Grupo ThinkEPI
Estrategia y Prospectiva de la Información

Text mining versus redes neuronales. Dos métodos de análisis aplicados al caso de las políticas de las revistas sobre datos

Alicia García-García, Xavier García-Massó, Antonia Ferrer, Luis-Millán González, Fernanda Peset, Miguel Villamón y Rafael Aleixandre

Agradecimientos a Rosaura Santos Serra

4ª Conferencia internacional sobre calidad de revistas de ciencias sociales y humanidades (CRECS 2014)

Madrid, 8-9 de mayo de 2014.
15'45-16'15 Casa del Lector de la Fundación Germán Sánchez Ruipérez
http://www.thinkepi.net/crecs2014

GOBIERNO DE ESPAÑA · MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

UNIVERSITAT POLITÈCNICA DE VALÈNCIA · CSIC

DATASEA DATOS ABIERTOS DE INVESTIGACIÓN

Planteamiento del problema

Importancia del tema: En la actual sociedad de la información cada día a día se multiplica la cantidad de datos almacenados casi de forma exponencial. Contienen valiosa información que puede resultar muy útil para determinados procesos. Actualmente se buscan las vías para explotarlos de forma eficaz con diferentes objetivos. Además, estos datos pueden ser reutilizados

LÍMITACIONES COGNITIVAS Y DE RECURSOS nos empujan a buscar esos métodos de análisis del BigData

Nuestro objeto de estudio

- Políticas de revistas: cómo las revistas indicaban a los autores qué debían hacer con el material adicional, es decir con los datos que subyacen a las publicaciones.

- El almacenamiento de datos junto a las revistas es una vía reconocida (junto a subirlos a repositorios y bancos). Como ejemplo: Data Conservancy, IEEE y Portico han recibido una subvención (600.000 $ para dos años, mayo 2014) de la Alfred P. Sloan Foundation para conectar publicaciones y sus datos.

- Nosotros estudiamos las políticas en el proyecto DATASEA sobre las vías de almacenamiento y reutilización de datos de investigación *OPENDATASCIENCE, centro de recursos para la preservación y gestión de datos abiertos de investigación*", CS02012-39632-C02-02.

- El objetivo final de DATASEA es crear un buscador que permita recuperar conjuntos de datos de investigación

- El problema de la **recuperación de la información** es complejo, sobre todo siendo tan específica

Factores determinantes de la eficiencia en la RI

A) existencia de estándares aceptados con los que trabajar

B) recopilación de la información

C) creación de algoritmos específicos

Grupo ThinkEPI
Estrategia y Prospectiva de la Información

Verdadero problema (al menos en revistas)

Que las ideas están expresadas en lenguaje natural (NLP)

Las ideas tienen identidad respecto a algo que no son. No hay relación directa con el soporte que las contienen



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CSIC

DATASEA
DATOS ABIERTOS DE INVESTIGACIÓN

a) Estándares

- Si los datos estuvieran estructurados y disponibles de forma uniforme universalmente sería "fácil" analizarlos, reutilizarlos, recombinarlos...
- Pero...

- UNREALISTIC: cada grupo empresarial tiene sus propias políticas de expansión
- El acceso abierto o no a la información científica aun no tiene universalidad; mucho menos en datos

b) Recopilación de fuentes de origen de datos

- La búsqueda de información excelente y puesta a disposición en algún sistema lo hacemos sistemáticamente los documentalistas (directorios, bibliotecas, inventarios...). Pero cuando cambian, hay que volver a revisar.

- En 2011 creamos ODiSEA, como inventario de bancos de datos de investigación. Desde la aparición de Databib (2012) y Re3data (2013), unidos desde 2014, lo reenfocamos a la recopilación de sedes web de revistas de mayor impacto de todas las disciplinas que aceptan datos

- Esto nutre el futuro **OpenDataScience** por parte de expertos

- La interpretación de las políticas por un analista hace que los métodos sean muy dependientes de quién realiza esa calificación; al tiempo que están limitados a los recursos humanos que se dediquen a ello.

c) creación de algoritmos específicos

- Algoritmo es una secuencia de órdenes que tienen una lógica interna entre sí.
- El de Google usa palabras y luego rankiniza resultados, pero no parece que utilice modelos avanzados.
- Como ejemplo, en ajedrez se ha de usar inteligencia artificial en los simuladores, pues no puede calcular tanto número de variantes hasta final del juego. Así que le introducen algoritmo de análisis con premisas lógicas adaptadas al juego, por ejemplo asignando peso mayor al centro del tablero
- Hay modelos matemáticos que refinan las búsquedas cuando se escriben en lenguajes de programación estándar como matlab. Por ejemplo, asignan determinada fuerza a ciertas palabras, o indican que x término sea la palabra central dentro del grupo de palabras recuperadas

