HADOOP, como una plataforma de procesamiento masivo de datos

ANALISIS DEL SITIO WEB DE LA ESPOL

Esquema de Trabajo

- Objetivo
- Alcances
- Limitaciones
- Fundamentos Teóricos

Análisis Preliminar

- Herramientas
- Estructura del Cluster
- Resultados

Desarrollo del Estudio

- Análisis
- Conclusiones
- Recomendaciones

Resultados

Objetivo

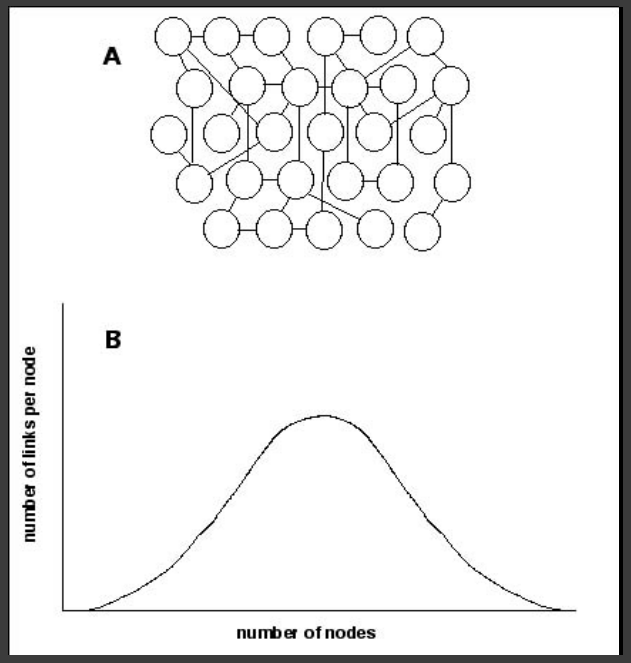
- Realizar un análisis de la estructura de la Web de la ESPOL, para determinar si tiene una forma ***Pequeño Mundo***, con miras a mejorar su navegabilidad y el valor a los usuarios de la misma

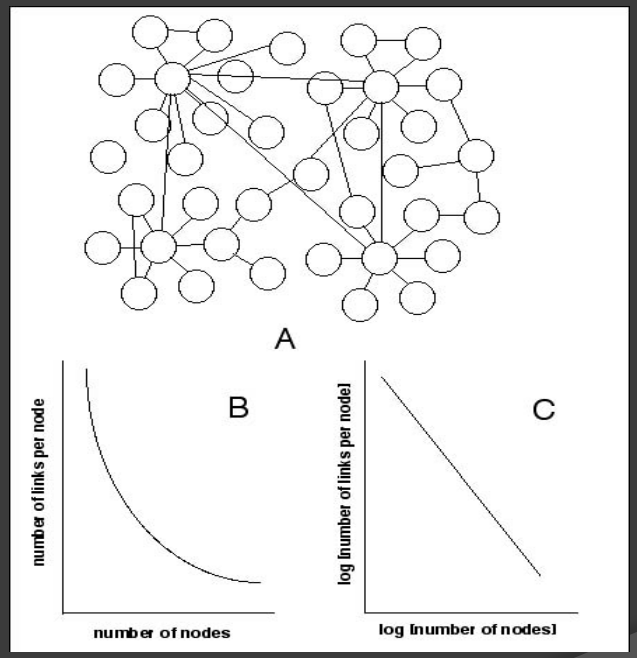
Fundamentos Teóricos: Redes Pequeño Mundo

- Grafo en el que mayoría de los nodos no son vecinos entre sí
  - Pero, la mayoría de los nodos pueden ser alcanzados desde cualquier nodo a través de un número relativamente corto de saltos
- Posee valores altos de coeficiente de agrupamiento (clustering coefficient)
  - Aunque dos nodos en la red no estén conectados de forma directa, existe una gran probabilidad de que se conecten a través de otros nodos
- Estructura libre escala (scale free network)

