Análisis y Visualización de Comunidades Científicas con Información Extraída de la Web

F. de la Rosa T., S. Pozo y R. M. Gasca

***Resumen*--El objetivo del presente trabajo es el de representar gráficamente las estructuras de una determinada área de conocimiento, a partir de datos recopilados de Internet. El actual desarrollo de la red Internet ha facilitado el acceso a nuevos recursos, esto ha contribuido a crear nuevas disciplinas que pretenden explotar estos recursos. Los principales problemas para su explotación son su localización y su procesamiento. A lo largo de este trabajo se combinarán técnicas de extracción, análisis y visualización de datos, para definir una arquitectura que solucione los problemas de procesamiento que plantean estos recursos. En concreto, en este trabajo exploraremos dos de las estructuras más importantes que definen una comunidad, los temas que generan mayor interés y su red social. Como ejemplo de aplicación de la arquitectura y de las técnicas utilizadas, se presenta un estudio sobre la comunidad de las JISBD, a partir de datos extraídos de Internet.**

***Índice de Términos*--análisis de palabras asociadas, análisis de redes sociales, cibermetría, content web mining, data clearing, extracción de información, visualización.**

I. Introducción

Extraer conocimiento de los recursos disponibles en Internet es una de las áreas de investigación que más interés está generando en la comunidad científica. Existen muchas áreas implicadas: Webmining, Cybermetría, Webometría, PLN, etc. El objetivo de este trabajo es la implementación de una arquitectura que permita las labores de extracción o descubrimiento de conocimiento. Entendiendo la transformación de información en conocimiento como una "extracción de información implícita, no trivial, previamente desconocida y potencialmente útil" [1]. Como veremos, para conseguir este objetivo, la arquitectura ha necesitado integrar técnicas pertenecientes a distintas áreas, entre estas se pueden destacar: *Extracción de Información*, *Recuperación de Información*, *Cienciometría*, *Técnicas de Visualización* y *Técnicas de la Teoría de Grafos*.

El presente trabajo se centra en analizar dos de las estructuras más importantes que definen una comunidad de investigadores: la red de social y las áreas temáticas que la definen. Para realizar estos análisis ha sido necesario conocer la información bibliográfica de la comunidad. Esta información permitirá analizar la red social a partir de su red de co-autorías, así como aplicar la técnica de las co-palabras, para analizar sus áreas temáticas. Como ejemplo ilustrativo de estas técnicas, presentamos un estudio aplicado sobre la comunidad de las *Jornadas de Ingeniería del Software y Bases de Datos* (JISBD). Para ello se ha extraído de Internet la información bibliográfica de la comunidad, información que recientemente ha sido publicada en la base de datos DBLP *(Digital Bibliograph & Library Proyect)*. DBLP es una base de datos cuyos servidores proporcionan información bibilográfica sobre las revistas y congresos más importantes sobre *Computer Science*.

Aunque estas técnicas puedan parecer sólo aplicables a la información bibliografica, lo cierto es que tiene gran aplicabilidad en cualquier sistema que pueda modelarse como una red o grafo, donde los nodos son entidades (autores, palabras, páginas web, artículos, clientes, empleados, empresas, productos, etc) y las aristas nos informan sobre las relaciones existentes entre las distintas entidades (publicación conjunta o colaboración, aparición conjunta, referencia, amistad, subordinado, compra, etc). Por ejemplo, la técnica de las Co-Autorías puede ser de utilidad para analizar las relaciones que se producen entre los miembros de un foro de discusión o entre los empleados de una empresa y la técnica de las Co-Palabras permite el análisis de las temáticas tratadas en cualquier corpus de documentos o campo con información textual (páginas HTML, mensajes de email o campos de una base de datos).

Como se describe en el trabajo [2], para crear nuevo conocimiento a partir de la creciente cantidad de información que disponen los sistemas, es necesario la utilización de *herramientas que permitan enfocar la información desde distintas perspectivas*. La construcción de estas perspectivas estaría guiada por nuestros *objetivos o necesidades informativas*. Por ello en este trabajo, proponemos una arquitectura que permita alcanzar nuestras necesidades informativas mediante el adecuado modelado de dichas redes. Por ejemplo, para el problema que planteamos, el análisis de las estructuras que forman las JISBD, no tienen las mismas necesidades de información un experto en Ingeniería del Software o en Bases de Datos que un investigador novel. El experto en Ingeniería del Software o en Bases de Datos estaría interesado en la búsqueda de autores afines o de líneas de investigación emergentes dentro de su área, mientras que un investigador novel estaría más interesado en conocer las temáticas que forman las jornadas o conocer a los autores más

Este trabajo ha sido parcialmente financiado por el proyecto LJC/GGM-55706 de Ayudas de Acciones Coordinadas entre Grupos de Investigación de la Consejería de Educación y Ciencia de la Junta de Andalucía.

