REDES- Revista hispana para el análisis de redes sociales
Vol.12,#9, Junio 2007
http://revista-redes.rediris.es

Sistemas de Inteligencia Web basados en Redes Sociales

Fco. Fernando de la Rosa Troyano[1] y Rafael Martínez Gasca[2],
Universidad de Sevilla

Resumen

El Análisis de las Redes Sociales (ARS) es un área que está emergiendo como imprescindible en los procesos de toma de decisiones. Su capacidad para analizar e intervenir una red social puede ser aprovechada para implantar tareas de vigilancia en los sistemas de inteligencia de un centro de investigación o una empresa de base tecnológica. El objetivo de este trabajo es realizar una propuesta para diseñar sistemas de inteligencia web basados en redes sociales. El primer obstáculo para implantar un sistema de estas características es el proceso de recolección de datos. Con objeto de resolver este problema se presenta una metodología para extraer redes sociales. El proceso de extracción se realiza analizando los resultados ofrecidos por los motores de búsqueda. Las consultas realizadas a los motores son construidas en base a direcciones de correo electrónico. A través de la red de extraída también se analiza su distribución espacial, el impacto global de una temática y las relaciones institucionales subyacentes. Como ejemplo concreto se analiza la estructura social de la comunidad que forma la lista de distribución REDES.

Palabras clave: Análisis de Redes Sociales, Co-autorías, Visualización, Extracción de Conocimiento, Internet, Sistema de Inteligencia, Centro de Interés, Relaciones Institucionales, Distribución espacial.

Abstract

Social Network Analysis (SNA) is an emerging area, essential in decision making processes. Its capacities to analyze and intervene in a social network can be used to implant surveillance tasks in research centers or technological-based businesses. The aim of this work is to make a proposal to design intelligence web systems based on social networks. The first obstacle to implant these systems is the data gather process. In order to solve this problem, an extracting social networks methodology is presented. The extraction process is carried out by analyzing the search engine results. Queries are based on electronic mails. From the extracted network, its spatial distribution of social relationships, the global thematic impact and the institutional relationships are also analyzed. The social structure of REDES email distribution list is analyzed as an example.

Key words: Social Network Analysis, Co-authorships, Visualization, Knowledge Discovering, Internet, Systems Intelligence, Institutional Relations, Spatial Distribution.

---

[1] Dpto. de Lenguajes y Sistemas Informáticos. Enviar correspondencia a ffrosat@lsi.us.es En la página http://revista-redes.rediris.es/webredes/google/nogoogle.htm puede visualizar los resultados de los experimentos realizados en el trabajo.

[2] Dpto. de Lenguajes y Sistemas Informáticos. Enviar correspondencia a gasca@lsi.us.es

Introducción

El Análisis de las Redes Sociales (Wasserman 1994; Scout 2000) es un área que está emergiendo como imprescindible en los procesos de toma de decisiones por su capacidad para analizar e intervenir en las redes sociales. Este potencial puede ser aprovechado para implantar tareas de vigilancia en los sistemas de inteligencia (Luhn 1958; Palop 1999; Escorsa 2001) de un centro de investigación o una empresa de base tecnológica. El objetivo de este trabajo es realizar una propuesta para diseñar sistemas de inteligencia Web basados en redes sociales, para ello se analizan los trabajos publicados hasta la fecha y se valida la propuesta realizando numerosos experimentos.

El primer obstáculo que hay que salvar para implantar un sistema de inteligencia de estas características es la recogida de los datos necesarios para construir la red social de interés. Por ejemplo, para obtener redes sociales de terroristas (Valdis 2002; Rodríguez 2004:1; Miralles 2005) o de poder (Mariano & Pizarro 1985; Rodríguez 2004:2; Merino 2006) se analizan noticias de prensa. Otra herramienta fundamental son las encuestas que permiten construir gran variedad de redes, como las redes de transmisión de enfermedades (Liljeros 2001). La construcción de redes también puede estar asistida por bases de datos, así en los estudios bibliométricos se analizan las estructuras de co-autoría (Kretschmer 1994; Chen, 2003) que emergen en las comunidades científicas consultando bases de datos bibliográficas como por ejemplo ISI (Price 1965; Small 1973; Egghe & Rousseau 1990).

Una de las consecuencias que provoca la expansión de la red Internet es que la cantidad de información pública crece exponencialmente. Por si sola esta sobreabundancia limita la utilidad de la información, por *infoxicación* (Wurman 2000), pero también oculta estructuras emergentes. Los procesos que descubren estas estructuras permiten inferir conocimiento y son conocidos como procesos de extracción de información. El análisis de la información extraída permite adquirir o descubrir conocimiento. En trabajos previos se han definido procesos de extracción de redes sociales para diversas fuentes de información: motores de búsquedas (Kautz & Selman, 1997; Matsuo, 2003; Mika, 2004), chats (Mutton, 2004), DBLP (De la Rosa T., 2005:1), archivos FOAF (Mika, 2004, 2005), SourceForge (Crowston & Howison 2004), listas de distribución (McCallum 2004), sitios de contactos (Heer & Boyd, 2005), etc. A lo largo de este trabajo se presenta una novedosa metodología para extraer redes sociales a partir de consultas a motores de

búsquedas. Estas consultas se diseñan especialmente para extraer la red social tras un posterior análisis de las respuestas recibidas del motor de búsqueda.

Entre los sistemas que se apoyan en la Web podemos encontrar el sistema REFERRAL WEB (Kautz & Selman, 1997) que fue diseñado para extraer redes sociales a partir de consultas al motor Altavista, la red obtenida se centra en una persona determinada (red egocéntrica). Para ello sólo necesita conocer el nombre del ego, que denotaremos con la variable ***X***. Mediante un reconocedor de entidades (Named Entity Recognition o NER) el sistema extrae del contenido de los ***k*** primeros documentos indexados por el motor de búsqueda, donde aparece el nombre ***X***, una lista de personas con las que se relaciona. Para medir la relevancia de la relación entre ***X*** e ***Y*** (la variable ***Y*** contiene cualquiera de los nombres de la lista de personas relacionadas con ***X***) utiliza el coeficiente Jaccard (Jaccard, 1901). Este coeficiente es calculado dividiendo el número de páginas indexadas por el motor con los nombre ***X*** e ***Y***, consulta ***"X and Y"***, por el número de páginas que contiene cualquiera de los nombre ***X*** o ***Y***, consulta ***"Y or X"***. El proceso descrito puede repetirse recursivamente con cada uno de los nombres que aparecen relacionados con ***X***, obteniéndose redes con distintas profundidades. En la Figura 1 se puede observar un ejemplo de cómo calcular el coeficiente Jaccard, medida que llamaremos afinidad.

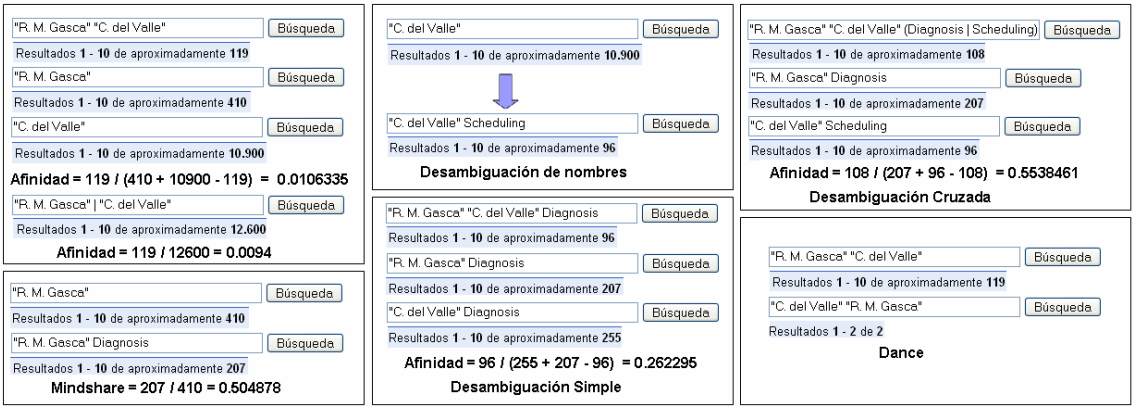


Figura 1. Ejemplos de distintos usos de los resultados ofrecidos por un motor de búsqueda.

Recientemente se han presentado dos sistemas POLYPHONET (Matsuo 2003, 2006) y FLINK (Mika 2004) que generan la red social de una comunidad científica partiendo de una lista con los nombres de sus miembros. El algoritmo básico de ambos sistemas construye la matriz de afinidad realizando consultas para cada pareja ***X*** e ***Y*** (***X*** e ***Y*** son dos nombres distintos de la lista). La relevancia de la relación se calcula utilizando el método propuesto en Kautz & Selman (1997). Matsuo (2006) propone otras medidas alternativas al índice Jaccard: *matching*

coefficient, coeficiente de información mutua, *overlap coefficient*, coeficiente de Dice y el coseno. Ambos sistemas utilizan un umbral de corte para determinar cuándo las relaciones son significativas.