Fundamentos Teóricos: Redes Pequeño Mundo

REPRESENTACION DE ESTÁNDARES DE REDES





ALEATORIA

LIBRE DE ESCALA

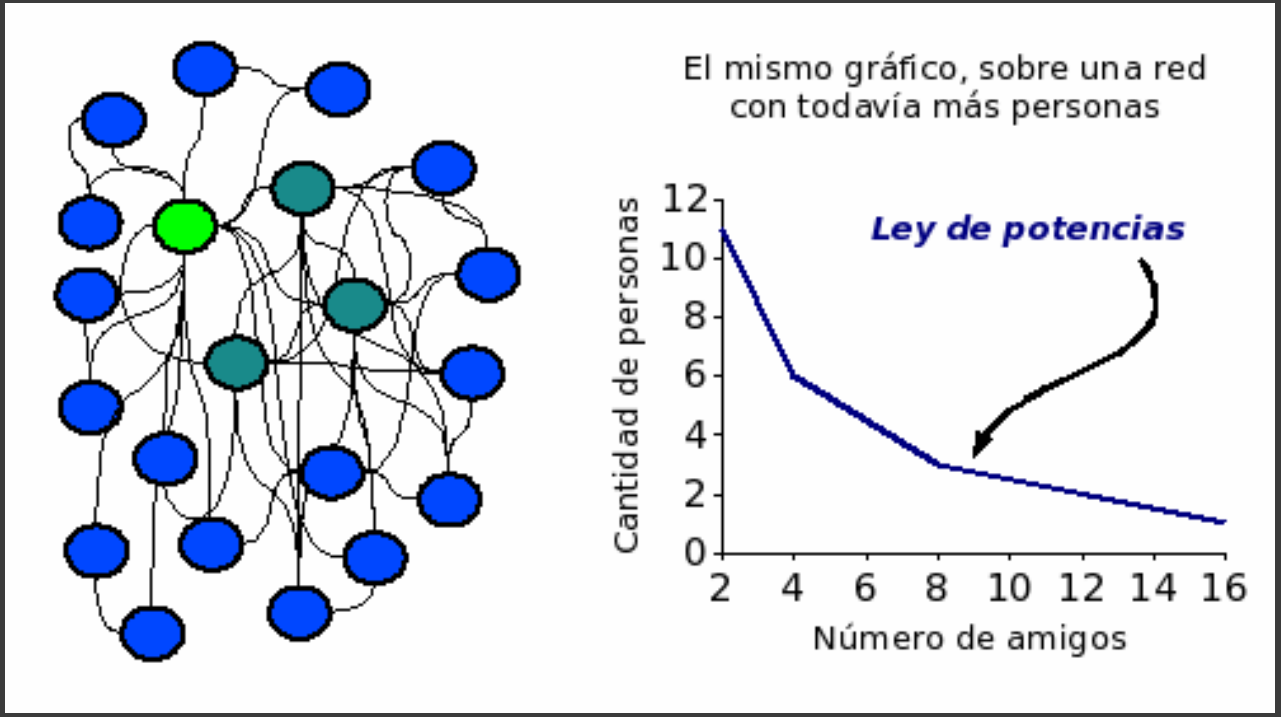
Redes Libre Escala

- Distribución de enlaces sigue a la de la Ley de las Potencias
  - La fracción P(x) de nodos en una red que tienen K conexiones hacia otros nodos, para grandes valores de K,

$$P(k) \approx k^{-\gamma}$$

  - γ: constante, generalmente está entre 2 y 3
- Algunos nodos se encuentran altamente conectados, aunque el grado de conexión de casi todos los nodos es bajo

Redes Libre Escala



CARACTERÍSTICA DE LA RED LIBRE DE ESCALA

Ley de Potencias

- Tipo de relación matemática entre dos cantidades
- Si una cantidad es la frecuencia y la otra el tamaño del evento en sí, entonces la relación es una distribución ***Ley de Potencia***s si el tamaño del evento incrementa de forma en que la frecuencia del evento decrece lentamente
- Una relación en forma de ***Ley de Potencias*** entre dos escalares cuantitativos X y Y es aquella que puede expresarse como sigue:

$$y = ax^k$$

  - **a**: constante de proporcionalidad
  - ***k***: exponente de la potencia (constante)