F. de la Rosa T. (e-mail: ffrosat@lsi.us.es).
S. Pozo (e-mail: sergio@lsi.us.es).
R. M. Gasca (e-mail: gasaca@lsi.us.es)

Depto de Lenguajes y Sistemas Informáticos, Universidad de Sevilla.
Avd. Reina Mercedes S/N, 41012, Sevilla, España. http://www.lsi.us.es/

influyentes capaces de orientarle en sus investigaciones. También una empresa podría tener necesidades diferentes a las anteriores, con la idea de obtener alguna ventaja competitiva. Ésta podría por ejemplo estar interesada en buscar expertos que colaborasen en algún proyecto conjunto o conocer los intereses de los investigadores de empresas competidoras.

Como vemos, las necesidades informativas son muy amplias y la información a analizar es muy extensa y difusa, a veces *sólo disponibles desde Internet, caso de las comunicaciones de las JISBD*. Por ello es necesario el desarrollo de aplicaciones que nos proporcionen medios para descubrir la información que necesitamos. Aplicaciones que deberán conjugar técnicas procedentes de diversas áreas. Para describir la arquitectura y las técnicas utilizadas para resolver estos problemas, el artículo se ha dividido en las siguientes secciones: en la sección 2 se describe la arquitectura desarrollada, en la sección 3 se detalla el proceso de aclarado de errores de los datos extraídos y en las secciones 4 y 5 se describe las técnicas utilizadas para analizar las estructuras de las JISBD, la red social y las áreas temáticas.

II. Arquitectura

En un trabajo previo [3] se presentó una arquitectura que integraba todos los procesos necesarios para descubrir conocimiento, desde la extracción de información, hasta los procesos de análisis y de visualización. En la Fig*. 1* podemos observar las distintas fases de la arquitectura. En aquella ocasión se aplicó la arquitectura para producir resúmenes de prensa con información extraída en distintos periódicos digitales y se expuso ampliamente, una metodología para construir un crawler/wrapper utilizando *WebL* [4]. Esta metodología se basa en dos políticas: la primera, indica al sistema de extracción cómo navegar y la segunda, cómo extraer los contenidos. Una de las características más interesantes de la arquitectura es el *desacoplamiento* del proceso de extracción de información, del proceso de explotación de los datos extraídos (análisis y visualización). Esto se ha conseguido haciendo uso de una *Base de Datos Semiestructurada (BDSE)*. En este trabajo, se ha enriquecido la arquitectura, introduciendo nuevos módulos que permiten, corregir errores en los datos extraídos y analizar visualmente estos datos, aplicando la arquitectura a un problema más complejo, como es la exploración de áreas de conocimiento.

Siguiendo la arquitectura, se ha construido un crawler/wrapper, al cual se le ha indicado cómo navegar a través de DBLP y cómo extraer la información de sus páginas web. Posteriormente esta información se ha almacenado en una BDSE, la cual recoge la información sobre los distintas ediciones de las JISBD, en concreto se ha almacenado la información sobre los artículos, los títulos, los autores y las fechas en que se publicaron dichas comunicaciones.

Una vez extraída la información de DBLP, se han realizado tres procesos para conseguir la visualización de la información. El primer proceso ha consistido en filtrar los datos erróneos. Aunque los datos que ofrece DBLP sobre los autores son muy precisos, sobre todo si son comparados con otros sistemas bibliográficos como CiteSeer, hemos podido comprobar, que siguen existiendo identificadores de autores duplicados. Para la eliminación de estos errores se ha desarrollado una herramienta que será descrita en la sección 3.

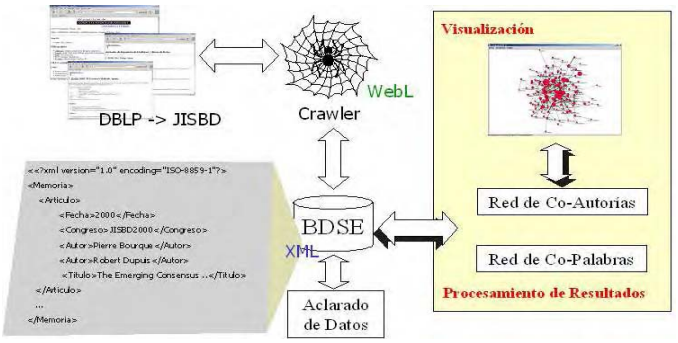


Fig. 1. Arquitectura del sistema.