A continuación presentamos los inconvenientes de estos sistemas:

- ***Clasificación de relaciones:*** Los índices de afinidad calculados mediante consultas ***"X and Y"*** son de difícil interpretación. La co-ocurrencia de nombres en las páginas indexadas puede deberse a múltiples factores: publicación conjunta, simple casualidad, participación en un mismo evento (comité de programa), trabajos referenciados en un mismo artículo, etc. En los trabajos de Matsuo (2005) se utiliza el algoritmo C4.5 (Quinlan1993) para obtener reglas que permitan clasificar distintos tipos de relaciones obtenidos por estos índices (co-autoría, miembros de un mismo laboratorio o proyecto, participación en una misma conferencia). Para realizar esta clasificación aplica el algoritmo de aprendizaje a los ***k*** primeros documentos que aparecen en la consulta, ***"X and Y"***, del motor de búsqueda. El problema de aplicar esta técnica es que requiere un corpus de entrenamiento (no siempre disponible) y una apropiada selección de los atributos que deben considerarse en el proceso de aprendizaje.
- ***Nombres ambiguos:*** Al utilizar nombres de personas en las consultas se está agregando ambigüedad a sus resultados, la probabilidad de que el nombre o la firma haga referencia a más de una personas es alta. Para evitar este problema (Matsuo 2006; Mika 2005) utilizan consultas de tipo **"(X and Y) and w"** y **"(Y or X) and w"**, esperando que la introducción de una palabra clave, ***w***, relacionada con el área de la persona permita desambiguar los nombres. La utilización de esta técnica puede producir pérdidas relevantes de información si la palabra clave no está bien seleccionada (Figura 1).
- ***Variedad de nombres:*** Suele suceder que se haga referencia a una misma persona utilizando varios nombres, por ejemplo "Rafael Martínez Gasca" o "R. M. Gasca". Debido a la pérdida de información este problema puede afectar negativamente al cálculo de los índices de afinidad.
- ***Escalabilidad:*** Los sistemas que utilizan listas cerradas de nombres no son escalables. Para conseguir la escalabilidad de estos sistemas Kautz & Selman (1997) y Matsuo (2006) proponen utilizar reconocedores de entidades que extraigan nombres de autores relacionados a partir de las los ***k*** primeros documentos devueltos por el motor de búsqueda para cada persona. A posteriori esta técnica requiere la verificación del significado de las relaciones encontradas haciendo uso de algún índice de afinidad calculado con un motor de

búsqueda. Otro trabajo afín (McCallum 2004) extrae estos nombres pero de sus páginas personales y no realiza ninguna verificación de la significación de la relación.

- ***Complejidad:*** En el peor caso los algoritmos revisados son de $O(|N|^2)$, donde N es el número de miembros de la red, para cada pareja de actores debe realizarse una consulta. Esto es un problema relevante teniendo en cuenta que las licencias de uso que distribuyen los motores de búsquedas pueden tener limitaciones. Por ejemplo Google no permite realizar más de 1000 consultas diarias (para completar una red de 500 actores serían necesario 125 días). La implementación de algoritmos escalables (Matsuo 2006) ha permitido reducir sensiblemente la complejidad respecto del número de consultas necesarias para completar la red. La complejidad de la propuesta de (Matsuo 2006) es de $O(|N|)$ (según datos del autor para 503 actores es necesario 19.852 consultas, 20 días). McCallum (2004) describe un algoritmo de $O(|N|)$ aunque hay que tener en cuenta que las condiciones iniciales son sensiblemente diferentes y por tanto la comparación con los algoritmos revisados no procede.

En este trabajo se propone un algoritmo de extracción de redes sociales que mejora sustancialmente la complejidad de los algoritmos revisados. Además utiliza un modelo de consulta que lo hace robusto ante la ambigüedad y permite una clara interpretación de las relaciones extraídas sin necesidad de utilizar algoritmos de aprendizaje. Otra ventaja que ofrece la metodología presentada en este trabajo es la capacidad de analizar las relaciones institucionales subyacentes en la red y el impacto global de una temática.

De acuerdo con todo ello, el trabajo se distribuye en las siguientes secciones, inicialmente se presenta la notación y el esquema de extracción de redes sociales, para posteriormente concretar el proceso de extracción automática. Previamente se delimita tanto la red de interés sobre la cual se realizan los experimentos como el ámbito de actuación del proceso de extracción. En las siguientes secciones se presenta el concepto de decoración de red social y los experimentos realizados sobre la red de interés: segmentación por dominios, cartografía temática, impacto temático, localización geográfica y análisis de relaciones institucionales. Finaliza este trabajo con las conclusiones, donde además se expondrán los trabajos futuros y algunas aplicaciones prácticas de la metodología. En el Anexo V encontrará un resumen de este trabajo en forma de mapa conceptual.

Notación

A lo largo de este trabajo modelaremos las redes sociales como un *grafo no dirigido* que denotaremos como ***G***. Los nodos de ***G*** se corresponden con los actores de la red y los lazos entre los actores son representados mediante las aristas de ***G***. Denotaremos:

- G_N como el conjunto de nodos de ***G***
- G_E como el conjunto aristas del grafo ***G***
- $n \in G_N$ indica que ***n*** es un nodo del grafo ***G***
- $e \in G_E$ indica que ***e*** es una arista de ***G***. De forma alternativa se representa ***e*** como (n_1, n_2), siendo n_1 y n_2 dos nodos de ***G***.

Tanto los nodos como las aristas podrán disponer de atributos, se denota con n_x y e_x al atributo ***x*** de ***n*** y ***e*** respectivamente, donde $\mathbf{n} \in G_N$ y $\mathbf{e} \in G_E$. A continuación se presentan los atributos utilizados a lo largo de este trabajo:

- n_{hits} acumulador de ***n***
- n_{email} dirección de correo electrónico de ***n***
- n_{domain} dominio de la dirección de correo electrónico de ***n***
- n_{login} cuenta de la dirección de correo electrónico de ***n***
- e_{hits} acumulador de ***e***

Para realizar transformaciones sobre los grafos se definen las primitivas:

- ***Nodo(email, G)***: devuelve el nodo $n \in G_N$ cuyo atributo n_{email} sea igual al parámetro ***email***. Si no existe crea un nodo asignando el valor cero y el valor del parámetro ***email*** a los atributos ***hits*** y ***email*** del nodo respectivamente.
- ***Arista(n1,n2, G)*** devuelve la arista $(n1,n2) \in G_E$, si no existe la crea y asigna el valor cero al atributo ***hits*** de la arista.

Esquema de extracción de una red social

El proceso de extracción de una red social se realiza mediante un procedimiento iterativo que actúa como una bola de nieve. En cada iteración, un proceso expande la red social representada por el grafo ***G***, actuando sólo sobre un subconjunto de nodos, $S \phi G_N$. Los elementos del conjunto finito ***S***, conjunto de nodos semillas, son seleccionados de forma automática o por el usuario. Este proceso se repite ***t*** veces con distintos conjuntos de nodos semillas, $t0$ 7. A continuación se muestra el esquema del procedimiento iterativo:

```
fun ProcesoIterativo(G, t)
  T ← { }
  foreach i in 1 .. t
     S ← SelectorDeSemillas(G, T)
     G ← ProcesoDeExpansión(G, S)
     T ← T ∪ S
  return G
```

En la llamada inicial el parámetro ***G*** del procedimiento iterativo debe contener al menos un elemento, a partir del cual construir la red. Se distinguen dos tipos de procesos de expansión:

- **Procesos extractores:** amplían los límites de la red con nuevos miembros y/o lazos.
- **Procesos decoradores:** no modifican la estructura de la red pero si agregan nueva información.

En los procesos extractores, el número de iteraciones es una variable dependiente del usuario y en los procesos decoradores solo se produce una iteración, ***t=1***. El proceso automático de selección de las semillas se define como:

```
fun SelectorDeSemillas(G, T)
    return G_N - T
```

La explicación de la función ***ProcesoDeExpansión*** se realizará más adelante en combinación con los experimentos realizados en este trabajo.

Ámbito de los experimentos realizados

Para realizar los experimentos que se presentan en este trabajo era necesario disponer de una lista de direcciones de correo. Dicha lista fue cedida por el moderador de la lista de distribución REDES y sólo aparecen los miembros que decidieron participar. Se utilizó la lista para alimentar el proceso iterativo descrito en el apartado anterior. Se extrajeron redes sociales que se plasmaron en mapas, utilizando hasta 5 iteraciones. Por tanto no todos los miembros de la red que aparecen en los mapas pertenecerán a la lista REDES.

Los experimentos realizados se clasifican en dos tipos: 1) los que analizan la red completa; 2) los que se centran en analizar la estructura social de la lista de distribución REDES. Estos segundos experimentos analizan un subconjunto de la red social extraída. La red analizada por los experimentos de tipo 2 está formada por los miembros de la lista REDES y los que están directamente relacionados con ellos. Para evitar confusiones se indicará en el título de la sección que presenta

cada experimento el tipo de red utilizada: red completa (experimentos tipo 1) o red parcial (experimentos tipo 2).

A la hora de interpretar los resultados obtenidos también hay que considerar que:

1) No se pretende analizar el flujo de correos internos de la lista de distribución de REDES, sino las relaciones de colaboración que se producen.

2) Un alto porcentaje de miembros de la lista REDES utilizan direcciones de correo proporcionada por los ISP (*Internet Service Provider*), estas direcciones suelen utilizarse para ofrecer anonimato al usuario y por tanto la metodología propuesta no suele arrojar datos sobre ellos.

