocurrencias)**: un tópico puede estar enlazado a uno o más recursos externos de información, los cuales son relevantes en relación al tópico en cuestión. Estos recursos son llamados ocurrencias del tópico. Técnicamente, una ocurrencia es una conexión que relaciona un tópico con un recurso.

Asimismo, este modelo de datos enriquece su funcionalidad con otras características:

- cada *topic* posee propiedades individuales tales como identificador, tipo, nombre, indicador de tema.
- en las asociaciones entre dos o más tópicos, cada uno de ellos tiene asignado un rol (*role*) como miembro (*member*) de esa *association*. Este rol indica en qué manera el *topic* participa de esa asociación y lo caracteriza según el contexto o *scope* que se determine.
- la declaración de *scopes* o contextos se usa para definir diferentes perspectivas o puntos de vista sobre un mismo conjunto de información (idiomas, audiencia, tiempo o época, autor).
- las ocurrencias o recursos externos de información se encuentran enlazados mediante una referencia que sirve para su localización. Estos recursos no son almacenados en el topic map en sí, por lo tanto forman una "capa separada" del mismo, permitiendo que aquellos sean de cualquier tipo, formato y localización.

Material y método

El universo de tesis de grado recogidas para este estudio es de 146, todas pertenecientes al área de Nutrición de la Universidad FASTA de Mar del Plata. Dado que solo se publica en la web un listado de las mismas (no un catálogo) se recurrió a los títulos para la extracción de los términos representativos de los contenidos. Se asume que el título de este tipo de documento guarda una relación directa con el objeto de la investigación desarrollado.

Para el análisis de co-ocurrencia de términos se creó un archivo de texto ASCII conteniendo el listado de las expresiones extraídas de los títulos. Posteriormente se calcularon las frecuencias absolutas y, en base a estos resultados, se procedió a la lematización del índice de términos con la intervención de una especialista en el área temática abordada.

El resultado de este análisis se procesó a fin de calcular la co-ocurrencia de términos por medio del software Bibexcel[1]. El índice de parejas resultante se utilizó para generar una red social de representación semántica a través del software Pajek[2].

Para el diseño y estructuración del Topic Map se tomó como base la red social de co-ocurrencia de términos, la cual brindó una imagen real de los principales nodos temáticos (topics) y las relaciones (tipos e intensidad) entre los tópicos representados. El mapa final fue generado con el editor TM4L[3] desarrollado por la Departamento de Informática de la Universidad Estatal Winston-Salem (EE.UU.) que trabaja con el estándar XML para Topic Maps (XTM).

Para la visualización del topic map se utilizó Omnigator[4], uno de los navegadores (*browsers*) de topic maps más difundidos en la comunidad investigadora de la Web Semántica. Omnigator es un producto que forma parte de OKS (Ontopia Knowledge Suite), conjunto de herramientas para trabajar con topic maps creada por Ontopia[5], empresa noruega que ofrece servicios de consultoria relacionados con tecnologías de la web semántica.

Resultados y discusión

El primer paso de este análisis, entre todos los procesos llevados a cabo hasta llegar a la conformación del Topic Map, fue la generación de un índice de términos ordenados por su frecuencia (tabla 1). Es importante recordar que los mismos fueron extraídos de los títulos de las tesis con la ayuda de una especialista en el área lo cual permitió, por un lado, obtener la precisión necesaria en los conceptos seleccionados y, por otra, realizar los agrupamientos semánticos adecuados (lematización) para su representación.

[1] http://www.umu.se/inforsk/Bibexcel/
[2] http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[3] http://compsci.wssu.edu/iis/nsdl/
[4] http://www.ontopia.net/omnigator/models/index.jsp
[5] http://www.ontopia.net

Expresiones	Frec.
Niños	35
Evaluación-del-estado-nutricional	28
Conducta-alimentaria	25
Mar-del-Plata	21
Alimentos	16
Nutrientes	16
Adolescentes	14
Ingesta-alimentaria	13
Asistencia-alimentaria	12
Terapéutica-nutricional	10
Obesidad	9
DBT	9
Adultos-mayores	8
Adultos	8
EAN	8
Embarazo	7
Deportistas	7
Estrato-socioeconómico	6
Productos-alimenticios	6
Antecedentes-maternos	5
Actividad-física	4
Bromatología	4
Ejercicio-profesional	4
Enfermedad-celíaca	4

Tabla 1. Índice de frecuencias de las expresiones extraídas de los títulos.