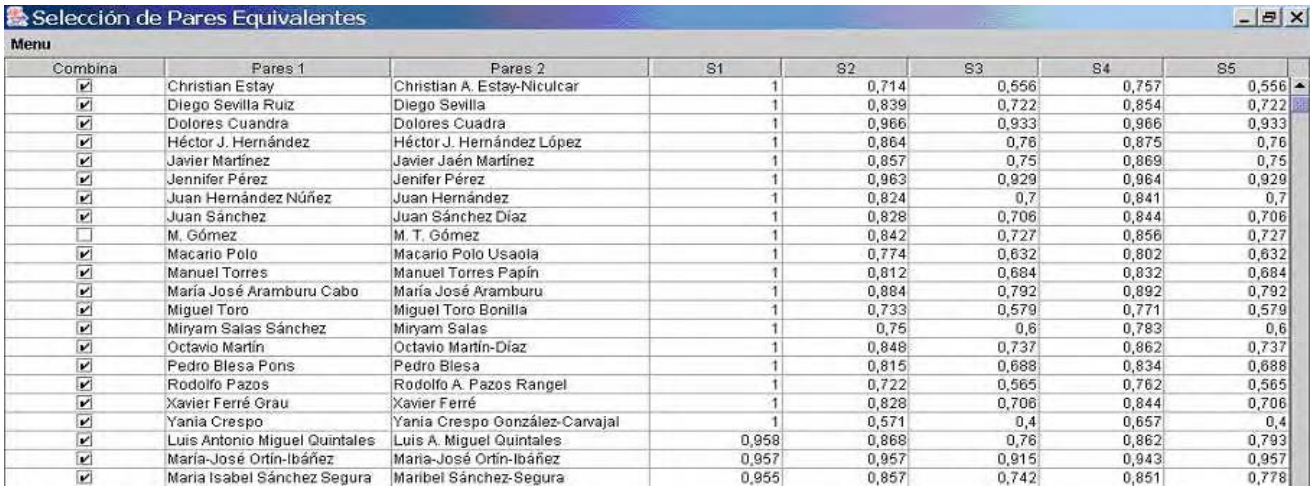
Una vez la base de datos ha sido filtrada de errores, en la fase de procesamiento y análisis, dos procesos se encargan de construir la red de Co-Autorias y de Co-Palabras. Finalmente estas redes se analizan visualmente. Para llevar a cabo esta última tarea, se ha desarrollado una herramienta que de forma visual asiste al usuario en el desarrollo y la exploración de nuevas perspectivas. Los procesos de análisis y visualización de las redes, serán descritos detalladamente en las secciones 4 y 5.

III. Data Clearing o aclarado de datos

Tras la extracción de la información de DBLP sobre las comunicaciones de las JISBD, descubrimos que existían algunos identificadores sinónimos de autores (identificadores escritos de formas distintas pero que hacían referencia al mismo autor). Por ejemplo *"Rafael Ceballos"*, era un identificador sinónimo de *"R. Cevallos"*. Este proceso de eliminación de errores es conocido como *Data Clearing*, en [5] podemos encontrar una revisión de las distintas técnicas de aclarado y así como su taxonomía.

Para eliminar estos errores se ha desarrollado una herramienta que calcula, a partir de los identificadores (en este caso, cadenas de caracteres), varias medidas de similitud entre pares de ellos. El objetivo de esta herramienta es facilitar al usuario la selección de *identificadores sinónimos*. Para ello la aplicación muestra los conjuntos de pares de identificadores, ordenados según su similitud. En la gran mayoría de los casos los identificadores con una alta similitud son identificadores similares, aunque no siempre. A partir de estas selecciones, la herramienta forma *grupos de identificadores sinónimos*, que luego son sustituidos por un *identificador representativo* de todos ellos, por defecto este identificador es el más frecuente, aunque el usuario puede cambiarlo por otro o crear uno nuevo. Salvo la selección de los pares, el proceso es totalmente automático.

Inicialmente la BDSE disponía de 539 autores y se detectaron 38 autores con identificadores sinónimos (un 7%). Esta herramienta ha permitido la corrección de dichos identificadores y como consecuencia se ha conseguido disminuir aproximadamente en un 10% de las relaciones erróneas que poblaban la BDSE. Este porcentaje consideramos que es bastante significativo.



Combina	Pares 1	Pares 2	S1	S2	S3	S4	S5
☑	Christian Estay	Christian A. Estay-Niculcar	1	0,714	0,556	0,757	0,556
☑	Diego Sevilla Ruiz	Diego Sevilla	1	0,839	0,722	0,854	0,722
☑	Dolores Cuandra	Dolores Cuadra	1	0,966	0,933	0,966	0,933
☑	Héctor J. Hernández	Héctor J. Hernández López	1	0,864	0,76	0,875	0,76
☑	Javier Martínez	Javier Jaén Martínez	1	0,857	0,75	0,869	0,75
☑	Jennifer Pérez	Jenifer Pérez	1	0,963	0,929	0,964	0,929
☑	Juan Hernández Núñez	Juan Hernández	1	0,824	0,7	0,841	0,7
☑	Juan Sánchez	Juan Sánchez Díaz	1	0,828	0,706	0,844	0,706
☐	M. Gómez	M. T. Gómez	1	0,842	0,727	0,856	0,727
☑	Macario Polo	Macario Polo Usaola	1	0,774	0,632	0,802	0,632
☑	Manuel Torres	Manuel Torres Papín	1	0,812	0,684	0,832	0,684
☑	María José Aramburu Cabo	María José Aramburu	1	0,884	0,792	0,892	0,792
☑	Miguel Toro	Miguel Toro Bonilla	1	0,733	0,579	0,771	0,579
☑	Miryam Salas Sánchez	Miryam Salas	1	0,75	0,6	0,783	0,6
☑	Octavio Martín	Octavio Martín-Díaz	1	0,848	0,737	0,862	0,737
☑	Pedro Blesa Pons	Pedro Blesa	1	0,815	0,688	0,834	0,688
☑	Rodolfo Pazos	Rodolfo A. Pazos Rangel	1	0,722	0,565	0,762	0,565
☑	Xavier Ferré Grau	Xavier Ferré	1	0,828	0,706	0,844	0,706
☑	Yania Crespo	Yania Crespo González-Carvajal	1	0,571	0,4	0,657	0,4
☑	Luis Antonio Miguel Quintales	Luis A. Miguel Quintales	0,958	0,868	0,76	0,862	0,793
☑	María-José Ortín-Ibáñez	María-José Ortín-Ibáñez	0,957	0,957	0,915	0,943	0,957
☑	Maria Isabel Sánchez Segura	Maribel Sánchez-Segura	0,955	0,857	0,742	0,851	0,778