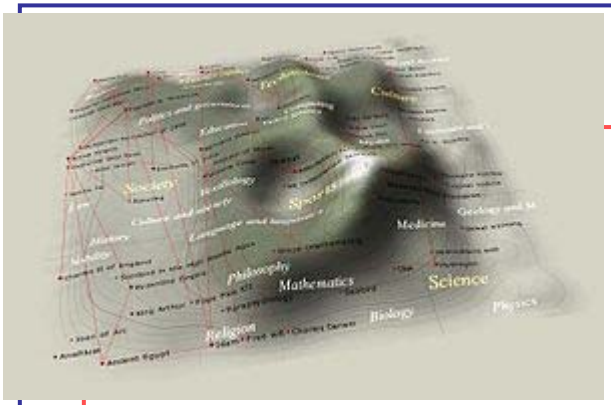
- Indagamos qué sistemas existen en el mercado a nivel matemático o informático para hacer más inteligente el futuro buscador. Que no sea la simple recuperación en las fuentes indicadas. Ambos son complementarios. Pero no buscan de igual manera. La colaboración con otros expertos/disciplinas nos permite alimentarlo

- Pretende crear algoritmos de búsqueda que nutran el modelo de recuperación contemplando premisas lógicas especiales para el problema estudiado.

- De esa forma, filtran la información hoy, y en el futuro

¿Qué es lo que estamos buscando?

- Método de análisis más eficaz para encontrar datos que puedan ser reutilizados
- En este momento estamos probando diferentes soluciones matemáticas para conseguir el objetivo: un buscador avanzado
- El año pasado ya se presentó un tipo de método basado en redes neuronales



Grupo ThinkEPI
Estrategia y Prospectiva de la Información

Aproximación al problema a partir de los mapas auto-organizados

- SOM-Self Organizing Maps es un Clasificador avanzado de características o vectores (**Teuvo Kohonen** en 1981-82)
- Es una clase de algoritmos de redes neuronales competitivas en la categoría de aprendizaje no supervisado. Construye 1º una red de nodos con un peso aleatorio. Iterativamente el algoritmo analiza los distintos casos y los resitúa en nodos (neuronas). Produce una neurona ganadora y modifica la forma de la red.
- Se representa en un mapa de dos dimensiones, que puede ser interpretado de forma cualitativa. A diferencia del análisis de clusters, SOM toma en cuenta todas las variables al tiempo

Empleado para la clasificación de información, redes (Moya, Guerrero-Bote, Herrero-Solana, Campanario, Olmeda, Ortiz-Repiso...)

Primeros resultados con SOM

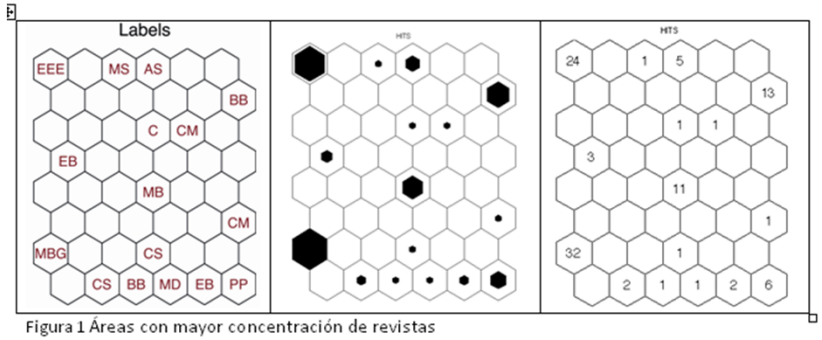


Figura 1 Áreas con mayor concentración de revistas

AS = Agricultural Science	EC = Enviroment, Ecology	**MD** = Multidisciplinary
BB= Biology & Biochemistry	GC = Geoscience	NB = Neuroscience Behaviour
C = Chemistry	IM = Immunology	**PT** = Physics
CM = Clinical Medicine	**MS** = Material Science	PS = Plan, animal Science
CS = Computer Science	MT = Mathematics	PAS = Psichiatry, Psichology
EB = Economic & Business	**MB** = Microbiology	**PP** = Social Science
EEE = Engineering	**MBG** = Molecular, Biology, Genetics	**SSG** = Space Science