Ámbito de la propuesta metodológica

Los procesos propuestos en este trabajo son susceptibles de aplicarse a cualquier sistema KWIC (Key Words in Context) (Luhn 1959). Típicamente estos sistemas indexan un corpus de documentos y permiten realizar consultas sobre el corpus utilizando palabras claves. La característica de los sistemas KWIC es que almacenan las palabras claves junto con su contexto, que suelen ser las ***N*** palabras a la izquierda y a la derecha. Por tanto, cuando se realiza una consulta muestran los documentos que contienen las palabras claves utilizadas en las consultas junto a sus contextos. Esto permite al usuario del sistema evaluar la importancia del documento sin necesidad de explorarlo. Con objeto de poder comparar los resultados obtenidos con otras propuestas de extracción de redes sociales a través de la Web en este trabajo se utiliza como sistema de referencia el motor de búsqueda Google. Este motor está compuesto por un conjunto de agentes de búsquedas o crawlers que rastrean la Web e indexan los documentos que encuentran a su paso. Posteriormente mediante el uso de palabras claves permite consultar los documentos indexados junto a los contextos de las palabras claves (Figura 2).

Figura 2. Ejemplo de consulta al motor de búsqueda Google. En negrita las palabras claves utilizadas y alrededor su contexto.

Extracción automática de redes sociales

A diferencia de propuestas anteriores destacar como aportación de este trabajo el uso de direcciones de correo electrónico, como palabras claves de las consultas, en vez de los nombres de autores. Si bien es cierto que una persona puede tener más de una dirección de correo (por ejemplo, la personal y la institucional), también es cierto que una dirección de correo suele identificar a una única persona. Esto permite eliminar los problemas de ambigüedad que presentan otras metodologías, aunque no los problemas de variedad.

Como se expone en el apartado anterior el proceso de extracción de la red es iterativo. En cada iteración se realiza una consulta al motor de búsqueda con cada una de las direcciones de los nodos del conjunto de semillas, ***S***. Estas direcciones de correo serán denominadas direcciones semillas y se componen de una cuenta y un dominio. Por ejemplo para la dirección ***ffrosat@lsi.us.es*** la cuenta será ***ffrosat*** y el dominio será ***lsi.us.es***. El esquema de la consulta por cada dirección semilla es:

"cuenta" "dominio" filetype:pdf

La consulta se compone de dos partes: la parte 1) ***"cuenta" "dominio"*** fuerza al sistema a buscar los contextos en que aparece la dirección de correo 2) y la parte ***filetype:pdf***, restringe la búsqueda a los documentos pdf (en la Figura 2 se muestra un ejemplo). Una vez obtenido el resultado de la consulta se analizan los contextos mediante expresiones regulares. Se analizan sólo los contextos donde aparece la dirección semilla. Cada nueva dirección de correo que aparezca en el contexto formará parte de la red social como un nuevo nodo. Y se añadirá un lazo entre la dirección semilla y la nueva dirección. Los nodos y los lazos están

asociados con un contador que representará su importancia dentro de la red. Si alguna dirección vuelve a repetirse en algún contexto los contadores de los nodos y los lazos correspondientes se incrementarán en una unidad. Podemos ver un ejemplo en la Figura 3A.

Figura 3. Transformación de los contextos de la consulta mostrada en la Figura 3 en una red social. Los nodos rojos son los actores añadidos al sociograma y los azules los nodos semillas. Los cuadrados celestes representan el contador asociado.

En los contextos pueden aparecer algunas expresiones de correo especiales como por ejemplo:

{ceballos,ffrosat,gasca}@lsi.us.es

En estos casos especiales la dirección de correo forma un cluster donde todos se relacionan con todos, podemos ver un ejemplo de la transformación en Figura 3B.

Mediante estas transformaciones podemos expandir la red social como si de una bola de nieve se tratase. Si una dirección era semilla en la iteración i deja de serlo para la iteración i+1 y si una dirección aparece por primera vez en la iteración i pasará a ser semilla en la iteración i+1. Por tanto para iniciar el proceso bastará con disponer de un conjunto finito de direcciones de correos semillas. El número de iteraciones dependerá de los recursos disponibles por el usuario y por sus preferencias, a modo orientativo, con tres o cuatro iteraciones se consiguen redes muy completas (Figura 4).

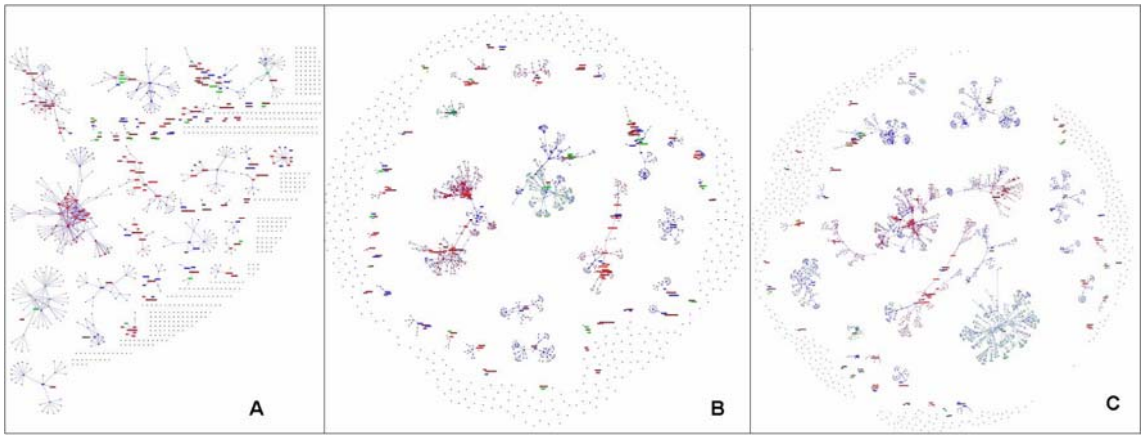


Figura 4. Resultados obtenidos con 3, 4 y 5 iteraciones utilizando las direcciones facilitadas por el moderador de la lista de distribución REDES. En el Anexo I se amplía la figura 4A.

El proceso de extración de redes sociales descrito se formaliza como:

fun ProcesoDeExpansión(G, S)
  foreach s ∈ S do
    <L, D> ← <s_{login}, s_{domain}>
    contexts ← ExtractContexts("L D filetype:pdf")
    foreach c ∈ contexts
      emails ← ExtractEmails(c)
      if (s_{email} ∈ emails)
        if (IsCluster(c))
          foreach <em_1, em_2> ∈ { emails **x** emails } | em_1 < em_2 do
            <n_1, n_2, e> ← <Nodo(em_1, G), Nodo(em_2, G), Arista(n_1, n_2, G) >
            <$n_{1\ hits}$, $n_{2\ hits}$, e_{hits}> ← < $n_{1\ hits}$ + 1, $n_{2\ hits}$ + 1, e_{hits} + 1>
        else
          foreach em ∈ {emails – s_{email}} do
            <n, e> ← <Nodo(em, G), Arista(n, s, G)>
            <s_{hits}, n_{hits}, e_{hits}>← <s_{hits} + 1, n_{hits} + 1, e_{hits} + 1>

Donde el operador ***{emails X emails}*** representa el producto cartesiano de los elementos del conjunto ***emails*** y el operador ***em_1 < em_2*** establece una relación de orden entre dichos elementos. También se definen las siguientes primitivas:

- ***ExtractContexts("L D filetype:pdf")***: devuelve un conjunto con los contextos extraídos de una consulta de tipo ***"cuenta" "dominio" filetype:pdf***.
  - ***ExtractEmails(c)***: devuelve un conjunto con las direcciones de correo electrónico que encuentra en el contexto **c**.
  - ***IsCluster(c):*** tras analizar el contexto ***c*** devuelve ***verdadero*** si las direcciones de correo extraídas de ***c*** forman parte de un cluster y ***falso*** en caso contrario.

Decorado de redes sociales

Una vez construida la red social esta puede ser decorada. Entendiendo por decorar una red al proceso de enriquecer la información de los distintos elementos que la componen (nodos y aristas) mediante el cruce de datos. Seguidamente se presentan los experimentos realizados, así como los resultados obtenidos. Los experimentos se distribuyen en las siguientes secciones: segmentación por dominios, cartografía temática, localización geográfica y análisis institucional. A lo largo de estas secciones se describen tres índices (***frecuencia de ocurrencia***, ***impacto*** y ***mindshare***) que miden la relevancia de una determinada temática para un miembro de la red. Basándonos en estos índices se propone el concepto de ***impacto temático global***. El impacto temático global es una medida que permite analizar el grado de relevancia de una determinada temática en la red completa. También se discute a lo largo de estas secciones las características destacables de la metodología para analizar: 1) no sólo las relaciones de colaboración entre los miembros de una lista de distribución sino las relaciones de colaboración con miembros ajenos a la lista de distribución; 2) las relaciones institucionales que subyacen a una red.

Segmentación por dominios (red completa)

Una forma de representar en los mapas la información es mediante el coloreado de sus nodos. Por ejemplo, analizando los dominios asociados a los nodos, es posible:

1) Asignar un color a cada dominio. De esta forma se identifica cada nodo con su organización y se obtiene una visualización de la red segmentada en las distintas instituciones que la componen.