Fig. 2. Proceso de selección de identificadores sinónimos, ordenados por la métrica s1.

Las medidas de similitud utilizadas en esta herramienta, se basan en el algoritmo de la subsecuencia común más larga a dos secuencias, técnica de programación dinámica de complejidad polinómica. En nuestro caso las secuencias eran las cadenas de caracteres de cada una de los identificadores. Seguidamente mostramos las definiciones de las métricas utilizadas:

$$s1 = \frac{|scl(x,y)|}{\min(|x|,|y|)}; \qquad s2 = \frac{2*|scl(x,y)|}{|x|+|y|};$$

$$s3 = \frac{|scl(x,y)|^2}{|x|*|y|}; \qquad s5 = \frac{|scl(x,y)|}{\max(|x|,|y|)};$$

$$s4 = 0.33*s1 + 0.33*s2 + 0.33*s3 \qquad (1)$$

En estas definiciones $scl(x,y)$ es el algoritmo que calcula la *subsecuencia común más larga* a los identificadores *x* e *y*, y el cardinal representa la longitud de los identificadores. También es posible utilizar otras medidas de similitud entre cadenas, como *Levenshtein distance* o *edit distance*, de coste polinóminco *O(n*m),* pero menos vulnerable a la permutación de palabras dentro de las cadenas, como por ejemplo el caso "Rafael Ceballos" y "Ceballos, R.", y la distancia *LikeIt* [6] de coste *O(n+m)* y que resulta apropiada para comparar cadenas de gran tamaño (*n* y *m* son las longitudes de las dos cadenas). En los trabajos [7,8,9,10,11] podemos encontrar una amplia revisión de las medidas de similitud entre cadenas.

El proceso descrito anteriormente se ha incorporado a la arquitectura y su utilización en diferentes BDSE ha sido muy efectiva, ya que el número de parejas de palabras con alto porcentaje de similitud es escaso y en cuestión de pocos minutos un usuario puede decidir que parejas de identificadores son sinónimas. En la Fig. 2 se puede observar el proceso de selección ofrecido por la aplicación.

Aunque el cálculo de las medidas de similitud para todas las parejas de identificadores en las BDSE tratadas es razonable, consideramos que este cálculo puede ser muy costoso en BDSE de mayor tamaño. Proponemos como trabajos futuros, la utilización de la información contextual que nos ofrecen estas redes, para mejorar el rendimiento del algoritmo, así como automatizar el proceso de toma de decisiones, utilizando técnicas de clustering [12].

IV. Red de co-autorías

Una vez filtrado los datos erróneos de la red, se desarrolló un proceso para calcular la *red de co-autorías*, posteriormente esta red fue analizada visualmente. El objetivo de esta técnica es la representación de los *colegios invisibles* que se forman en las distintas áreas en su etapa de madurez. En el trabajo [13] se definen los colegios invisibles como *"círculos de investigadores influenciados por unos pocos investigadores de alta productividad"*. Existen varias métricas para representar los colegios invisibles a través de las relaciones entre los autores: métricas basadas *co-citas* [14,15] (número de citas conjuntas que reciben dos autores), métricas basadas en la *centralidad sociométrica* [16] y métricas basadas en la *distancia geodésica* [17]. En contraste con estos trabajos el componente interactivo, tanto de la exploración como de la visualización de la red, resulta fundamental. Por ello se decidió que la metáfora más apropiada para representar la red era el sociograma [18], esta representación nos permite visualizar al mismo tiempo el impacto de los autores (2) y sus relaciones de co-autorías (3).

$$\text{impacto}(a_i) = \#\,\text{Publicaciones de a}_i \qquad (2)$$

$$\text{coautoría}(a_i, a_j) = \#\,\text{Colaboraciones conjuntas entre los autores a}_i \text{ y a}_j \qquad (3)$$

donde a_i y a_j representan a dos autores distintos y # representa el número de publicaciones o de colaboraciones conjuntas. Para una revisión de las técnicas de visualización de las co-autorías se puede consultar [19].