Esquema de Trabajo

- Objetivo
- Alcances
- Limitaciones
- Fundamentos Teóricos

Análisis Preliminar

- Herramientas
- Estructura del Cluster
- Resultados

Desarrollo del Estudio

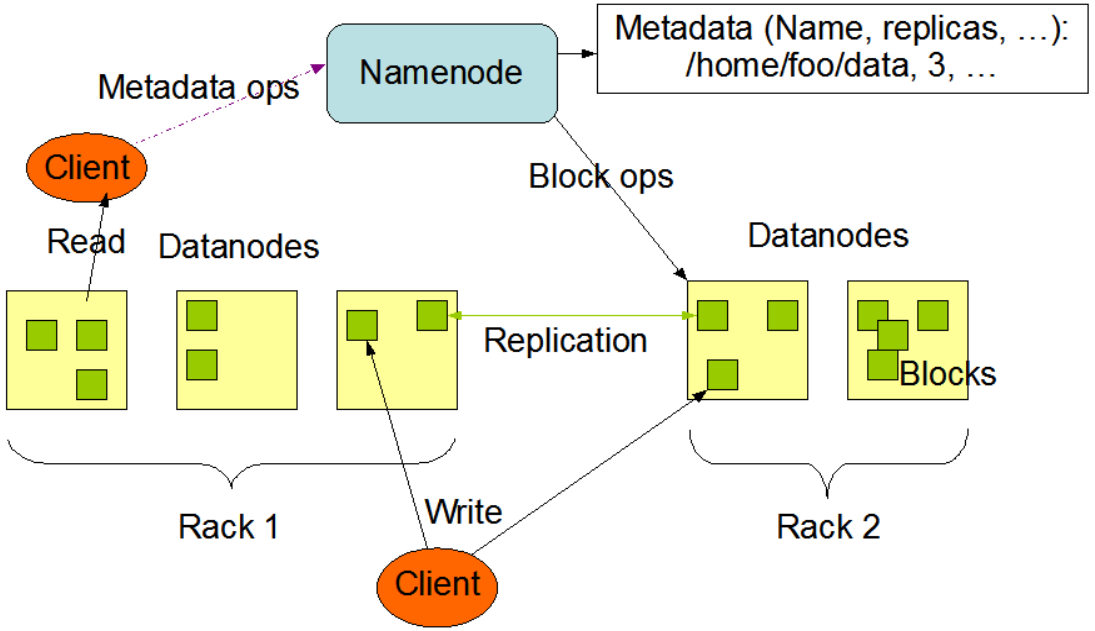
- Análisis
- Conclusiones
- Recomendaciones