Falta U-Matrix

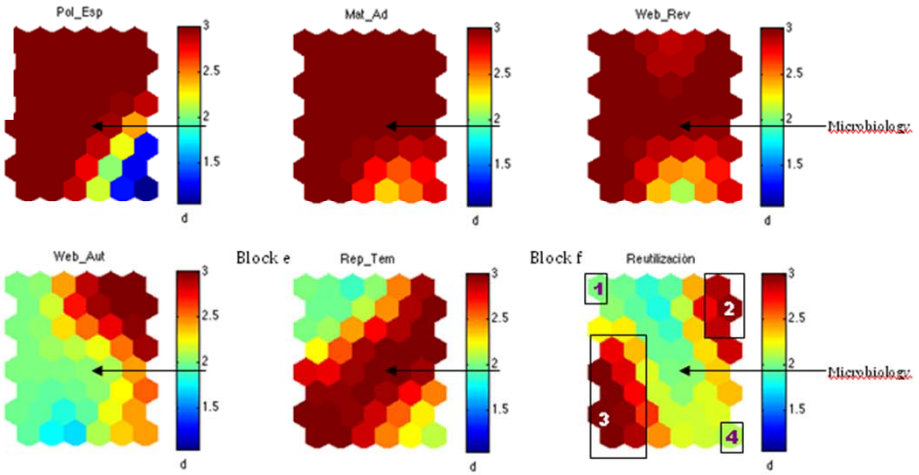


Figura 3. Self-organising map, ordered by variables.

Resumen de SOM

- VENTAJAS
  - Menos subjetivo que la recopilación de fuentes
  - Más automatizado que la interpretación de las fuentes recopiladas
  - Puede clasificar no sólo los contenidos buscados sino los no buscados/los contrarios y por tanto el grado de cumplimiento del criterio que se analiza
  - Emplea lenguaje matemático que puede ser implementado en algoritmos de búsqueda
- DESVENTAJAS
  - Necesita cierta acción del recopilador del material
  - Más lento que un buscador normal: el proceso de computación puedes llevar a durar horas

Grupo ThinkEPI

Estrategia y Prospectiva de la Información

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CSIC

Aproximación al problema a partir de la minería de texto

Definición: MINERÍA DE DATOS se define como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de datos.

- Minería de texto: ofrece con sistemas automáticos una solución, sustituyendo o complementando el trabajo de personas, sin que importe la cantidad de texto. Analiza grandes colecciones de texto para descubrir información antes desconocida. Pueden ser relaciones o patrones escondidos que de otro modo serían extremadamente difícil o imposibles de descubrir.

http://sitecore.jisc.ac.uk/publications/briefingpapers/2008/bptextminingv2.aspx

- Conjunto de técnicas pero nosotros solo vamos a explicar una: co-word
- Definición. Recuenta las veces en que dos palabras se relacionan en un entorno de lenguaje natural. Se convierte en un método de procesamiento de lenguaje natural para encontrar algo de nuestro interés

Courtial, J. P. A coword analysis of scientometrics.

Coword

analysis is a tool for classification of objects, taking into account only primarily existing (not calculated, for instance through profiles) links between objects.

- Punto crítico es la sinonimia del término datos (en revistas).

- Para este experimento
  - Hemos tenido que identificar las fuentes y los párrafos exactos donde se habla de datos ya que está en diferentes lugares: instrucciones para autor, políticas generales o específicas de revista...
  - Hemos unido en un concepto único las palabras clave que hemos identificado o generalizado como datos ya que utilizan todo tipo de términos: supplemental information o supplementary material. En el futuro cualquier otra que se utilice

Aproximaciones previas

- Piwowar y Chapman utilizan métodos de análisis del lenguaje natural para identificar los autores que están utilizando un tipo determinado de datos, los microarrays, para después consultar en los bancos de datos que almacenan los microarrays. De esta manera, cuantifican cómo los están depositando.