Existe un amplio rango de herramientas de análisis bibliográfico que utilizan distintas metáforas de visualización, podemos destacar: *Bibexcel*, *Brookes ToolBox* [21], *Metric*,

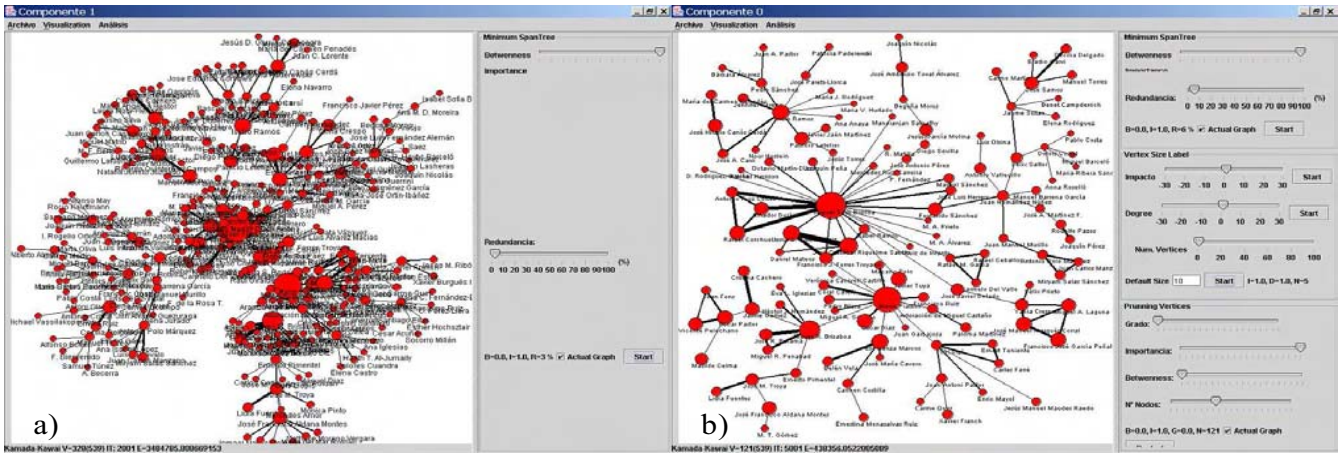


Fig. 3. a) Imagen holística; b) Filtrado de los 121 autores más importantes.

Dataview y más recientemente *DIVA* [22], *CiteSpace* [23] y *DBL-Browser* [20]. Esta ultima herramienta permite la consulta offline de bases de datos bibliográficas digitales, como DBLP, esta herramienta no sólo permite la consulta de las fichas bibliográficas de los autores y navegar a través de ellas, sino también implementa algunas metáforas visuales basadas en el ego, por ejemplo, permite visualizar la producción temporal, diagrama de barras, y las colaboraciones realizadas por el autor, sociograma. A diferencia de esta herramienta, la herramienta de visualización que se ha desarrollado en este trabajo, permite una visión holística, a partir de la cual podemos modelar la red, pudiendo crear distintas perspectivas de la misma. Seguidamente presentamos los procesos utilizados para obtener los mapas que se exponen en este trabajo. Nuestro primer objetivo fue construir un mapa holístico representativo de la comunidad de las JISBD, para ello seguimos los siguientes procesos:

- Partiendo de la red completa, observamos que debido al gran número de nodos y relaciones, alrededor de 539 nodos y 1190 aristas, era imposible visualizar íntegramente la red. Nuestra primera decisión fue calcular las redes conexas, y de entre ellas, seleccionamos la red con más nodos, 328 de los 539 autores, considerándola como representativa del núcleo de la comunidad.
- Aún así el número de relaciones, aproximadamente 814, era excesivo para una visualización adecuada. Por tanto para visualizar correctamente la red, generamos un árbol de expansión mínima, permitiendo un 3% de redundancia en las relaciones. La selección de las relaciones se realizaron por orden de importancia. En la Fig. 3a podemos observar la escena final.

Nuestro segundo objetivo fue observar los autores con más impacto en las JISBD, así como los autores que trabajan con ellos, para ello seguimos un proceso semejante al anterior, que resumimos a continuación:

- Para obtener una visión óptima de la escena, seleccionamos los 121 autores con más impacto del núcleo y realizamos un filtrado de las relaciones, generando un árbol de expansión mínimo con un 6% de redundancia Fig. 3b.
- Para finalizar decidimos aumentar el tamaño de las etiquetas de los 7 autores con más impacto. El resultado final de la escena lo podemos observar en la Fig. 4.

Respecto al algoritmo para la distribución de los nodos en la pantalla, se ha utilizado el algoritmo de Kamada-Kawai [24].

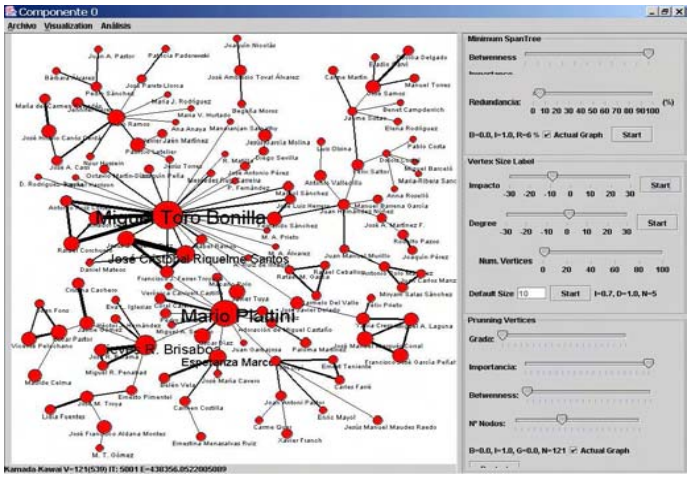


Fig. 4. Escena final de los 121 autores con más impacto de las JISBD.