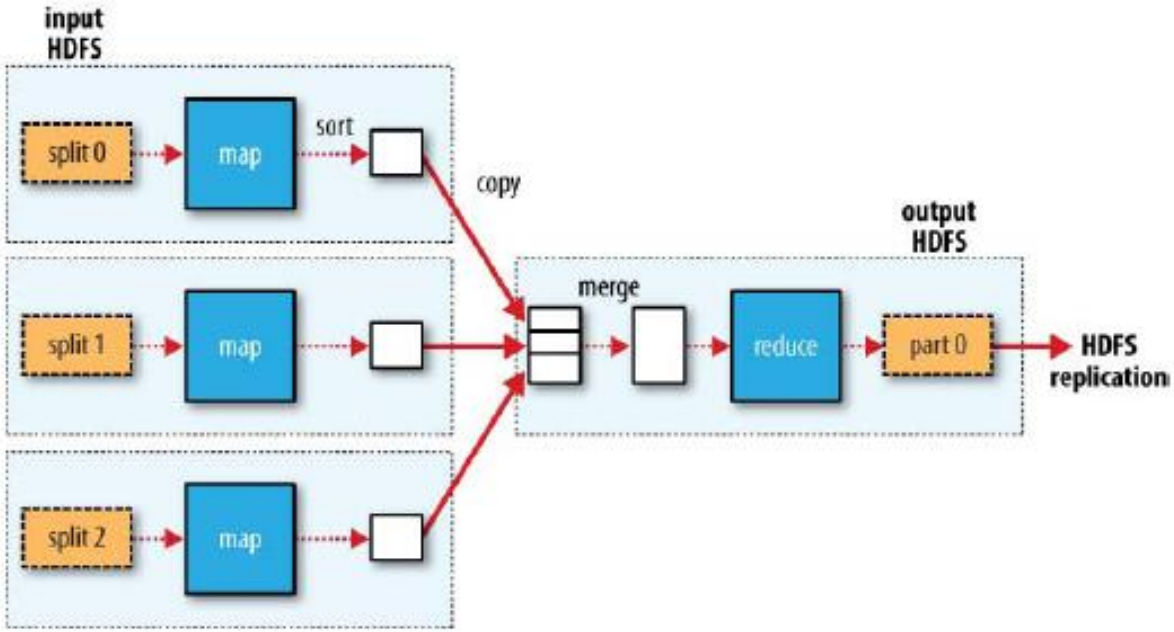
Resultados

Herramientas: HADOOP

- Plataforma de procesamiento de datos masivos, de código libre
  - Del Apache Software Foundation
- Basada en el paradigma de programación distribuida MapReduce
  - Similar a Dividir y Vencer pero que se aplica a grandes volúmenes de datos
- HDFS: sistema de archivos distribuido
- Altamente escalable y tolerante a fallos

HDFS Architecture

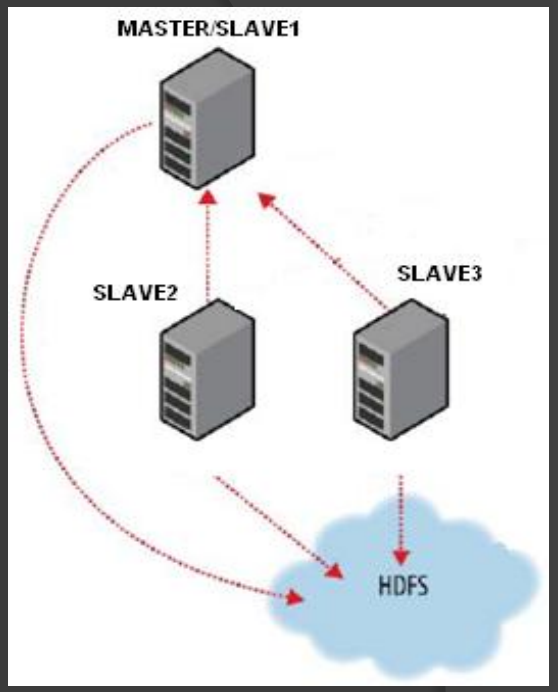




Herramientas: Otras

- **Ubuntu Studio Works**
- **NUTCH**
- **ECLIPSE**
- **JAVA**
- **TOMCAT**
- **R-project**
- **Visualización:**
  - **Graphviz**
  - **Lanetvi**
  - **Cytoscape**

Desarrollo del Cluster



- Instalamos y configuramos un clúster Hadoop
- Descargamos Web de ESPOL
- Escribimos código Map-Reduce para procesar los Web y generar los resultados
- Resultados de procesados con el lenguaje R (R-project), para el análisis estadístico de las propiedades del grafo

Resultados

- Obtuvimos los enlaces de toda la red de la ESPOL
- Generamos una tabulación con los enlaces de salida y de entrada, de manera que visualizamos el grado de nodos y la frecuencia
- Procesamos estos datos en la herramienta R-project para determinar si la distribución de los enlaces de salida y entrada seguía una ley de potencias

Esquema de Trabajo

- Objetivo
- Alcances
- Limitaciones
- Fundamentos Teóricos

Análisis Preliminar

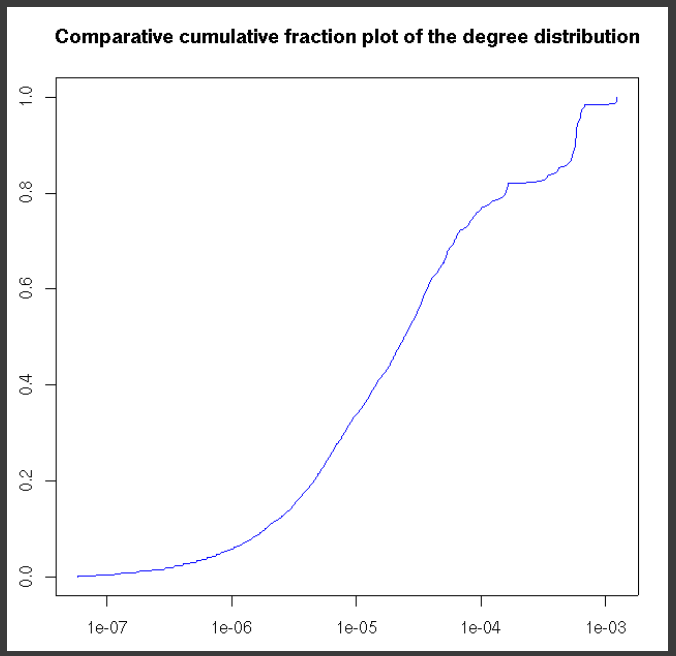
- Herramientas
- Estructura del Cluster
- Resultados