Como puede observarse, muchas expresiones están unidas con un guión con el fin de que al momento de ser procesadas cada una de estas entradas sean tomas como sintagmas y no como términos simples. La frecuencia obtenida permitió establecer la centralidad que cada concepto tendría en la representación.

A continuación de este proceso se continuó con el análisis de co-citación de todas las expresiones utilizadas dando por resultado un índice con las frecuencias de co-ocurrencia de todas las parejas de conceptos existentes (tabla 2).

Parejas de expresiones		Frec.
Niños	Mar-del-Plata	8
Evaluación-del-estado-nutricional	Niños	8
Conducta-alimentaria	Niños	7
Evaluación-del-estado-nutricional	Mar-del-Plata	7
Obesidad	Niños	7
Asistencia-alimentaria	Mar-del-Plata	6
Asistencia-alimentaria	Niños	6
Conducta-alimentaria	Nutrientes	6
Conducta-alimentaria	Adolescentes	5
Evaluación-del-estado-nutricional	Adultos	4
Ingesta-alimentaria	Evaluación-del-estado-nutricional	4
Adultos-mayores	Mar-del-Plata	4
Evaluación-del-estado-nutricional	Asistencia-alimentaria	4
DBT	Conducta-alimentaria	3
Conducta-alimentaria	Mar-del-Plata	3
Nutrientes	Niños	3
Terapéutica-nutricional	Deportistas	3
Sobrepeso	Obesidad	3

Tabla 2. Principales frecuencias de co-ocurrencias de conceptos.

Tal como se aprecia, las frecuencias más altas de co-ocurrencia determinarán los conceptos centrales de la representación. Dados estos resultados, la *niñez*, las evaluaciones relativas a los *estados nutricionales* y el lugar geográfico predominante en los estudios de campo llevados a cabo (*Mar del Plata*) constituirán el principal núcleo sobre el que se conformará nuestra red semántica. Sin embargo, más allá de estas apreciaciones objetivas, no es posible establecer a través de este índice la real dimensión de los vínculos existentes entre todos los conceptos presentes en el conjunto. Como mencionábamos en la introducción, es necesario recurrir a un método de representación gráfica que nos permita visualizar el tejido de relaciones existentes. Para ello se recurrió a una representación de redes sociales (figura 1).