V. RED DE CO-PALABRAS

El objetivo de la técnica de las co-palabras, es la identificación de focos o centros de interés, así como las relaciones que existen entre ellos. Estos focos o centros de interés son asimilables a las áreas temáticas que componen las JISBD. Y nuestro objetivo es producir una serie de mapas, a partir de la información obtenida con esta técnica, que permitan comprender la estructura temática de las JISBD. Esta técnica han sido ampliamente analizada en [25,26,27] y en [28] podemos encontrar la descripción de una herramienta, *Leximap*, que utiliza esta metodología junto con técnicas de *análisis de lenguaje natural (NLA)*. Los fundamentos de la técnica, se basan en la construcción de una red de co-palabras, a partir de un conjunto de documentos o de sus títulos y resúmenes. La red de co-palabras se compone de nodos que representan la importancia de las palabras (4) y de aristas

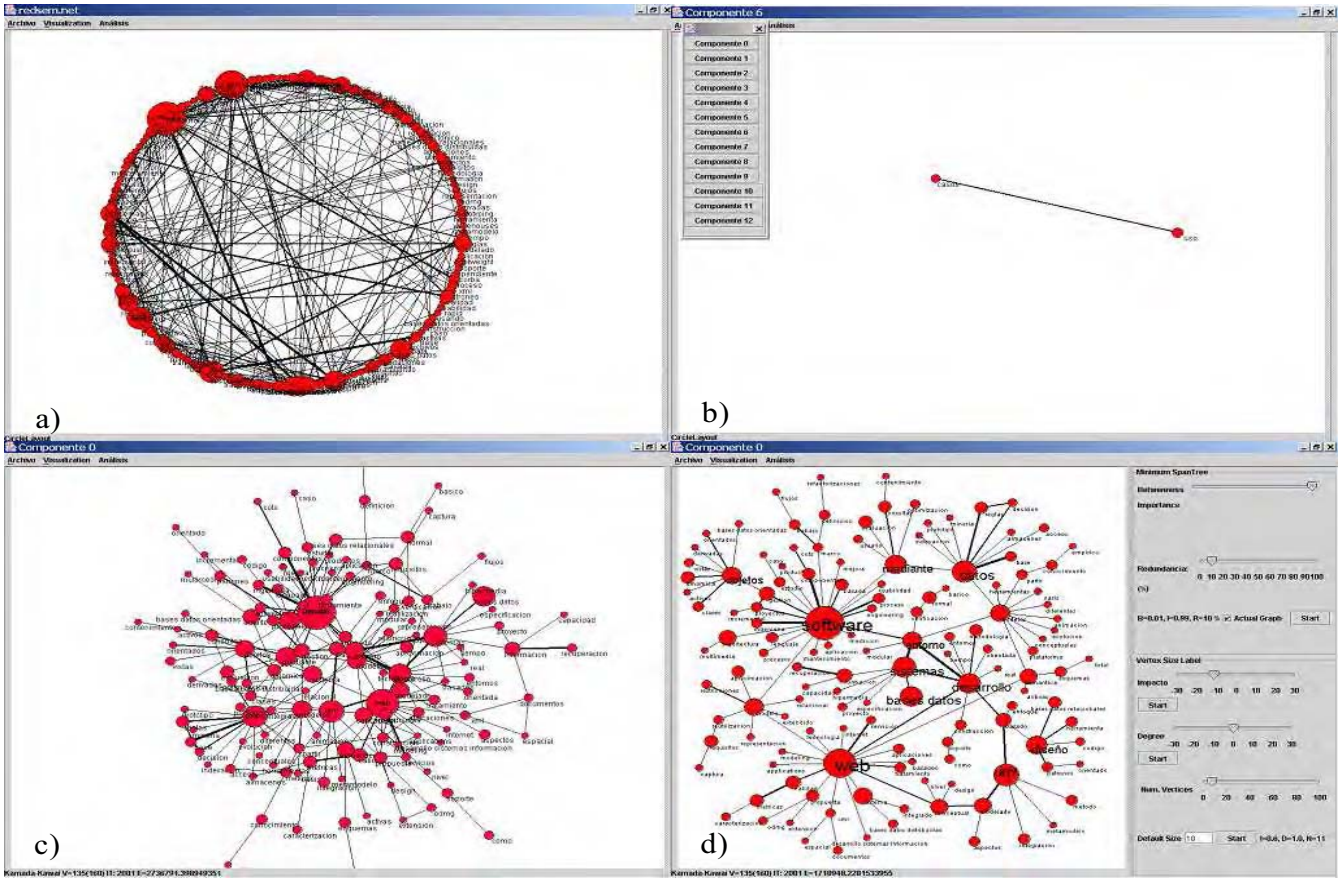


Fig. 5. a) Red de términos; b) Proceso de selección de las redes conexas. Selección de la red conexa 6; c) Red conexa 0, núcleo de la red de términos; d) Escena final de la red de términos.