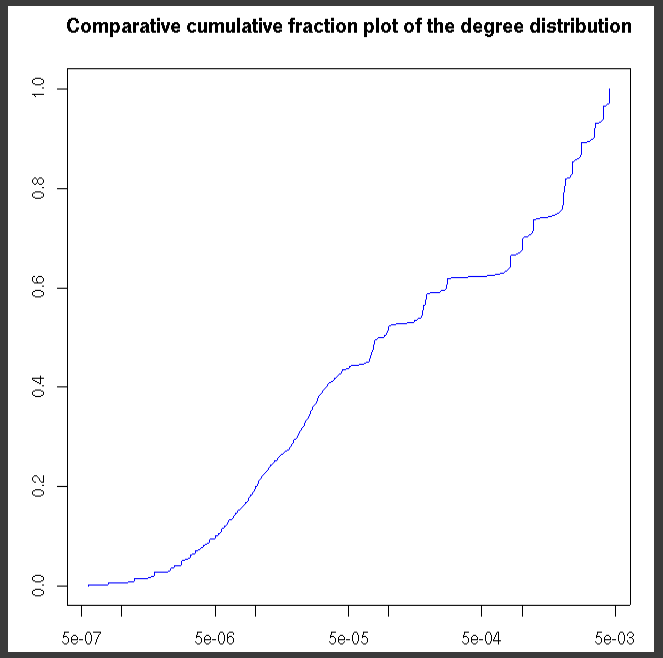
Desarrollo del Estudio

- Análisis
- Conclusiones
- Recomendaciones

Resultados

Análisis: Sitio Web ESPOL





- Web de ESPOL no sigue una distribución de ley de potencias
  - Coeficiente de agrupamiento alfa < 1

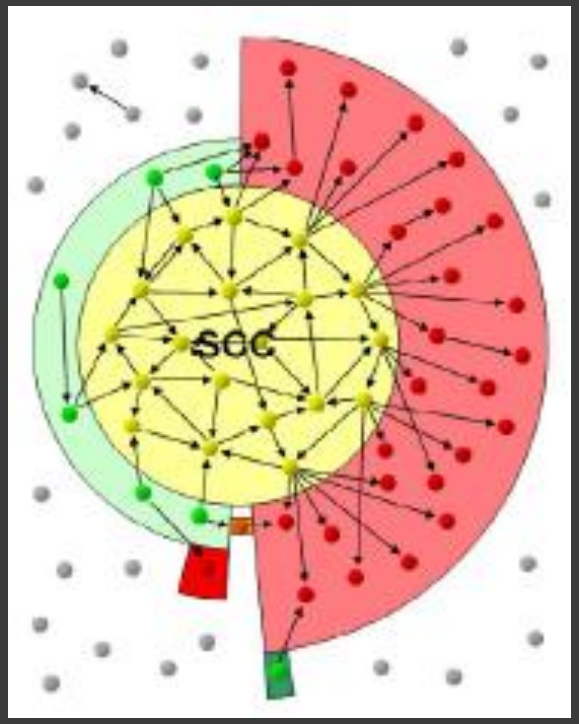
➔ Web de ESPOL no es una pequeño mundo

Análisis: Otros Sitios

- Para dar mayor realce a este estudio se creyó conveniente agregar un estudio de otros sitios, para su comparación
- Se escogió las universidades de Reino Unido (UK)
  - Estudio demuestra sus Web tienen la estructura pequeño mundo y muestra los beneficios de esta estructura en la navegabilidad de las mismas

Björneborn, Lennart. "Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach" Tesis Doctoral, del Departamento de Estudios de la Información, Royal School of Library and Information Science, Dinamarca.