Nuestro procesamiento

Grupo ThinkEPI
Estrategia y Prospectiva de la Información

1. Obtener el **material**: los textos (manualmente) y **refinar** el texto con stop words (de bibexcel)

2. Generar las redes de **co-ocurrencia** (bibexcel) con la intención de automatizarlo con lenguaje de programación como matlab

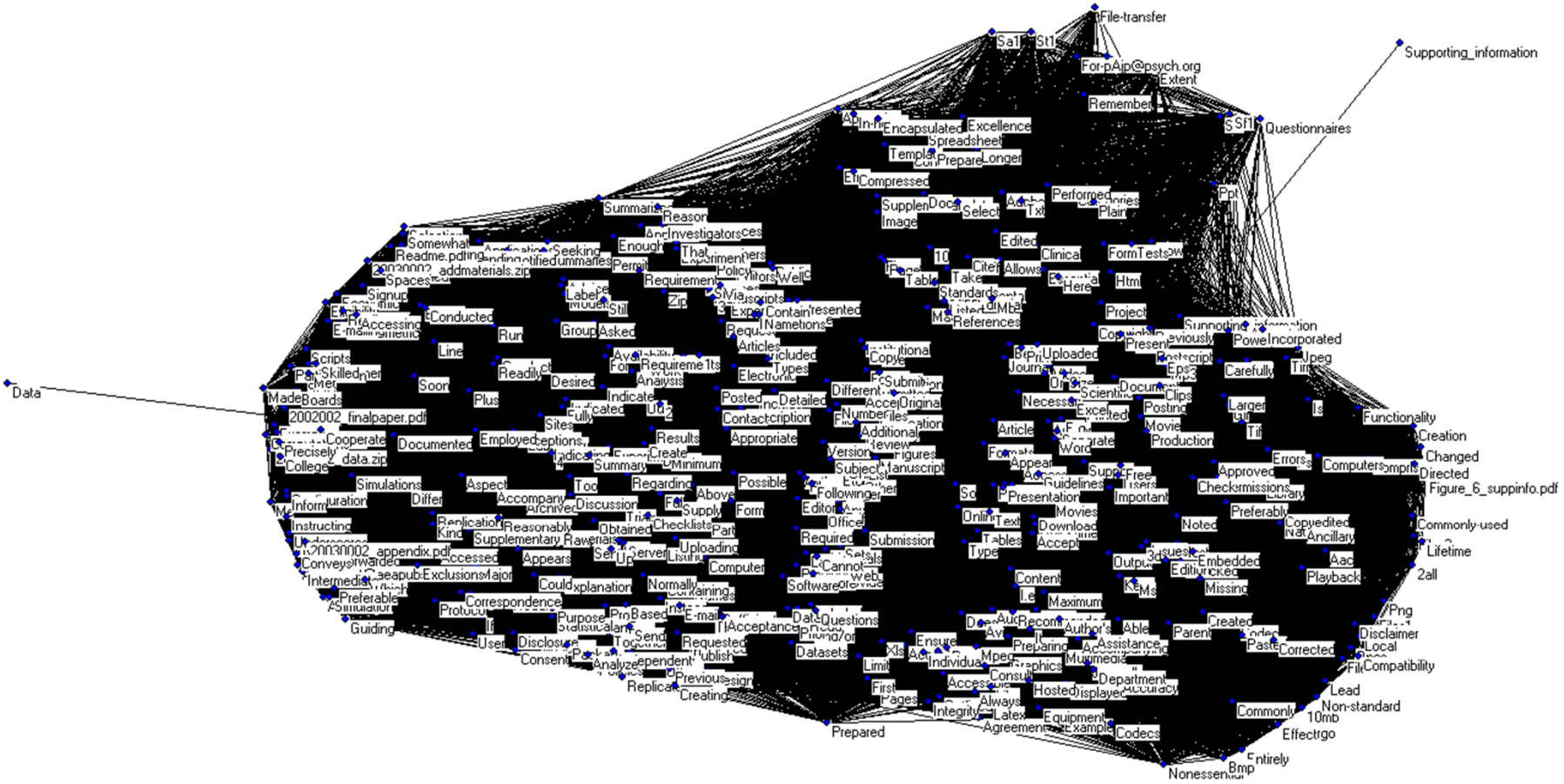
3. Establecer unos **umbrales** de interés: reducir la red (pajek o bibexcel)