cuyos pesos representan las co-ocurrencias entre dos palabras (5). A diferencia de la técnica clásica y debido a la baja redundancia de los datos disponibles (solamente disponemos de los títulos de las comunicaciones recopiladas) asumimos en este trabajo que los términos más frecuentes son los más importantes. Las métricas utilizadas son las siguientes:

$$\text{impacto}(w_i) = \#\,\text{Apariciones en títulos de artículos} \qquad (4)$$

$$\text{coaparición}(w_i, w_j) = \#\,\text{Apariciones conjuntas de las palabras } w_i \text{ y } w_j \text{ en títulos de artículos} \qquad (5)$$

donde w_i y w_j son dos palabras o tokens distintos y # representa el número de apariciones o de apariciones conjuntas. El proceso de producción de estos mapas temáticos también difiere de la técnica clásica y consta de las siguientes fases:

- Determinar y normalizar los tokens en que se dividen los títulos. Por ejemplo, el título *"BD-Web: una propuesta metodológica basada en UML y XML"* sería normalizado como *"db-web una propuesta metodológica basada en uml y xml"*.
- Eliminar las palabras huecas (palabras sin significado léxico) a partir de una lista de palabras huecas. Por ejemplo, el titulo normalizado *"db-web una propuesta metodológica basada en uml y xml"* se transformaría en *"db-web propuesta metodológica basada uml xml"*.
- Búsqueda de bigramas y trigramas, como estrategia para disminuir el número de nodos y relaciones implicadas con la objetivo de mejorar su compresión. Para ello, en un primer procesamiento de los títulos, se buscan las parejas de palabras con frecuencias de co-ocurrencia extraordinarias, para conseguir esto, se utiliza el test estadístico de detección de outliers (6), definido como:

$$\text{coaparición}(w_i, w_j) > \overline{X} + 3 * S_x \qquad (6)$$

donde $\overline{X}$ es el número medio de co-apariciones y S_x la desviación típica de las co-apariciones. A partir de esta información se calculan todos los posibles bigramas y trigramas., que posteriormente se convertirán en tokens.
- Paralelamente, al segundo procesamiento de los títulos, se calcula el impacto de cada uno de los tokens y el número de co-apariciones conjuntas en los títulos.
- Finalmente el proceso seguido para la generación de los mapas temáticos, es similar al descrito en el apartado anterior. En la Fig. 5 podemos observar los resultados.

VI. Conclusiones y trabajos futuros

Como conclusiones destacar dos, la primera, es que los recursos disponibles en Internet se encuentran aún poco explotados. La causa principal es la falta de integración de las distintas tecnologías implicadas en el proceso. Por ello, en este trabajo hemos desarrollado una arquitectura que soporta todas las etapas necesarias para analizar y visualizar información extraída de Internet. Denominamos a este tipo de arquitectura como *Internet knowledge summarization, analysis and visualization iK-SAV*.

La segunda de las conclusiones, es el potencial que tienen las técnicas de análisis visuales de redes para el desarrollo de aplicaciones interactivas, que faciliten las tareas del usuario final. Como ejemplo citamos: la exploración de foros de discusión, la construcción de crawlers especializados en buscar determinados tópicos, así como para el desarrollo de redes de contactos o la difusión de nuevas innovaciones.

Con respecto a los posibles trabajos futuros, esperamos incluir en la herramienta algoritmos de *clustering*, que permitan la construcción de mapas jerárquicos que faciliten la exploración de las redes. También será necesario realizar estudios con usuarios reales que permitan comparar el modelo interactivo que se propone con otras técnicas de análisis.

VII. Referencias

[1] W. J. Frawley, G. Piatetsky-Shapiro and C.J. Matheus. Knowledge "Discovery in Databases: An Overview," G. Piatetsky-Shapiro and C.J. Matheus. Knowledge Discovery in databases, pages 1-27, MIT press, 1991.

[2] Bent Hetzler and Paul Whitney and Lou Martucci and Jim Thomas. "Multi-faceted Insight Through Interoperable Visual Information Analysis Paradigms," Proceedings {IEEE} Symposium on Information Visualization 1998.

[3] F. de la Rosa T., Rafael M. Gasca, Carmelo Del Valle, Rafael Ceballos: "Arquitectura de un Crawler para Extraer las Estrcturas y Contenidos de Recursos Electrónicos. " JISBD 2002: 259-269.

[4] Thomas Kistler and Hannes Mariais. "WebL – A Programming language for the Web." Computer Networks and IDSN Systems (Procceding of the WWW7 Conference). Volume 30. pages 259-270. Elsevier.1998.

[5] E. Rahm and H.H. Do. "Data Clearing: Problems and Current Approaches." IEE Bulletin of the Technical Commitee on Data Enginnering. 23(4), 2000.