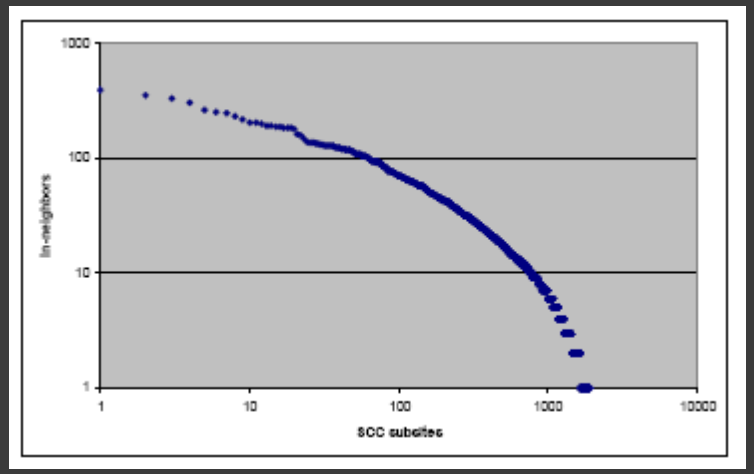
Análisis: Red de universidades del Reino Unido (UK)



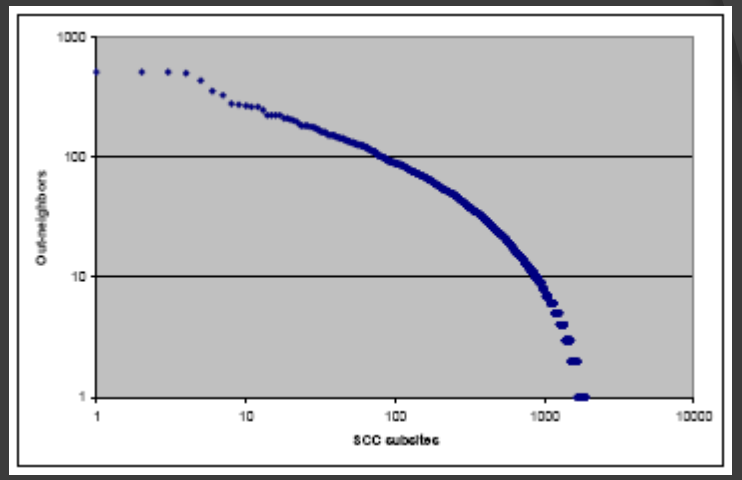
ANALISIS CENTRADO EN LA RED SCC

Björneborn, Lennart. "Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach" Tesis Doctoral, del Departamento de Estudios de la Información, Royal School of Library and Information Science, Dinamarca.

Análisis: Red de universidades del Reino Unido (UK)



DISTRIBUCIONES DE ENLACES ENTRANTES PARA 1893 SUB-SITIOS DE LA RED SCC EN ESCALA LOG-LOG



DISTRIBUCIONES DE ENLACES SALIENTES PARA 1893 SUB-SITIOS DE LA RED SCC EN ESCALA LOG-LOG

Björneborn, Lennart. "Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach" Tesis Doctoral, del Departamento de Estudios de la Información, Royal School of Library and Information Science, Dinamarca.

Análisis: Red de universidades del Reino Unido (UK)

LOS 15 SUB-SITIOS CON LA MAYOR CANTIDAD DE ENLACES ENTRANTES HACIA SUS VECINOS EN LA RED

rank	Id	subsite	comp.	# in-neigh-bors	# out-neigh-bors	affiliation
1	4124	src.doc.ic.ac.uk	OUT	411	0	'Sun Site' mirror site (Usenet archive) at Imperial College, London
2	1357	cbl.leeds.ac.uk	SCC	387	91	Computer Based Learning Unit, Univ. of Leeds
3	3017	scit.wlv.ac.uk	SCC	349	434	School of Computing and Info.Technology, Univ. of Wolverhampton
4	4129	sunsite.doc.ic.ac.uk	OUT	331	0	'Sun Site' mirror site (Usenet archive) at Imperial College, London
5	1821	users.ox.ac.uk	SCC	330	507	Personal web pages at Univ. of Oxford
6	2760	cs.ucl.ac.uk	SCC	300	265	Dept. of Computer Science, Univ. College London
7	4928	comlab.ox.ac.uk	OUT	289	0	Computing Laboratory (CS dept.), Univ. of Oxford
8	1866	info.ox.ac.uk	SCC	259	120	Former server with official web pages of Univ. of Oxford
9	3020	scitsc.wlv.ac.uk	SCC	249	8	School of Computing and Info. echnology, Univ. of Wolverhampton
10	3339	cup.cam.ac.uk	OUT	249	0	Cambridge University Press
11	325	cl.cam.ac.uk	SCC	246	141	Computer Laboratory (CS dept.), Univ. of Cambridge
12	2642	cogs.susx.ac.uk	SCC	231	268	School of Cognitive and Computing Sciences, Univ. of Sussex
13	1466	cs.man.ac.uk	SCC	216	224	Dept. of Computer Science, Univ. of Manchester
14	925	dcs.gla.ac.uk	SCC	203	511	Dept. of Computing Science, Univ. of Glasgow
15	3010	csv.warwick.ac.uk	SCC	202	354	Univ. of Warwick Information Service