4. Seleccionar las **conexiones** que tienen que ver con data (momento manual del experimento)

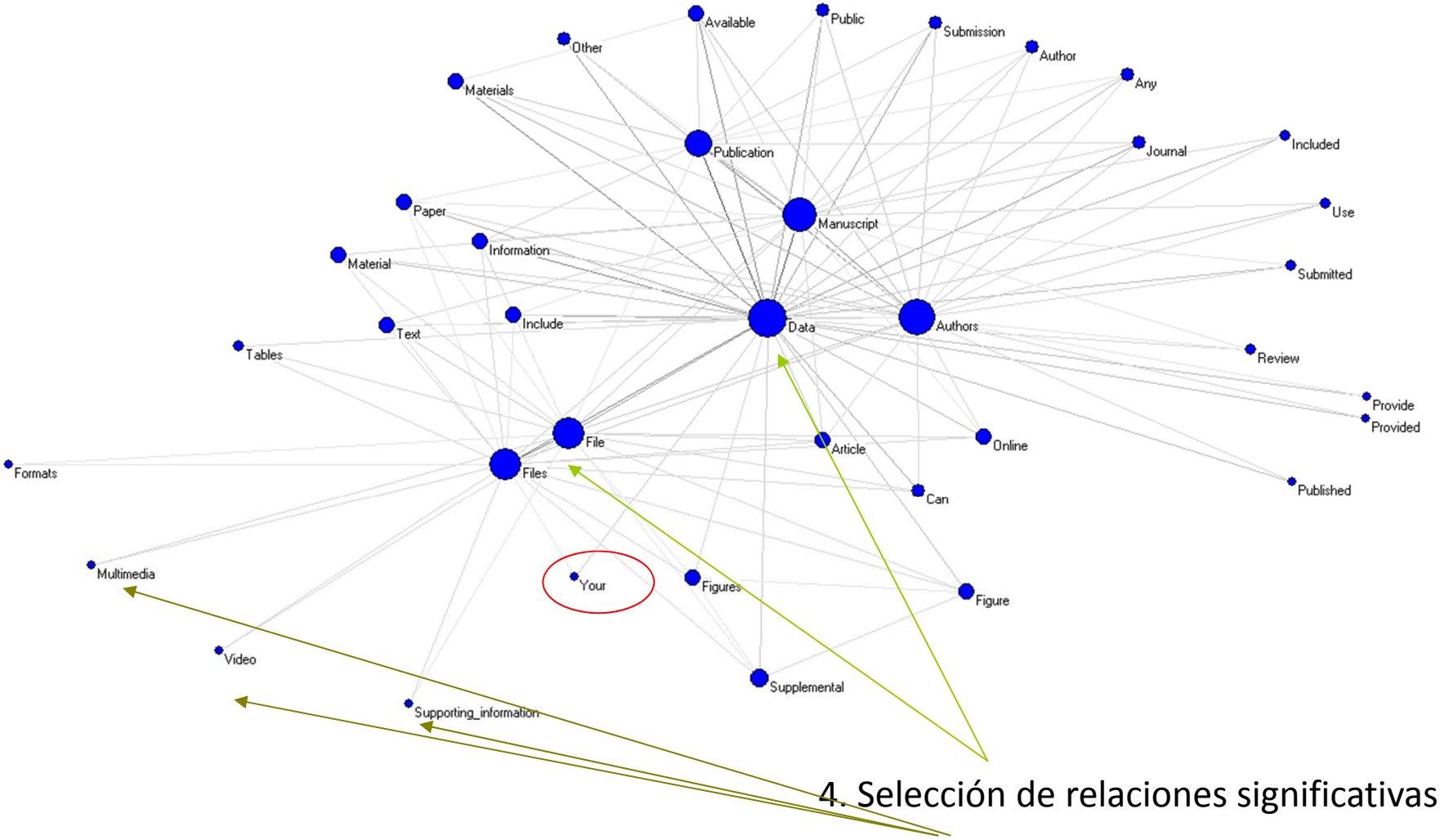
5. **Calcular** (registrar) la fuerza de las relaciones del término de interés (data) con otros significativos (file u otras similares). Existen muchos cálculos distintos por ejemplo centralidad de la palabra/s (pajek), intermediación, densidad (vosviewer) etc. Al buscador se le incluirían las palabras relacionadas, pero también sus ramificaciones, algo que nunca haría un buscador convencional

6. Con los cálculos realizados, **seleccionar la fuente** de donde provienen las palabras "premiadas" y en consecuencia las revistas