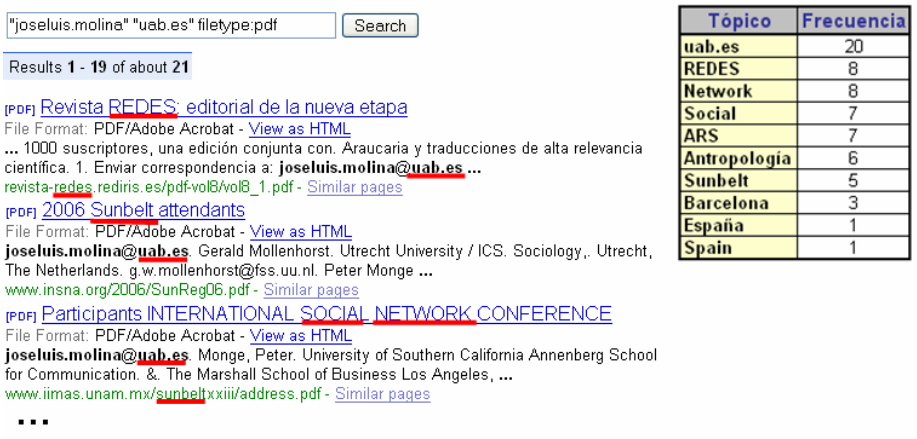
2) Asignar a los actores con dominios "*.es" el color rojo (actores españoles), los que no tienen dominios "*.es" de azul (miembros con características de internacionales), los "*.com", "*.org" y "*.net" de color verde (característica indeterminada) y finalmente representar a los miembros de la lista REDES en color sepia. De esta forma podemos caracterizar las redes sociales de integran a los miembros de REDES, Anexo I.

Cartografía temática de redes sociales (red completa)

Un problema que abordan los trabajos revisados es asociar a cada autor un tópico. En el trabajo de (McCallum 2004) se utiliza la ganancia de información de palabras extraídas de las páginas personales. El sistema FLINK utiliza una lista cerrada de tópicos y asocia los tópicos a los actores dependiendo del mindshare obtenido, número de documentos encontrados por la consulta ***"T and X"*** dividido por el

número de documentos encontrados por la consulta ***"X"*** (Figura 1), donde ***T*** es una palabra clave que se identifica con un tópico. En el caso del sistema POLYPHONET analiza las ***K*** primeras páginas devueltas por el motor de búsqueda, extrae de ellas un conjunto de palabras claves y asocia al actor las ***K'*** primeras palabras que más co-ocurren en las páginas analizadas.

Como alternativa a estos trabajos se propone otro índice diferente para asociar la relevancia de un tópico ***T*** a un actor de la red con dirección de correo ***cuenta@dominio***. Se define como ***frecuencia de ocurrencia*** de ***T*** al número de ocurrencias de ***T*** en el resultado de la consulta ***"cuenta" "dominio" filetype:pdf***. En la Figura 5 se muestra un ejemplo.



Tópico	Frecuencia
uab.es	20
REDES	8
Network	8
Social	7
ARS	7
Antropologia	6
Sunbelt	5
Barcelona	3
España	1
Spain	1

Figura 5. Resultado parcial de una consulta, donde están subrrayados en rojo algunos de los tópicos analizados. La tabla muestra la frecuencia de ocurrencia de los tópicos analizados en el resultado de la consulta.

A partir de este índice podemos decorar un mapa construido previamente por el proceso de extracción descrito, con la relevancia del tópico ***T*** en la red. Cada tópico se identificará con un color y la relevancia del tópico para un actor (***frecuencia de ocurrencia***) se reflejará mediante la intensidad del color.

Este índice tiene la ventaja de que permite representar la distribución de una gran variedad de tópicos aprovechando las consultas realizadas por el proceso de extracción de la red. Y como inconveniente que sólo puede representar los tópicos que aparecen en el contenido de los contextos de dichas consultas. Seguidamente describimos algunos de los experimentos realizados:

- Localizar los miembros de la red que pertenecen a un departamento, utilizando palabras claves como "lsi" (Lenguajes y Sistemas Informáticos), "sociología", "biblioteconomía", "documentación", etc. Anexo II y IV.
- Localizar los miembros de la red que participaron en alguna de las ediciones del congreso "sunbelt".

- Localizar documentos específicos, como tutoriales, sobre herramientas como “Ucinet” y “Pajek”.
- Localizar geográficamente a los miembros de la red utilizando palabras claves como “Sevilla”, “Andalucía”, “Extremadura”, “España”, etc.

Impacto temático en redes sociales (red completa)

Dado el inconveniente que tiene el índice anterior para representar tópicos, se propone como índice temático alternativo el ***mindshare*** de un actor. El concepto de *mindshare* se ha adaptado para los fines de este trabajo y se define como el número de documentos encontrados para la consulta ***“T” “cuenta” “dominio” filetype:pdf*** dividido por el número de documentos encontrados para la consulta ***“cuenta” “dominio” filetype:pdf*** (Figura 6A). En los experimentos realizados se ha detectado un caso particular donde la información aportada por este índice puede ser poco representativa (número de documentos encontrados para la consulta ***“cuenta” “dominio” filetype:pdf*** es igual a 1 y este es igual al número de documentos encontrados para la consulta ***“T” “cuenta” “dominio” filetype:pdf***, Figura 6B). Como alternativa y para evitar este caso anómalo se puede utilizar el número de documentos obtenido por la consulta ***“T” “cuenta” “dominio” filetype:pdf***, que llamaremos ***impacto***.

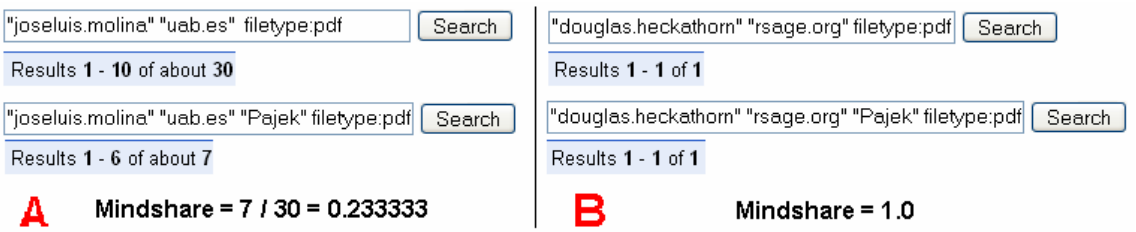


Figura 6. Ejemplo de cáculo de índice temático mindshare

Los dos índices anteriores (*mindshare* e impacto) tienen el inconveniente de necesitar una consulta extra por cada actor y tópico analizado. A cambio se obtienen las siguientes ventajas:

- Permiten explorar tópicos no recogidos en las consultas.
- No se restringen a los contextos obtenidos en las consultas, obteniendo resultados más precisos ya que se explora el contenido de los documentos mediante el sistema de indexación del motor de búsqueda.

Con objeto de estudiar el impacto temático de cuatro herramientas utilizadas para analizar redes sociales ("Ucinet", "Pajek", "Siena" y "Jung") se desarrolla el concepto de ***impacto temático global***. El impacto temático global es una medida que permite conocer la relevancia que tiene una determinada temática o centro de interés en una red de individuos. El impacto temático global se representa mediante un índice que se construye a partir de los índices temáticos (*mindshare* o impacto). Como prerrequisito a este análisis es necesario crear un mapa temático para cada tópico, el proceso es el siguiente:

- Se construye la red de interés con el proceso de extracción.
- Para cada actor de la red se calcula el impacto temático (*mindshare* ó impacto) del tópico. Esto implica que el sistema debe realizar una consulta extra para cada término.
- Se representa con un mismo color a los actores con impacto temático mayor que cero, la intensidad del color dependerá de la relevancia del tópico para el actor en cuestión. En el Anexo IV se puede observar el mapa temático para el término "Pajek".

A partir del mapa temático se calcula el *impacto temático global* de la red como la suma de los *impactos temáticos* (*mindshare* o impacto) de cada individuo:

$$\sum_{n \in G} n_{mindshare} \qquad \sum_{n \in G} n_{impacto}$$

Un aspecto importante que aporta este trabajo es el de comprobar la semántica de la información analizada. Para ello y para cada mapa generado se analizan los contextos que rodean cada tópico explorado en las consultas extras realizadas (Figura 7). De esta forma se comprobó que mientras los mapas de Pajek y Ucinet representaban la información requerida, los mapas sobre Siena y Jung no lo hacían. La información representada sobre el tópico Siena se refería mayoritariamente a Siena como ciudad o universidad y la información representada sobre el tópico Jung se refería a apellidos de investigadores, por tanto fueron desechados.

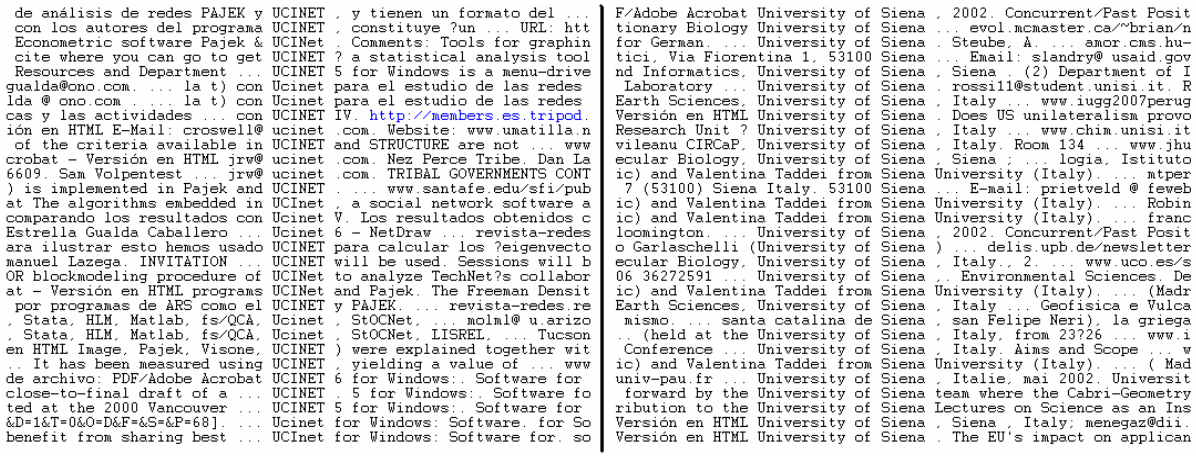


Figura 7. Contextos de Ucinet y Siena

Del análisis final de los *impactos temáticos globales* para los tópicos Siena y Pajek, Tabla 1, se concluye que el impacto de Ucinet en la red es un 20% mayor que el de la herramienta Pajek. Probablemente el componente diferenciador del impacto de ambas herramientas se deba a sus características de usabilidad.