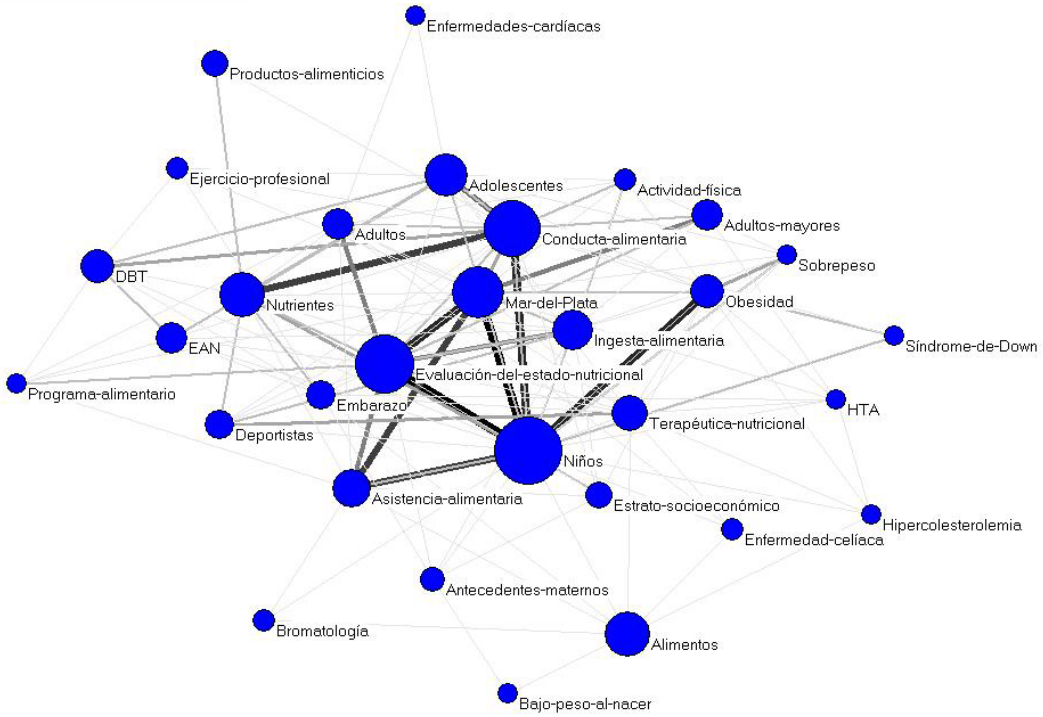


Figura 1. Representación por medio de una red social de la co-citación de expresiones.

La imagen resultante de la red social permite obtener una visualización concreta de la forma en que los conceptos analizados interactúan, es decir, nos proporciona la estructura en que este campo de conocimiento se articula desde el punto de vista de sus límites y conformación interna. La información que aporta esta gráfica se centra además en la intensidad de las relaciones, determinada por el grosor de los enlaces y de la centralidad de los conceptos intervinientes dada por el tamaño de los nodos (frecuencia de las expresiones) y la integración que los mismos tienen con el resto de los nodos de la red[6]. Al mismo tiempo, dado que se trata de la representación de tesis, es posible evaluar cuales son las principales líneas de investigación desarrolladas en esta campo dentro de la institución seleccionada.

A partir de la obtención de la red social, ésta pasó a constituir finalmente la base sobre la que se construyó el Topic Map, tomando como referencia las expresiones (topics) y las relaciones establecidas entre ellas a partir del patrón de co-ocurrencias existente. En este punto es importante aclarar que fueron respetadas las asociaciones dadas entre los conceptos a partir de la imagen del la red social, sin tipificar a que tipo pertenece cada una (*association type)*. La generación del Topic Map se llevó a cabo con un fin instrumental, es decir, para poder diseñar una interfaz gráfica de recuperación dinámica que pueda asociar las expresiones con las tesis correspondientes.

Para llevar a cabo este paso se trabajó con un editor de topics que posibilitó emular los resultados obtenidos (figura 2).

[6] La centralidad de los nodos o actores de una red social no debe asociarse con la posición espacial que éstos ocupan sino con el grado de intregación que tienen con el resto. Existen varias medidas de centralidad: de grado, de cercanía y de intermediación.

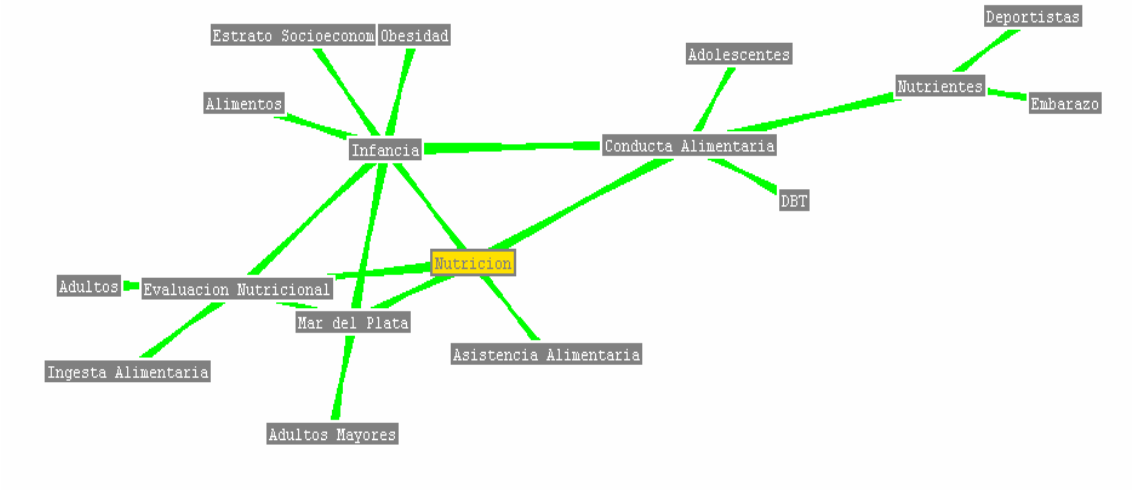


Figura 2. Topic Map generado a través del editor TM4L tomando como base la red social de co-ocurrencia de expresiones.