2. Primeros resultados coword



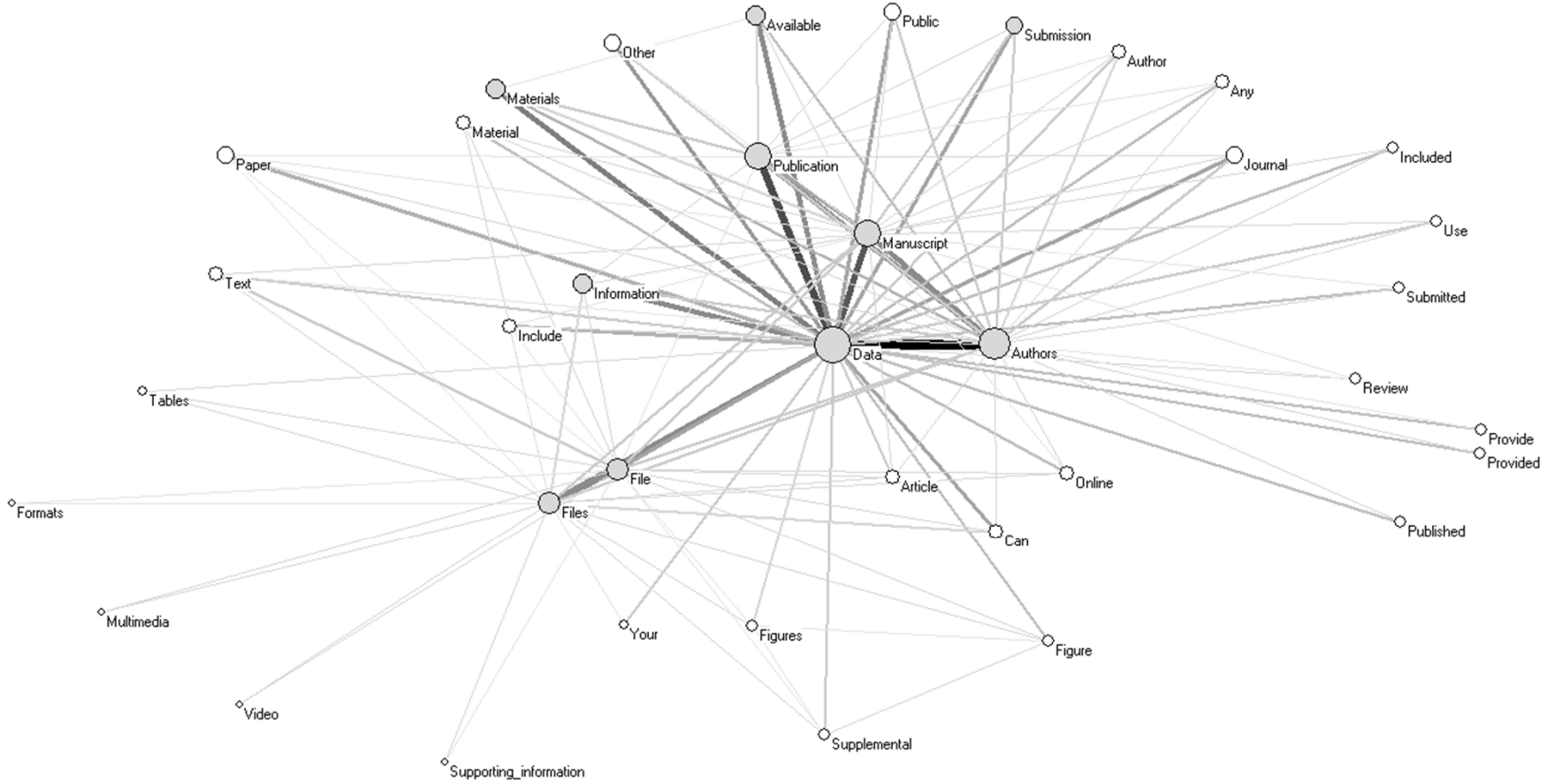
3. Umbral de interés



4. Selección de relaciones significativas

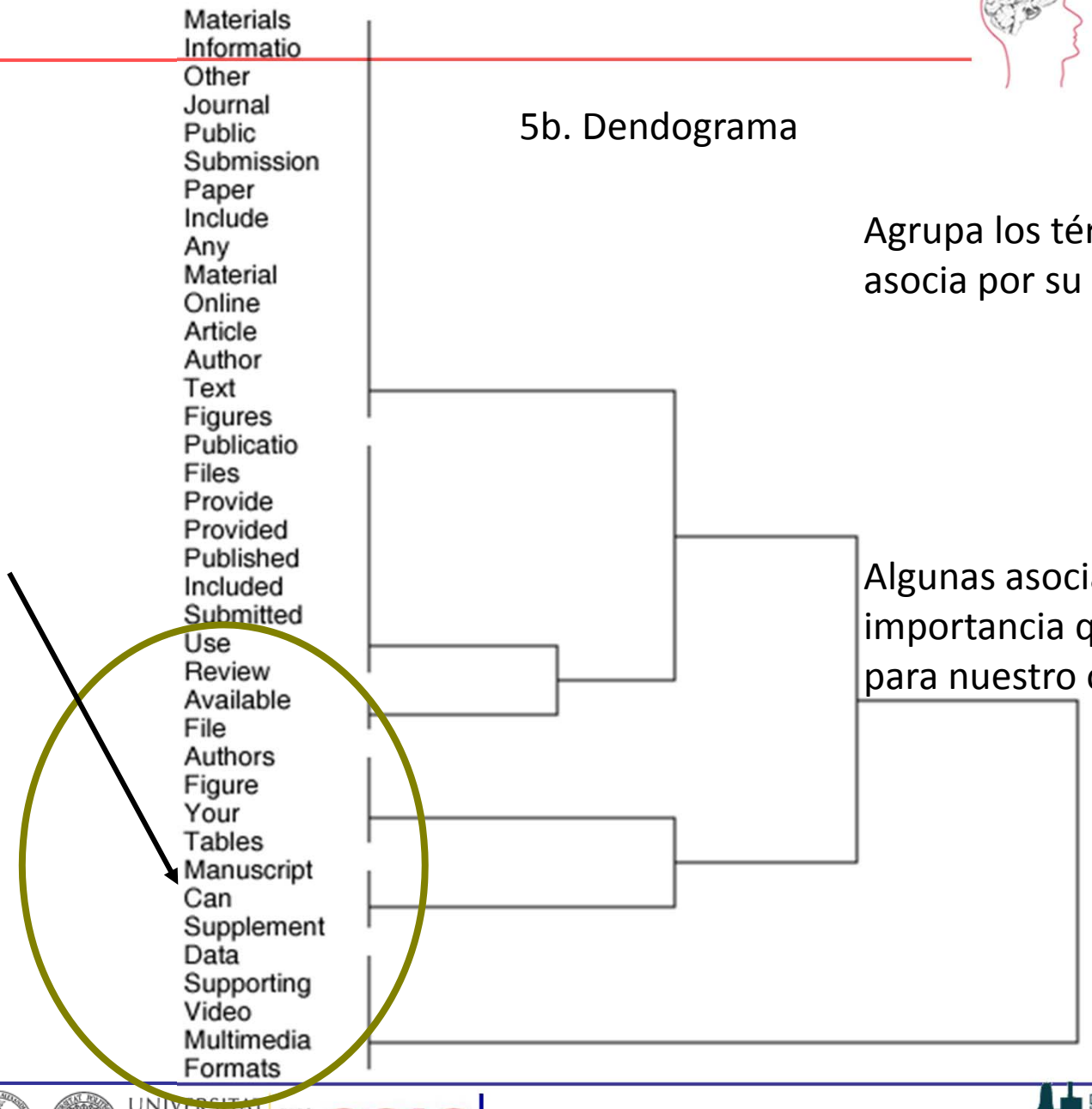
5a. Cálculo de Hubs and authority

Los grises son palabras autoridades y las blancas intermediadoras



5. Cálculo de fuerza de las palabras

5b. Dendograma



Agrupa los términos en clusters. Las asocia por su lógica

Algunas asociaciones tienen más importancia que otras asociaciones para nuestro objetivo, el buscador

Grupo ThinkEPI
Estrategia y Prospectiva de la Información

Recordemos:

1. Obtener el **material**: los textos (manualmente) y **refinar** el texto con stop words (de bibexcel)

2. Generar las redes de **co-ocurrencia** (bibexcel) con la intención de automatizarlo con lenguaje de programación como matlab

3. Establecer unos **umbrales** de interés: reducir la red (pajek o bibexcel)

4. Seleccionar las **conexiones** que tienen que ver con data (momento manual del experimento)

5. **Calcular** (registrar) la fuerza de las relaciones del término de interés (data) con otros significativos (file u otras similares). Existen muchos cálculos distintos por ejemplo centralidad de la palabra/s (pajek), intermediación, densidad (vosviewer) etc. Al buscador se le incluirían las palabras relacionadas, pero también sus ramificaciones, algo que nunca haría un buscador convencional

6. Con los cálculos realizados, **seleccionar la fuente** (URL) de donde provienen las palabras "premiadas" y en consecuencia las revistas