Herramienta	Impacto	MindShare
Pajek	133	747
Ucinet	178	890
Siena		
Jung		

Taba 1. Impacto temático global para cuatro herramientas utilizadas para analizar redes sociales

Localización geográfica de las redes sociales (red parcial)

La localización geográfica de las redes sociales es un área de interés reciente, dentro del Análisis de Redes Sociales. En (Molina *et al* 2005) se hace uso de una encuesta mediante EgoNet para construir la red de contacto de los miembros de la lista REDES y localizar geográficamente cada uno de sus miembros. En el ámbito de la web semántica el sistema FLINK determina las coordenadas geográficas de los miembros de la red explotando la información que aparece en los ficheros FOAF y utilizando el servicio ESRI Place Zinder Simple Web Service.

En la propuesta realizada en este trabajo se utilizan dos procedimientos para localizar geográficamente un miembro de la red a partir de su dirección de correo. El primero de ellos utiliza el dominio principal de la cuenta de correo (.es, .ar, .us, etc) para identificar el *país* de procedencia. Basta con tener una base de datos que asocie los dominios de cada país con su localización geográfica, esta información puede extraerse fácilmente del Think Fat Book. Está técnica tiene dos problemas:

- Sólo permite localizar geográficamente el país de procedencia del actor.
- No es adecuada para dominios especiales como .org, .net, .com o .edu. que no están asociados a ningún país.

Para superar estas carencias hay que recurrir a la integración de servicios. En este trabajo hemos integrado los servicios ofrecidos por Hostip y GeoIpTool disponibles en Internet. Estos servicios permiten utilizar una dirección HTTP para consultar la localización geográfica de una IP. El procesamiento de los resultados permite decorar la red con la localización geográfica de sus miembros. Para obtener la IP de una dirección de correo se utiliza el servicio DNS (Domain Name Server). Debemos destacar que se ha localizado geográficamente del 90% de países y 80% de ciudades de los miembros de la red. La mayoría de los fallos son provocados por direcciones de correo adscritas a ISP (Internet Service Provider) como Gmail, Hotmail o Yahoo; y en menor medida influye las direcciones de correos de universidades virtuales como la UOC a cuyos miembros no se les pueden asignar ningún lugar específico. En la Figuras 8, 9 y 10 presentamos los resultados obtenidos.

Figura 8. Localización geográfica de los países de procedencia de los miembros de la red parcial

Figura 9. Localización geográfica de las ciudades de procedencia de los miembros de la red parcial

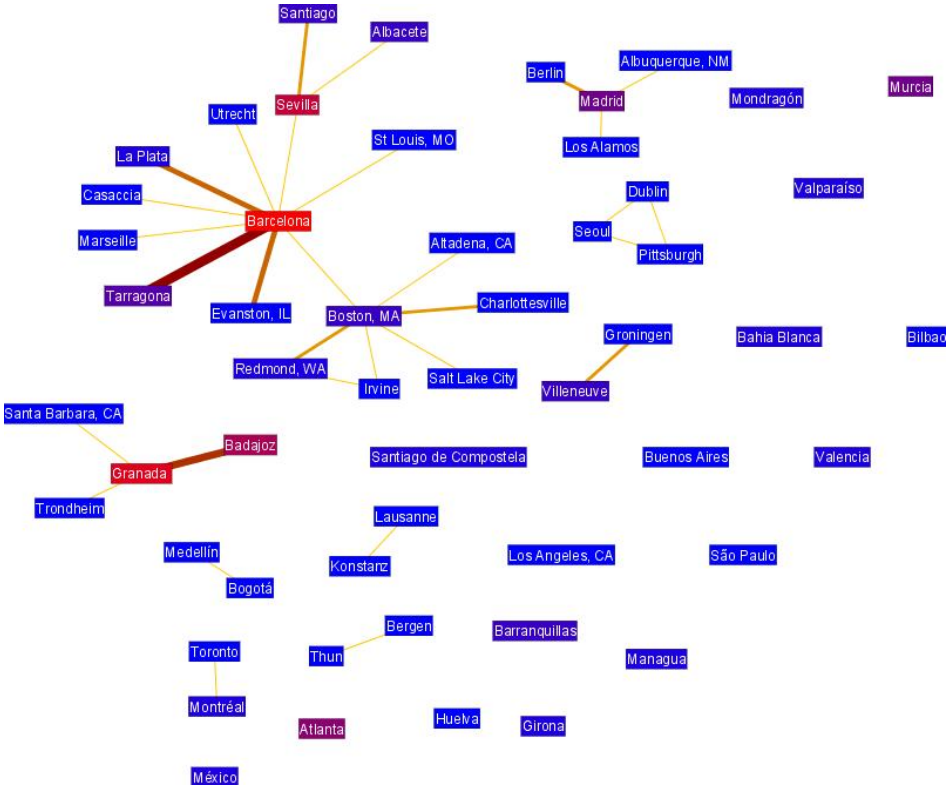


Figura 10. Relaciones entre las ciudades de procedencia de los miembros de la red parcial

Del análisis de estas figuras destacamos:

- El papel integrador de España y la centralidad de la ciudad de Barcelona en la red.

- Influencia espacial: La concentración de las relaciones en lugares geográficamente cercanos como ciudades o países (el 90% de la relaciones se producen entre miembros de un mismo país y el 80% entre miembros de una misma ciudad).
- La influencia de la cultura y la economía en la construcción de las redes personales. Escasez de relaciones entre los hemisferios norte y sur. Influencia de la lengua en la construcción de las redes personales.

Análisis de relaciones institucionales (red parcial)

Las redes personales traspasan las fronteras institucionales (países u organizaciones) permitiendo analizar las relaciones institucionales a través de ellas. El conocimiento de las instituciones a las cuales pertenece cada miembro de la red es un prerrequisito para este análisis. La utilización de listas de nombres no permite identificar esta información sin la utilización de técnicas de aprendizaje. No obstante la disponibilidad de las direcciones de correo de sus miembros permite obtener dicha información con un bajo coste de procesamiento. Aprovechando la disposición jerárquica de los dominios y resumiendo esta información se consigue emerger la red de países y organizaciones implicadas entre las relaciones personales, Figura 11 y 12.

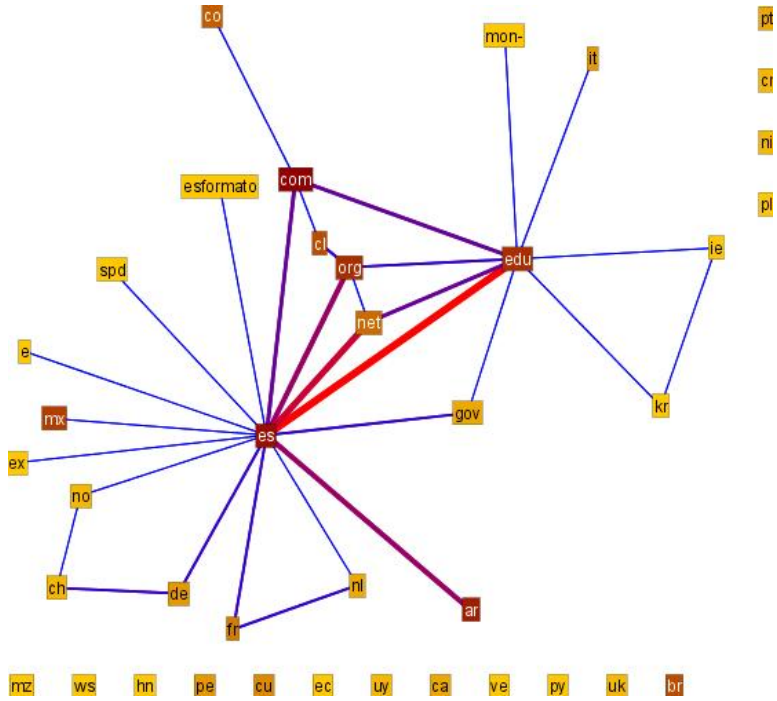


Figura 11. Relaciones entre los países de procedencia de los miembros del semgento 1