A continuación se utilizó la aplicación Omnigator que permite en un entorno web cargar y navegar sobre cualquier topic map a través de la interfaz de un *browser*. Para ser reconocido como tal y cargado correctamente, el topic map debe estar en notación *XTM* (XML Topic Map) o *LTM* (Linear Topic Map Notation). En nuestro caso, el topic map se desarrolló con el editor TM4L que respeta la especificación XTM 1.0 (figura 2). El resultado final de este proceso se plasma en la visualización del índice de tópicos disponibles para su consulta (figura 3), el grafo de la red semántica (figura 4) y, a modo de ejemplo, la página de enlace con una de las tesis trabajadas (figura 5).

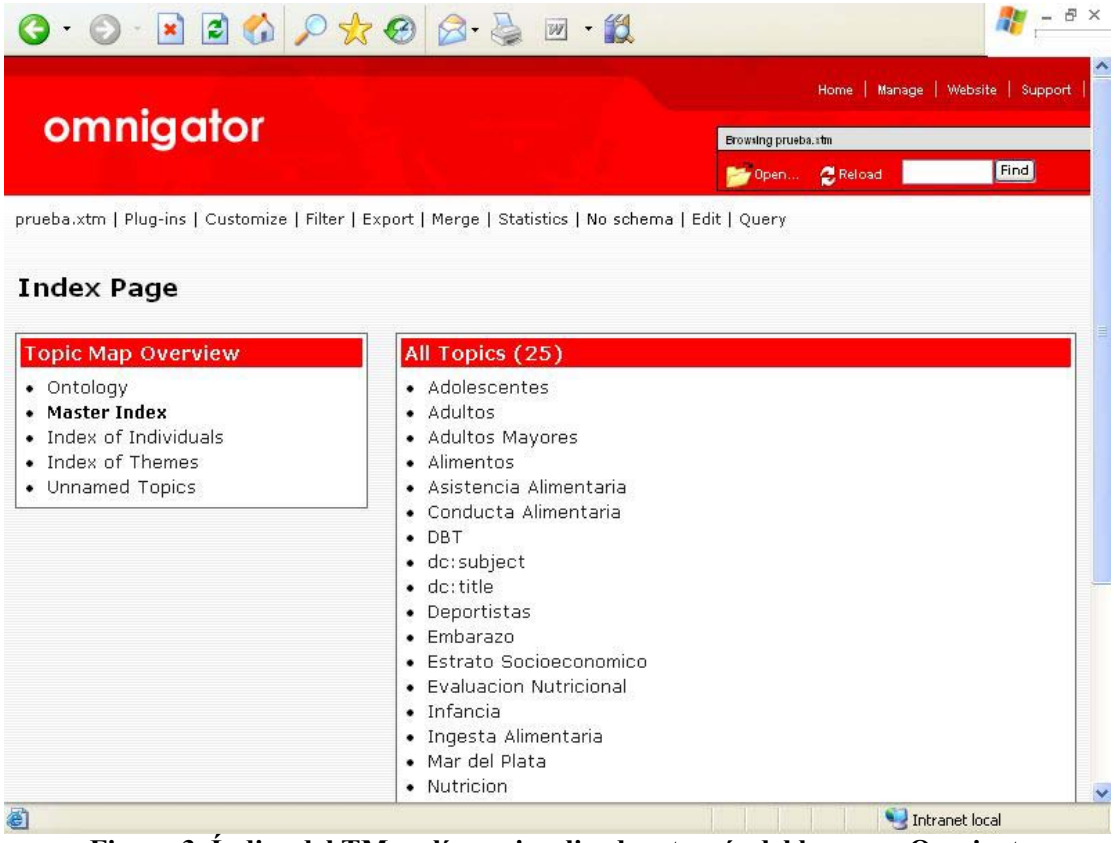


Figura 3. Índice del TM en línea visualizado a través del browser Omnigator.