Björneborn, Lennart. "Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach" Tesis Doctoral, del Departamento de Estudios de la Información, Royal School of Library and Information Science, Dinamarca.

Análisis: Red de universidades del Reino Unido (UK)

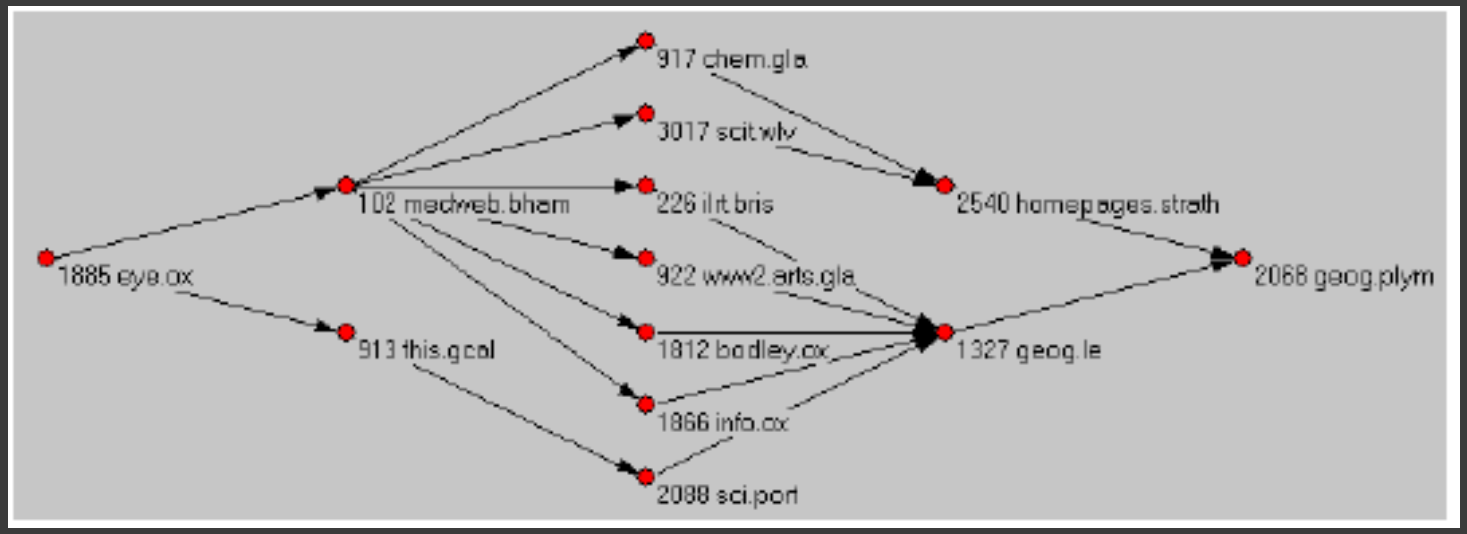
LOS 15 SUB-SITIOS CON LA MAYOR CANTIDAD DE ENLACES SALIENTES HACIA SUS VECINOS EN LA RED