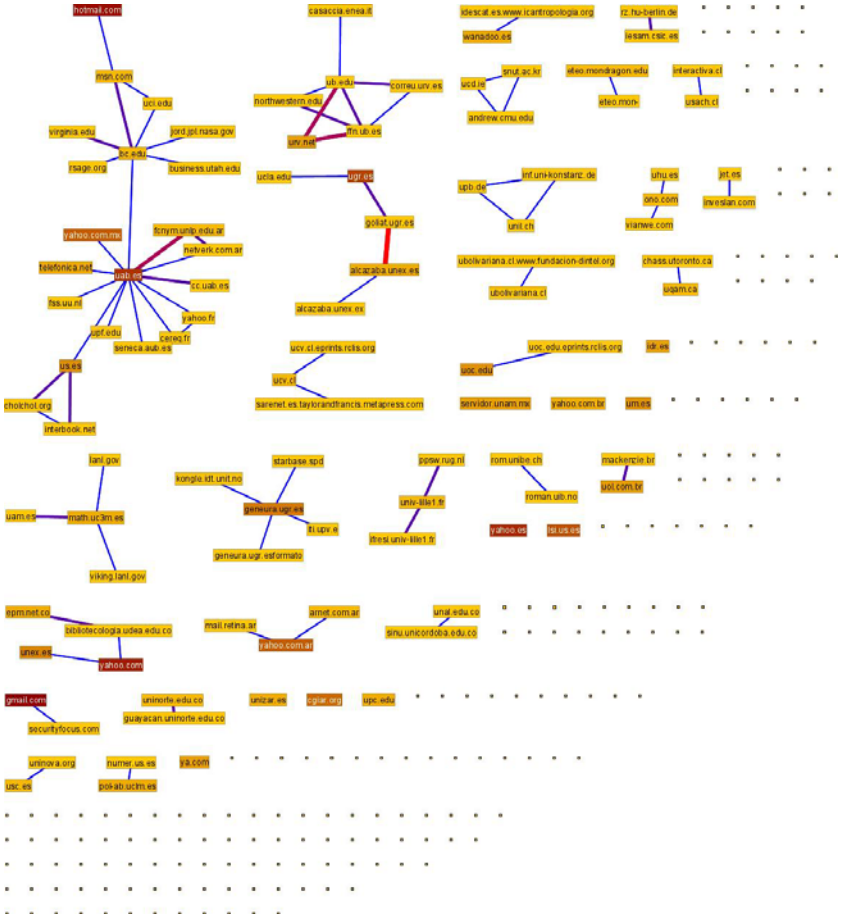


Figura 12. Relaciones entre las organizaciones que agrupan a los miembros de la red parcial

En los experimentos realizados con esta técnica podemos apreciar cómo el dominio .es, que hace referencia a miembros de la red localizados en España, emerge como eje central de la red. Otro polo de la red es el dominio .edu. Destacar que los dominios .com y .net suelen pertenecer a cuentas de correos gratuitas, por tanto no es recomendable asociarlas a ningún país.

Podemos profundizar en estas relaciones analizando la red a nivel de organizaciones. Es posible apreciar en el ámbito de la red analizada la fuerte intensidad de los lazos entre Universidad de Granada (ugr.es) y la Universidad de Extremadura (unex.es) así como los lazos entre la Universidad de Barcelona (ub.es) y la de Tarragona (urv.es). También destacable el alto grado de internacionalización de los lazos de la Universidad Autónoma de Barcelona (uba.es) con universidades de Argentina, Francia y Estados Unidos.

Conclusiones

En comparación con los trabajos revisados esta nueva metodología aporta los siguientes beneficios:

- Disminuye sensiblemente los problemas de ambigüedad inherente a las técnicas utilizadas en los trabajos publicados hasta la fecha.
- El coste computacional del algoritmo es lineal respecto al número de actores de la red y mejora sensiblemente la constante multiplicativa de la mejor propuesta realizada hasta el momento (Matsuo 2006) (según los datos aportados por el autor para 503 actores necesita 19.852 consultas, con esta propuesta serían necesarias 503 consultas).
- A diferencia de las propuestas realizadas hasta el momento, donde la confusión del significado de las relaciones extraídas es patente, esta propuesta permite una interpretación clara de los lazos que unen los actores de la red.
- Permite analizar las relaciones institucionales que emergen de la red de contactos.

De los experimentos realizados se constata que las redes obtenidas están influidas por los estilos de publicación, ya que estos afectan a la información contextual ofrecida por los motores de búsqueda. Aún siendo esto cierto, es de esperar que la redundancia de información minimice el impacto de este problema. También se constata que la calidad de los mapas temáticos depende de los términos utilizados. La ambigüedad en los términos puede influir negativamente en la información obtenida (por ejemplo la palabra “java” puede ser un lenguaje de programación pero también puede interpretarse como un café o una isla). Como se puede observar en numerosos estudios, el problema de la ambigüedad, es un denominador común de difícil solución. Debido a este problema, para generar mapas temáticos es imprescindible habilitar algún proceso que verifique la validez de los datos, como se ha realizado en esta metodología.

En este trabajo no sólo se defiende un proceso de extracción de redes sociales, sino que también se presentan varios experimentos enfocados a automatizar las tareas de inteligencia competitiva que puede tener un centro de investigación o una empresa de base tecnológica. La metodología propuesta complementa otras metodologías que restringen las tareas de inteligencia competitiva al análisis textual de documentos (Callon 1986, 1991; Coulter 1998). Es importante destacar

el bajo coste en recursos necesarios para implantar esta metodología. Como ejemplos de tareas de inteligencia que pueden realizarse, destacamos:

- La búsqueda de expertos.
- La determinación de colegios invisibles (Price 1963).
- El análisis de tendencias: Los mapas temáticos de la red social son una herramienta novedosa que permite analizar los centros de interés de la red social, permitiendo obtener respuestas a cuestiones bastante interesantes como qué tecnologías o técnicas de análisis tienen más impacto en la red.
- El análisis de relaciones institucionales.

En trabajos previos (De la Rosa 2005:2) los autores han constatado la importancia de implantar una fase de clareado de datos (*data clearing*) para eliminar las posibles variedades (direcciones de correo o nombres de personas asociadas a un mismo particular). También se plantea como trabajo futuro la definición de un proceso automático de extracción de redes sociales enfocado a una determinada temática y la generación de redes léxicas a partir de la información disponible.

Entre las aplicaciones prácticas de este trabajo destaca el uso de esta metodología como alternativa a los servicios de búsqueda de documentos mediante palabras claves ofrecidos por los motores de búsqueda. Aportando una perspectiva social a la búsqueda de información a través de una interfaz entre un usuario y un motor de búsqueda.

Bibliografía

Callon, M.; Law, J.; and Rip, A. (1986). *Mapping the dynamics of science and technology: Sociology of science in the real world*. London: MacMillan.

Callon, M.; Courtial, J.P.; and Laville, F. (1991). "Co-Word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry", *Scientometrics* 22(1): 155-205.

Chaomei Chen (2003). *Mapping scientific frontiers: The quest for knowledge visualization.* London: Springer-Verlag. ISBN: 1-85233-494-0.

CIA -The World Factbook
<https://www.cia.gov/cia/publications/factbook/index.html>

Coulter, N.; Monarch, I. and 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.

Culotta, A.; Bekkerman, R. and McCallum A. (2004). “Extracting Social Networks and Contact Information from Email and the Web”, CEAS. Conference on Email and Spam.

Crowston, K. and Howison, J. (2004). “The social structure of free and open source software development”.

De la Rosa T., F.; Pozo, S. and Gasca, R. M. (2005). “Análisis y Visualización de Comunidades Científicas con Información Extraída de la Web”, *IEEE Latin America Transactions*. ISSN: 1548-0992.

De la Rosa T., F.; Gómez-López, M. T. and Gasca, R. M. (2005). ”Analysis and Visualization of the DX Community with Information Extracted from the Web”, *DEXA. 16th International Conference on Database and Expert Systems Applications.* LNCS 3588. 726-735. ISSN: 0302-9743

De la Rosa T. (2007). “elgooG - Google en el espejo”, Web Redes <http://revista-redes.rediris.es/webredes/google/nogoogle.htm>.

Price, D. J. (1963). *Little Science, Big Science.* New York: Columbia Univ. Press.

Price, D. J. (1965). “Networks of Scientific Papers”, *Science* 149, 510-515.

Egghe, L. and Rousseau, R. (1990). *Introduction to Informetrics.* Eleservier Science Publihers. ISBN: 0-444-88493-9.

Escorsa, P. and Maspon, R. (2001). *De la vigilancia tecnológica a la inteligencia competitiva*. Madrid: Prentice Hall. ISBN: 84-2995-3057-3.

GeoIpTool http://www.geoiptool.com/

Heer, J. and Boyd, D. (2005). “Vizster: visualizing online social networks”, In *Proceedings of the 2005 IEEE Symposium on Information Visualization*: 33–40 <Hostip http://www.hostip.info/use.html>.

Jaccard P (1901). “Étude comparative de la distribution florale dans une portion des Alpes et des Jura”, *Bull Soc Vaudoise Sci Nat* 37:547-579.

Kautz, H.; Selman, B. and Shah, M. (1997). “The hidden Web”, *AI magazine* 18(2):27–35

Kretschmer, H. (1994). “Coauthorship networks of invisible-colleges and institutionalized. Communities”, *Scientometrics* 30(1): 363-369

Liljeros, F.; Edling, C. R.; Amaral, L. A. N.; Stanley, H. E. and Aberg, Y. (2001). "The web of human sexual contacts", *Nature*, 411(Jun)

Luhn, H. P. (1958). “A Business Intelligence System”, *IBM Journal*, October

Luhn, H. P. (1959). “Keyword-in-Context Index for Technical Literature (KWIC Index)”, Yorktown Heights, N. Y.: IBM

Mariano, B. and Pizarro, N. (1985) “The Structure of the Spanish Political Elite, 1939-1975”, en Gween Moore, (ed.), *Research in Politics and Society*, 1, JAI Press, Inc

Matsuo, Y.; Tomobe, H.; Hasida, K. and Ishizuka, M. (2003). “Mining Social Network of Conference Participants from the Web”, In Proceedings of *the International Conference on Web Intelligence*, 190-194.