[6] S. R. Bus, And P. N. Yianilos, "A bipartite matching approach to approximate string comparison and search" , NEC Research Institute Technical Report, 1995.

[7] D. Gusfield. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, New York, 1997.

[8] R. S. Boyer and J. S. Moore. A fast string searching algorithm. Communications of the ACM, 20(10):761-772, 1977. 12

[9] S. Wu and U. Manber. Fast text searching allowing errors. Communications of the ACM, 35(10):83-91, 1992.

[10] P. Sellers. The theory and computation of evolutionary distances: pattern recognition. Journal of Algorithms, 1:359-373, 1980.

[11] A Guided Tour to Approximate String Matching (1999) Gonzalo Navarro ACM Computing Surveys

[12] Cleansing Data for Mining and Warehousing (1999) Mong Li Lee, Hongjun Lu, Tok Wang Ling, Yee Teng Ko. Database and ExpertSystems Applications

[13] Derek J. de Solla Price. Little Science, "Big Science." Columbia Univ. Press, New York, 1963.

[14] Henry Small. "Co-citation in the scientific literature: a new measure of the relationship between two documents." Journal of the American Society for Information Sciences 24, pp.265-269, Jul-Aug 1973.

[15] Alan F. Smeaton, Gary Keogh, Cathal Gurrin, Kieran McDonald and Tom Sødring. "Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century ?". SIGIR Forum, Fall 2002, Volume 36 Number 2.

[16] Nascimento, M.A., Sander, J. and Pound, J. Analysis of SIGMOD's Co-Authorship Graph. ACM SIGMOD Record, 32(3). Sep./2003.

[17] José Luis Molina, Juan M. Muñoz Justicia y Miquel Domenech. "Redes de publicaciones científicas. Un análisis de la estructura de coautorías." Revistas Hispano Americana para el Análisis de Redes Sociales. Vol1. 2002.

[18] J. L. Moreno (1934). "Who shall survive?" New York: Beacon Press.

[19] Chen, C. (2003) "Mapping Scientific Frontiers: The Quest for Knowledge Visualization." Springer.

[20] Agarwal, Fankhauser, Gonzalez–Ollala, Hartmann, Hollfelder, Jameson, Klink, Lehti, Ley, Rabbidge, Schwarzkopf, Shrestha, Stojanovic, Studer, Stumme, Walter, Weber: Semantic Methods and Tools for Information Portals. SIP - 33rd Annual Conference of the German Informatics Society, SemIPort Project Symposium, 2003.

[21] The Bibliometric Toolbox, by Terrence A. Brookes -> John P. McLain: Bibliometrics toolbox. JASIS 41(1): 70-71 (1990)

[22] S. A. Morris, C. DeYong, Z. Wu, S. Salman, and D. Yemenu, "DIVA: a visualization system for exploring document databases for technology forecasting," Computers and Industrial Engineering, vol. vol 43, pp. 841-862, 2002.

[23] Chen, C. (2004) Searching for intellectual turning points: Progressive Knowledge Domain Visualization. Proceedings of the National Academy of Sciences of the United States of America (PNAS).

[24] Tomihisa Kamada and Satoru Kawai: "An algorithm for drawing general indirect graphs. " Information Processing Letters 31(1):7-15, 1989

[25] Callon, M., Law, J., and Rip, A. (1986). "Mapping the dynamics of science and technology: Sociology of science in the real world." London: Macmillan.

[26] Callon, M., Courtial, J.P. y Laville, F. "Co-Word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry." Scientometrics, 1991, vol. 22, nº 1, p. 155-205.

[27] Coulter, N., Monarch, I. & Konda, S. (1998). "Software engineering as seen through its research literature: A study in co-word analysis." Journal of the American Society for Information Science, 49(13), 1206-1223.

[28] Ira A. Monarch. "Information Science and Information Systems: Converging or Diverging?", Canadian Association for Information Science Proceedings of the 28th Annual Conference, 2000. http://www.slis.ualberta.ca/cais2000/monarch.htm.

VIII. Biografías

Fernando de la Rosa Ingeniero en Informática por la Universidad de Sevilla. Actualmente profesor en el departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla. Sus investigaciones están fundamentalmente orientadas al resumen, el análisis y la visualización de información extraída de Internet (iK-SAV), vigilancia tecnológica e inteligencia competitiva (VTIC), informetría, y análisis de redes sociales (ARS).

Sergio Pozo Ingeniero en Informática e Ingeniero Técnico en Informática de Sistemas por la Universidad de Sevilla. Actualmente es profesor del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla. Su línea de investigación se centra en Seguridad Informática: encontrar formas de detectar y recuperar sistemas ante fallos de seguridad, perfilado de intrusos utilizando técnicas de engaño y control, así como el modelado de su entorno social.

Rafael M. Gasca Doctor en Informática desde el año 1998. Este título lo obtuvo en la Universidad de Sevilla (España). Desde 1991 es profesor del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla. Actualmente es Profesor Titular de dicho Departamento e investigador responsable de varios proyectos de I+D+I en áreas relacionadas con la aplicación de la programación con restricciones en la resolución de problemas de ingeniería.