Resumen de textmining

VENTAJAS

- Muy bajo nivel de subjetividad
- No requiere recolección de información por parte de personas
- Emplea algoritmos de análisis en lenguajes de programación y puede ser implementado en buscadores

DESVENTAJAS

- El tiempo de computación es excesivamente lento y necesita de hardware sólido
- Que hasta la fecha muchas personas se dedican a la investigación con data mining pero no existe aplicaciones claras
- Requiere de buenas librerías de palabras (stop word, sinonimias, plurales... y de palabras clave como supplemental_information)

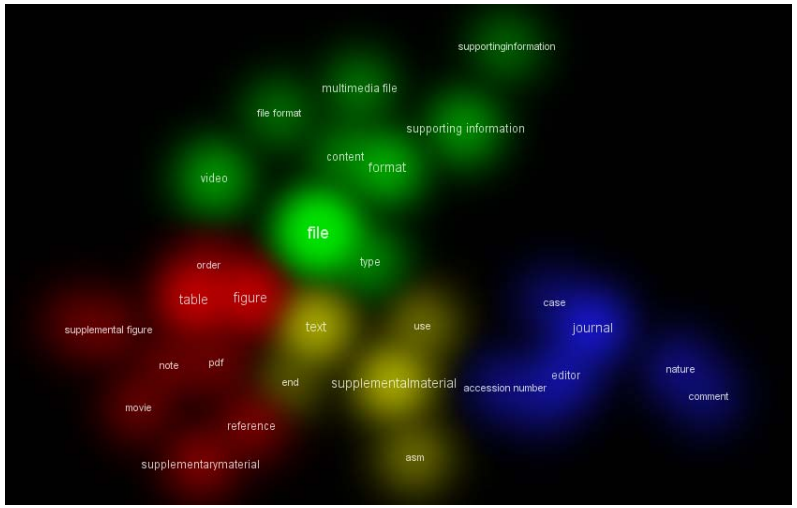
Grupo ThinkEPI

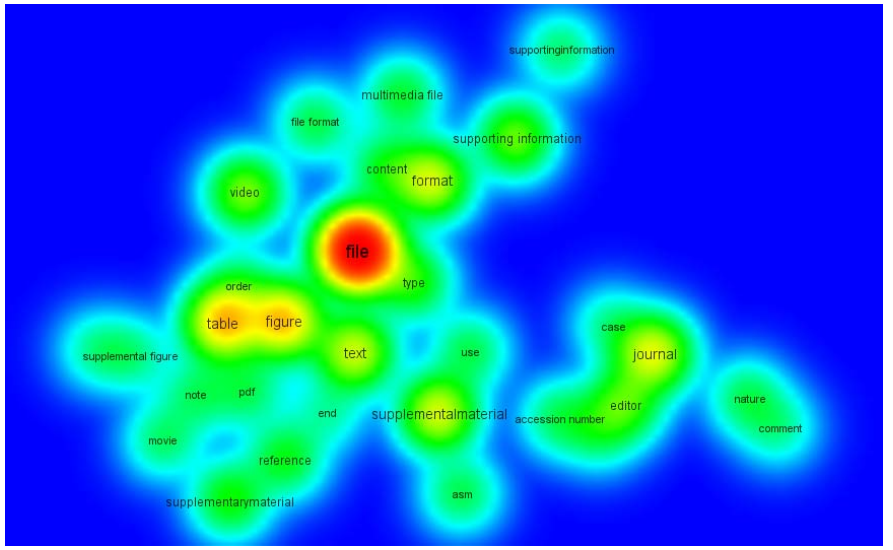
Estrategia y Prospectiva de la Información

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CSIC

Otros sistemas: clusters (vosviewer)





Otros sistemas: densidad (vosviewer)

No aparece data, no es frecuente

Conclusiones

- En este momento no tenemos conclusiones definitivas, pero si esta metodología es correcta, su uso contaría con varias ventajas. Por una parte puede acometerse el estudio de muchas más revistas, pues solo es necesario extraer los textos de sus políticas y realizar un mínimo tratamiento previo al análisis. Por otra, podría aplicarse a otro tipo de textos, como podría ser el apartado de material y método de los artículos publicados en busca de cómo han tratado esos datos.

- Datasea sigue por dos vías: recolectando fuentes y experimentando métodos que puedan llegar a sustituir el trabajo de personas

Grupo ThinkEPI
Estrategia y Prospectiva de la Información

Próximamente esperamos verlo publicado

http://www.datasea.es

Gracias por vuestra atención

mpesetm@upv.es