Matsuo, Y.; Mori, J.; Hamasaki, M; Takeda, H.; Nishimura, T.; Hashida, K. and Ishizuka, M. (2006). “Polyphonet: An advanced social network extraction system”, In *Proceedings of WWW2006*.

Merino J. (2006). “La asimetría de la información en las Organizaciones: Una propuesta metodológica desde el Análisis de Redes Sociales (ARS)”, *III Congreso online – Observatorio para la Cibersociedad.*

Mika, P. (2004). "Bootstrapping the FOAF-Web: An Experiment in Social Network Mining", Proc. 1st Workshop Friend of a Friend, *Social Networking and the Semantic Web.*

Mika, P. (2005). “Flink: Semantic web technology for the extraction and analysis of social networks”, *Journal of Web Semantics*, 3(2).

Miralles, J. (2005). “Internet, Análisis de Redes, Guerras de Cuarta Generación y Al Qaeda”, Revista Mundo Linux.

Molina, J. L.; Muñoz Justicia, J. M. and Doménech, M. (2002). “Redes de publicaciones científicas. Un análisis de la estructura de coautorías”, Redes. Revista Hispana para el Análisis de Redes Sociales, 1(3).

Molina, J. L.; Ruiz, A.; and Teves, L. (2005). “Localizando geográficamente las redes personales”, Redes. Revista Hispana para el Análisis de Redes Sociales, 8(5).

Mutton, P. (2004). “Inferring and Visualizing Social Networks on Internet Relay Chat”, InfoVis, Austin, TX, 35-43.

Palop, F. and Vicente, J. M. (1999). "Vigilancia Tecnológica e Inteligencia Competitiva: Su potencial para la empresa española”, Madrid: Fundación Cotec. Estudio Cotec nº15.

Quinlan, J. R. (1993). “C4.5: Programs for Machine Learning”, Morgan Kauffman.

Rodríguez, A. J. (2004). “La red terrorista del 11M”, VIII Congreso Español de Sociología. Alicante, 23-25 Sep.

Rodríguez, J. A.; Cárdenas, J. and Oltra, C. (2004). “Networks of economic power in Europe”, XXIV International Sunbelt Social Network Conference. Portoroz, Slovenia, May: 12–16

Scott, J. P. (2000). “Social Network Analysis: A Handbook”, Second edition. Sage Publications.

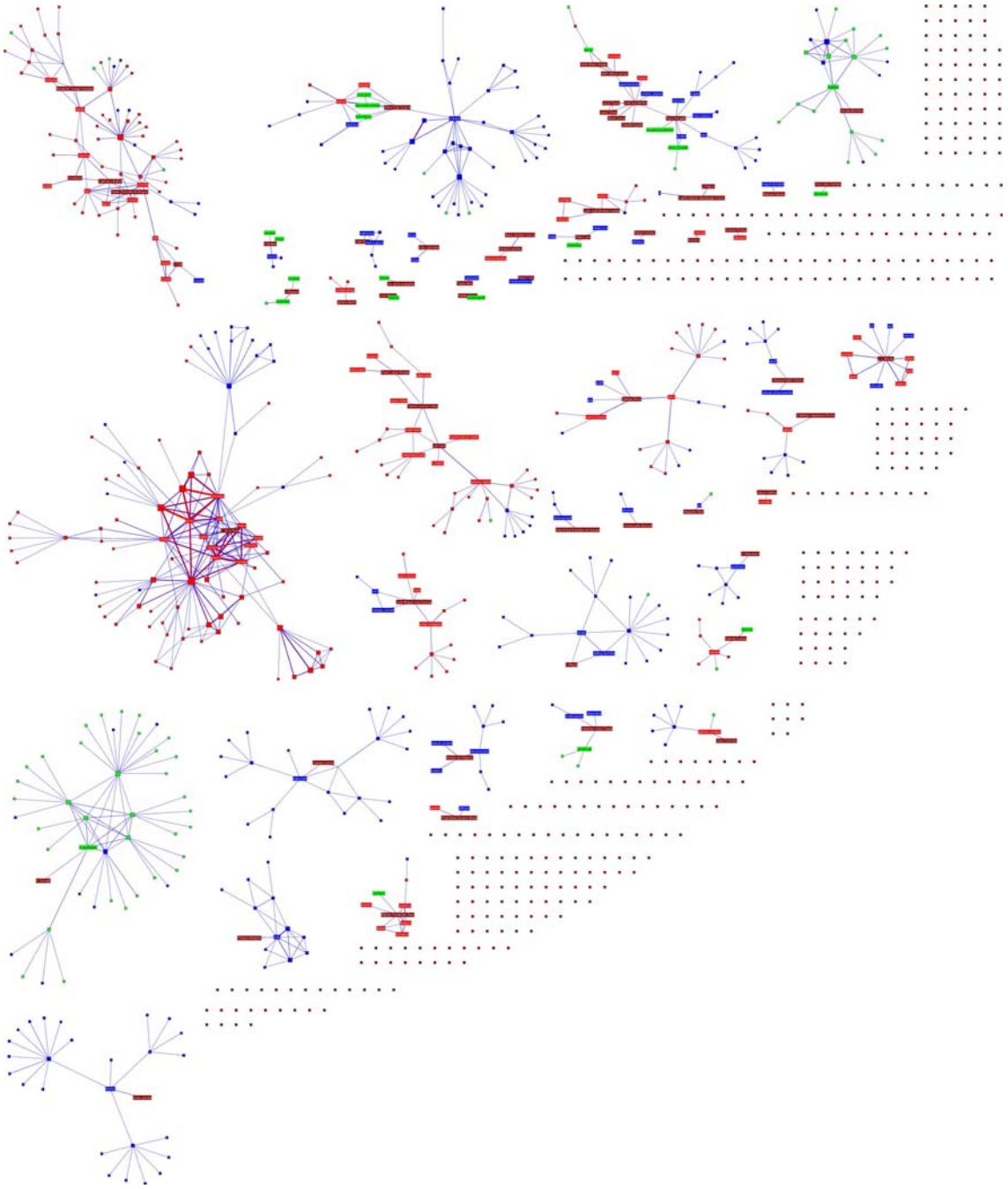
Small, H. (1973). “Co-citation in the scientific literature: a new measure of the relationship between two documents”, Journal of the American Society for Information Sciences 24(Jul-Aug): 265-269.

Valdis, K. (2002). “Mapping Networks of Terrorist Cells”, Connections 24(3): 43-52

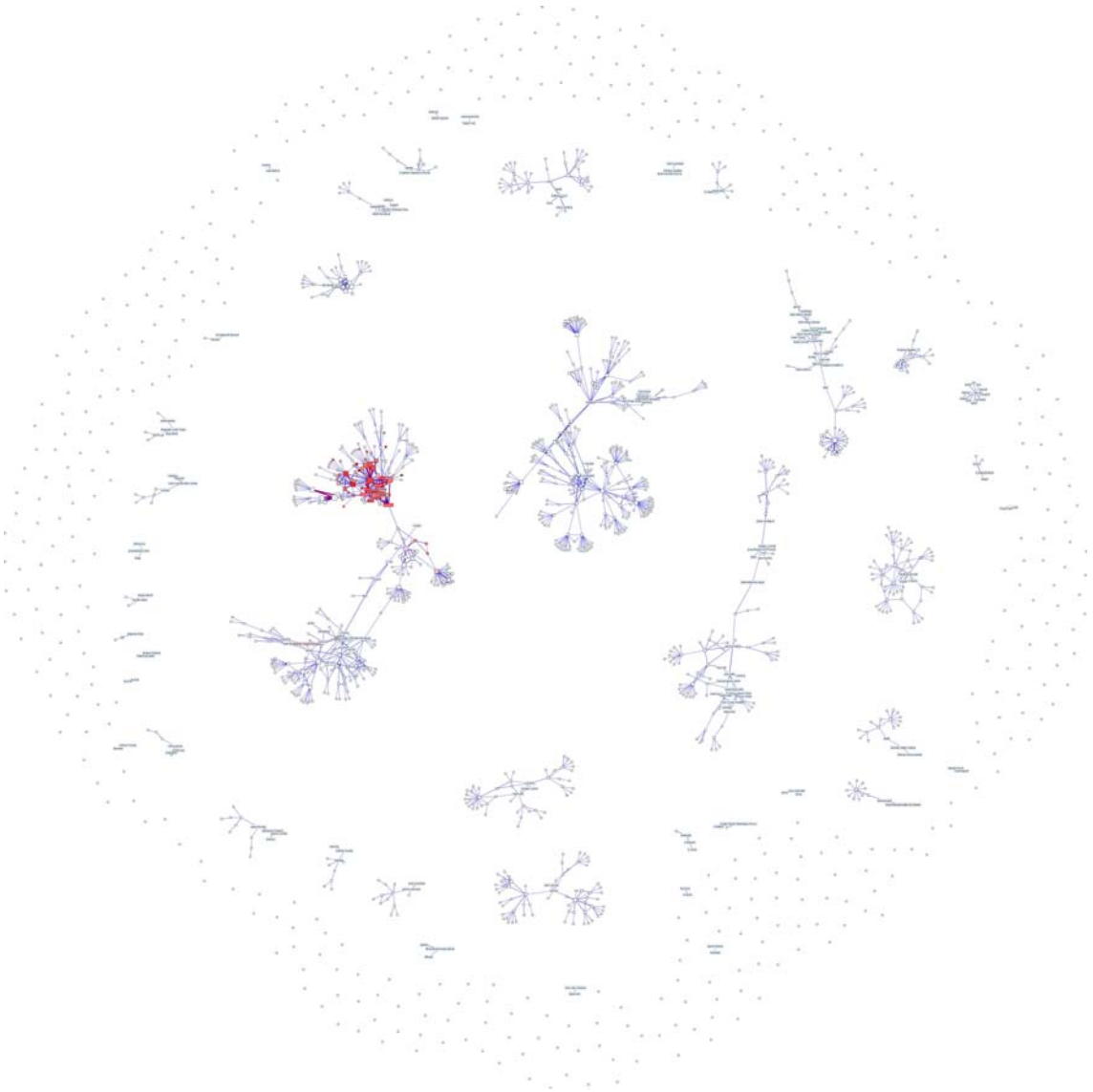
Wasserman, S. and Faust, K. (1994). “Social network analysis. Methods and Applications”, Cambridge University Press, Cambridge.