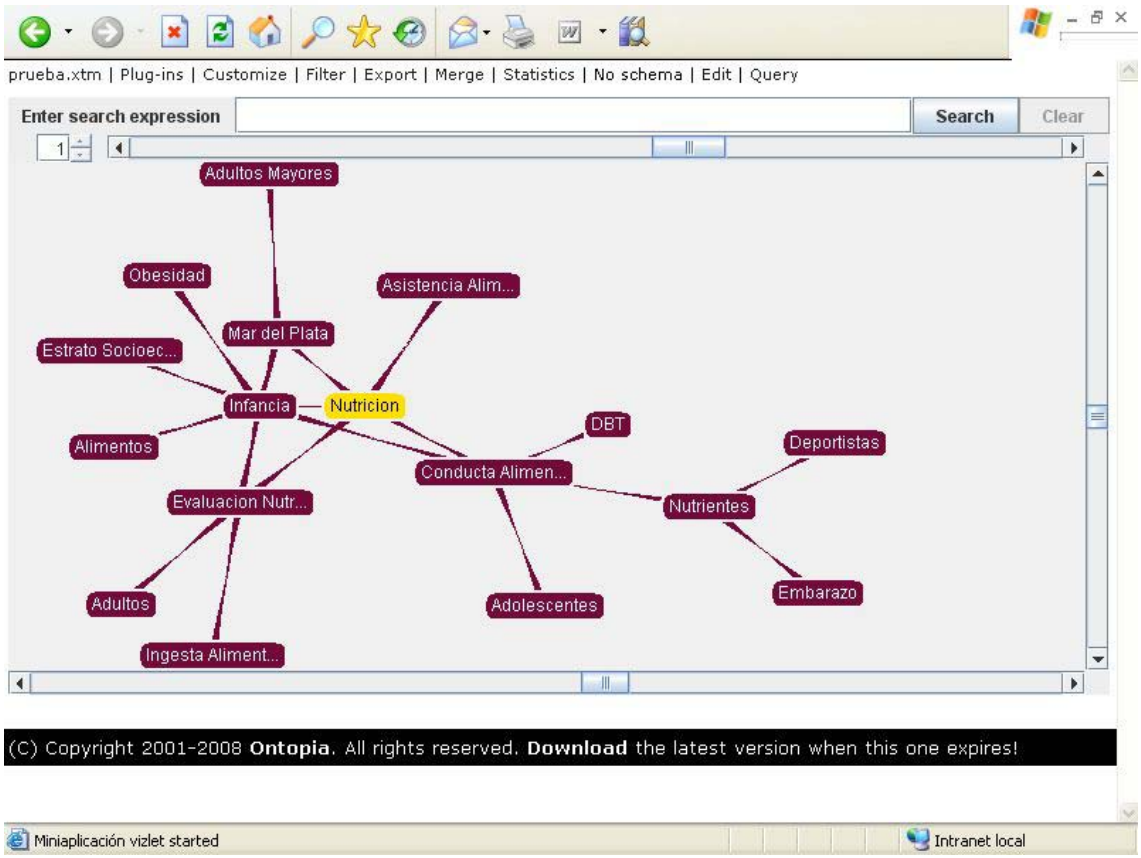


Figura 4. Pantalla del grafo del TM con enlaces a la colección de tesis.

Figura 5. Registro de una tesis (ocurrencia) enlazada a un topic (*Hábitos Alimentarios*).

Conclusiones

Las nuevas formas de visualización y recuperación de información presentan vías de desarrollo interesantes para mejorar la calidad de las representaciones de los opacs o colecciones de recursos de información. Al mismo tiempo, el camino para llegar a ellas puede tomar formas innovadoras que, como en este caso, conjugan diferentes metodologías de análisis en la búsqueda de un nuevo producto.