rank	Id	subsite	comp.	# In-neigh-bors	# out-neigh-bors	affiliation
1	1088	cee.hw.ac.uk	SCC	148	518	Dept. of Computing and Electrical Engineering, Heriot-Watt Univ.
2	1572	doc.mmu.ac.uk	SCC	127	514	Dept. of Computing and Mathematics, Manchester Metropolitan Univ
3	925	dcs.gla.ac.uk	SCC	203	511	Dept. of Computing Science, Univ. of Glasgow
4	1821	users.ox.ac.uk	SCC	330	507	Personal web pages at Univ. of Oxford
5	3017	scit.wlv.ac.uk	SCC	349	434	School of Computing and Info. Technology, Univ. of Wolverhampton
6	3010	csv.warwick.ac.uk	SCC	202	354	Univ. of Warwick Information Service
7	2367	ecs.soton.ac.uk	SCC	117	327	Dept. of Electronics and Computer Science, Univ. of Southampton
8	791	dai.ed.ac.uk	SCC	137	280	Department of Artificial Intelligence, Univ. of Edinburgh
9	2291	afm.sbu.ac.uk	SCC	2	277	'Virtual library', Centre for Applied Formal Methods, South Bank Univ
10	2642	cogs.susx.ac.uk	SCC	231	268	School of Cognitive and Computing Sciences, Univ. of Sussex
11	1268	comp.lancs.ac.uk	SCC	183	265	Computing Dept., Univ. of Lancaster
12	2760	cs.ucl.ac.uk	SCC	300	265	Dept. of Computer Science, Univ. College London
13	19	users.aber.ac.uk	SCC	34	250	Personal web pages at Univ. of Wales, Aberystwyth
14	3042	www-users.york.ac.uk	SCC	94	226	Personal web pages at Univ. of York
15	1597	dcs.napier.ac.uk	SCC	127	226	School of Computing, Napier Univ.

Björneborn, Lennart. "Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach" Tesis Doctoral, del Departamento de Estudios de la Información, Royal School of Library and Information Science, Dinamarca.

Análisis: Red de universidades del Reino Unido (UK)

PATH NET NH05. TODOS LOS CAMINOS CORTOS ENTRE eye.ex.ec.uk Y geog.plym.ac.uk



Björneborn, Lennart. "Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach" Tesis Doctoral, del Departamento de Estudios de la Información, Royal School of Library and Information Science, Dinamarca.

Análisis: Red de universidades del Reino Unido (UK)

Conclusiones:

- Longitud del camino y coeficiente de agrupamiento cumplen los requisitos de una red pequeño mundo
- Las distribuciones de enlaces entrantes poseen propiedades pequeño mundo
- Se encontró que la red de nodos centrales (SCC) poseían una distribución pequeño mundo

Björneborn, Lennart. "Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach" Tesis Doctoral, del Departamento de Estudios de la Información, Royal School of Library and Information Science, Dinamarca.

Conclusiones

- La Web de la ESPOL no posee propiedades pequeño mundo en este momento
- Gracias al interés interno en el mejoramiento de su información, la estructura ha mejorado recientemente
- Hemos podido observar una reestructuración en los sitios de la ESPOL
  - Tuvo un impacto negativo en el presente estudio ya que produjo inconvenientes en el proceso de indexación
  - Ha mejorado la navegabilidad de la ESPOL, ya que ahora existe al menos un enlace de entrada y salida en cada sitio de la ESPOL
  - Aún no existe una correcta navegabilidad en muchos sitios, pues estos poseen problemas en su interacción con el usuario y no manejan un correcto enlazado de sitios

Recomendaciones

- Dado que el ranking de universidades de Webometrics mide la usabilidad de los sitios Web y la cantidad de información investigativa que este proporcione a los usuarios, debemos coordinar una correcta estructura, de manera que se priorice la visibilidad de las investigaciones realizadas dentro de la institución y que esto sirvan para el desarrollo de actividades similares en otras instituciones
- Realizar un nuevo estudio en un periodo no mayor a un año, para observar efectos de nuevas políticas
- Siempre enlazar al sitio principal de la ESPOL y otros sub-sitios relacionados

Recomendaciones

- Evitar en todo momento la creación de sitios Web que no estén enlazados directamente con algún sito representativo de la universidad
  - Debería estar conectada a una unidad, facultad, centro de investigación, etc. que tenga alguna característica común con el sitio que se pretende crear.
- Hacer un mantenimiento de los sitios de las unidades u otras representaciones dentro del dominio de la ESPOL de manera que no nos encontremos con sitios que no están activos o que carezcan de enlaces de entrada o salida