Wurman, R.S. (2000) “Information Anxiety”, ISBN: 0-78972410-3.

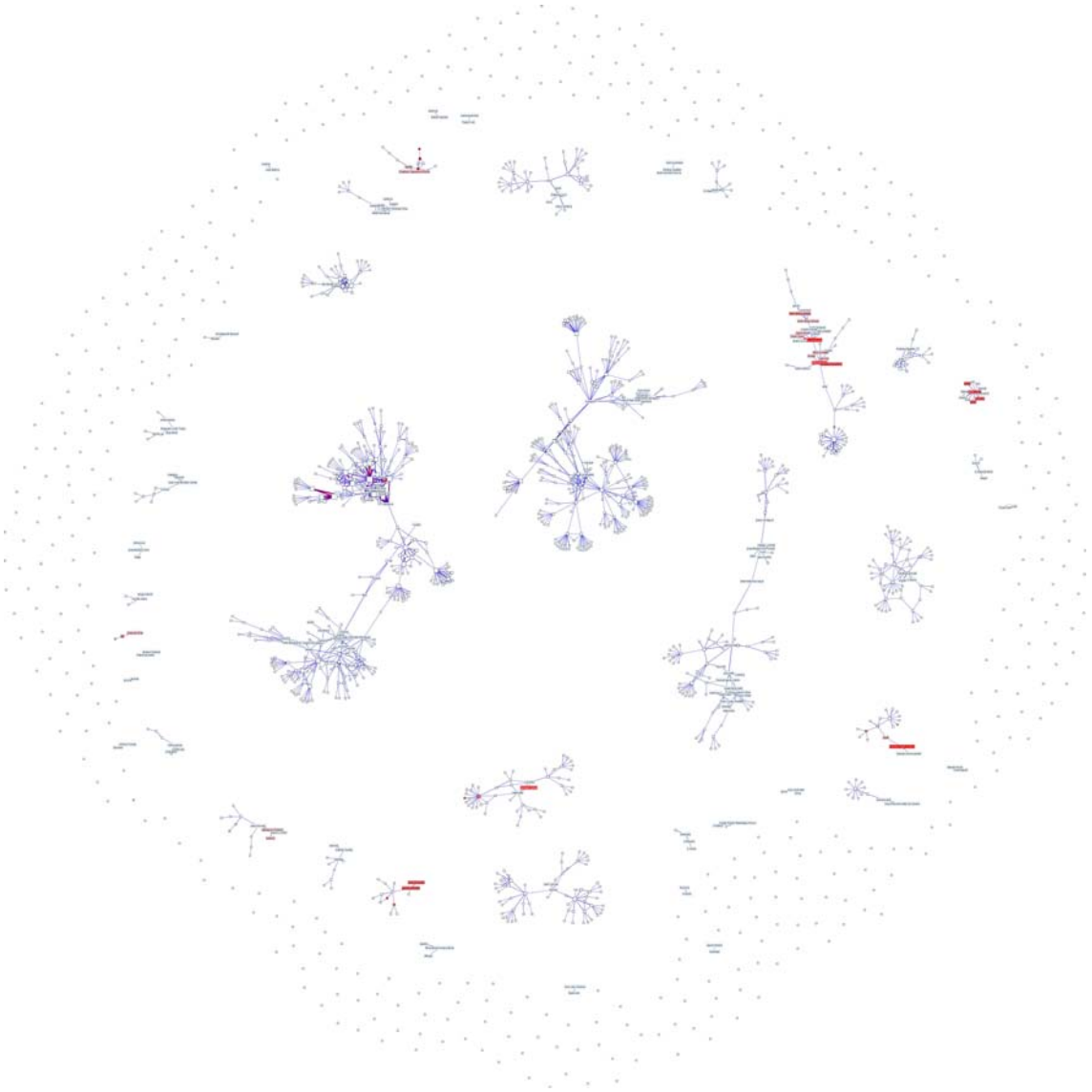
Anexo I. Mapa de la red social de la lista REDES (segmentado por dominios)



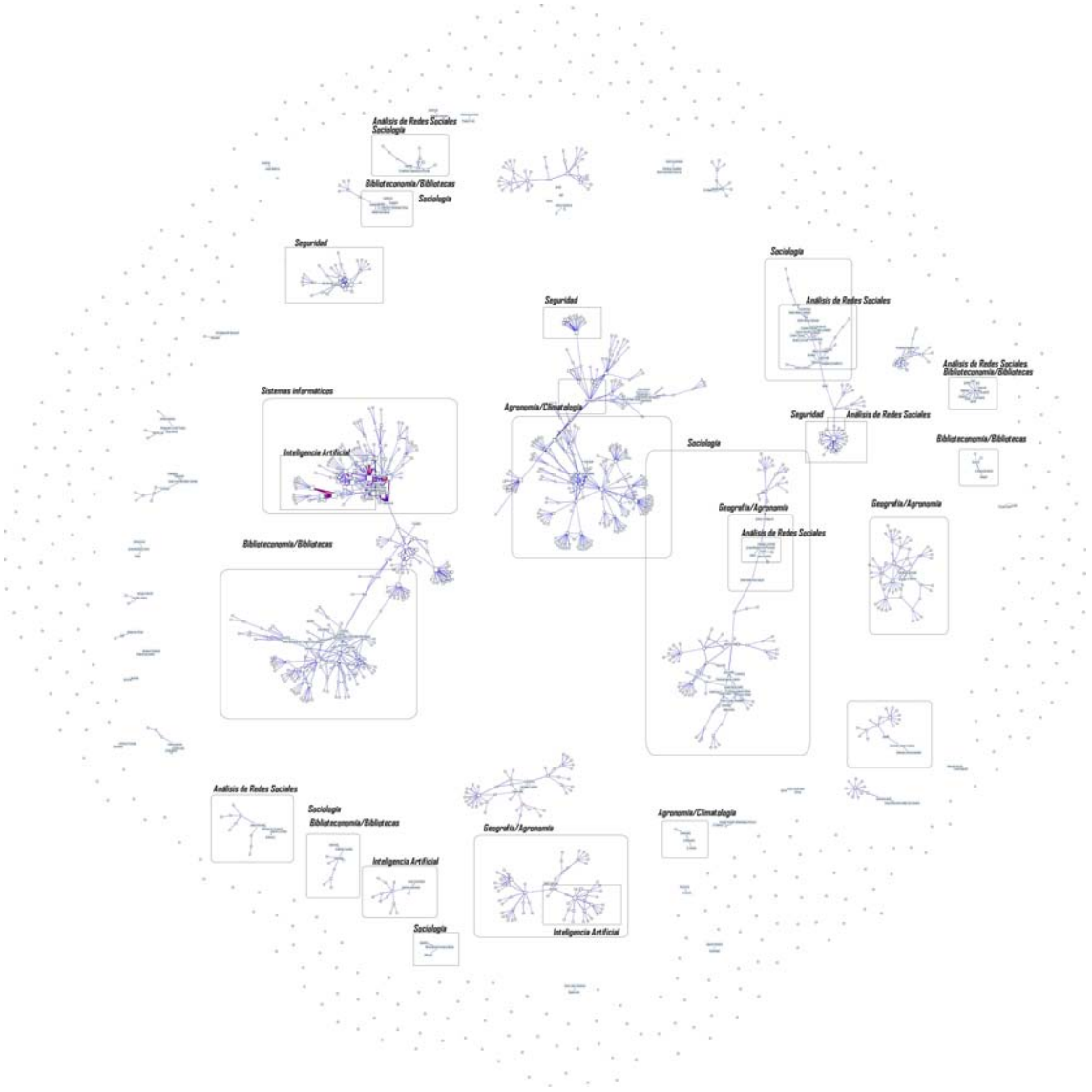
Anexo II. Mapa temático (frecuencia de ocurrencias). Departamento de lenguajes y sistemas informáticos



Anexo III. Mapa temático (mindshare). Herramienta Pajek



Anexo IV. Mapa temático (frecuencia de ocurrencias). Segmentación de la red en áreas de investigación.



Anexo V. Mapa conceptual