En el caso que se presentó aquí el análisis de co-citación de términos (*coword analysis*) resultó ser una herramienta de muy buen desempeño en la representación de la estructura conceptual del campo temático objeto de estudio. No sólo evidenció la trama de relaciones internas, sino que además aportó información real del verdadero peso de los temas presentes. Teniendo en cuenta que se trata de una colección de tesis, los resultados también demostraron las líneas de investigación más consolidadas, visibles a través de la intensidad de los vínculos entre los temas abordados claramente reflejados en la imagen de la red social. La información aportada por estos resultados en torno a un colectivo documental supera en mucho a cualquier forma de representación tradicional, dando al potencial usuario de esta aplicación datos precisos y reales sobre el espacio de conocimiento con el que va a interactuar.

Finalmente, la aplicación del Topic Map se enlaza como una técnica complementaria para hacer operativa la red semántica elaborada desde el punto de vista de la unión del concepto o expresión semántica con el o los recursos de información que la contienen. Si bien es cierto que en este producto quedan algunos cabos sueltos por resolver desde la perspectiva de la eficiencia en términos de recuperación de información, los resultados hasta aquí alcanzados inclinan la balanza hacia el signo positivo.

Bibliografía

- Bhattacharya, S.; Basu, P. K. (1998). Mapping a research area at the micro level using co word analysis. Scientometrics; 43 (3), 359-72.
- Boriana Ditcheva, Darina Dicheva. (2007). Visual Browsing and Editing of Topic Map-Based Learning Repositories. TMRA 2006 (Topic Maps Research and Applications), Leipzig, Germany, October 11-12, 2006. Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science) nº 4438, 44-55.
- Colmenero Ruiz, María Jesús. Introducción al modelo topic maps. Revista Digital de Biblioteconomia e Ciência da Informaçao, l (3), 77-102.
- Costa, J. (1998). La esquemática: visualizar la información. Barcelona: Paidós.
- Crosby, A. W. (1997). The Measure of Reality: Quantification and Western Society 1250-1600. London, Cambridge University Press.
- Dicheva, Darina; Dichev, Christo. (2006). TM4L: Creating and browsing educational topic maps. British Journal of Educational Technology; 37(3), 391-404.
- Ding, Y; Chowdhury, G G; Foo, S. (2000). Incorporating the results of co-word analyses to increase search variety for information retrieval. Journal of Information Science; 26 (6), 429-51.
- Fong, A C M; Hui, S. C. (2004). Document retrieval from a citation database using conceptual clustering and co-word analysis. Online Information Review; 28 (1). 22-32.

- Ibekwe-Sanjuan, Fidelia; Sanjuan, Eric. (2002). From term variants to research topics. Knowledge Organization; 29 (3/4), 181-197.
- Klovdhal, A. S. (1981). A note of images of social networks. Social Networks 3, 197-214.
- Liberatore, Gustavo y otros. (2004). Educación a distancia y topic maps: una aproximación a la problemática de la enseñanza de la indización. Biblios: Revista Electrónica de Ciencias de la Información, 6 (21/22).
- Librelotto, Giovani Rubert. (2005). Topic Maps: da sintaxe à semântica. Departamento de Informatica, Escola de Engenharia, Universidade do Minho, Braga. (Tesis presentada para el título de doctor en informática).
- Miguel, S.; Moya-Anegón, F.; Herrero-Solana, V. (2007). El análisis de co-citas como método de investigación en Bibliotecología y Ciencia de la Información. Investigación Bibliotecológica, 21 (43), 139-155.
- Rip, A. P.; Courtial, J. P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics, Scientometrics, 6, 381-400.
- Small, H. (1973). Cocitation in scientific literature – New measure of relationship between 2 documents. Journal of de American Society for Information Science, 24 (4), 265-269.
- Spinak, E. (1996). Diccionario de enciclopédico de bibliometría, cienciometría e informetría. Caracas, Unesco.
- Steve Pepper. (2000). The TAO of Topic Maps. In Proceedings of XML Europe 2000,Paris, France.
- Vargas-Quesada, Benjamín. (2005). Visualización y análisis de grandes dominios científicos mediante redes Pathfinder (PFNET). Tesis doctoral. Granada, La Universidad.
- Wasserman, S.; Faust, K. (1998). Social Network Analysis. Methods and Aplications. Cambridge: Cambridge University Press.