La próxima generación de motores de búsqueda

Julio Gonzalo

Grupo UNED en Recuperación de Información y Procesamiento de Lenguaje Natural

http://nlp.uned.es

El modelo clásico de IR...

Modelos formales

(solapamiento consulta/documento)

Necesidad información (precisa)

colección documentos (estática)

Expansión de la consulta

consulta

Comparación consulta/documento

Heurísticas de pesado (tf.idf)

Estructuras de datos

Docs *relevantes*

Filtrado

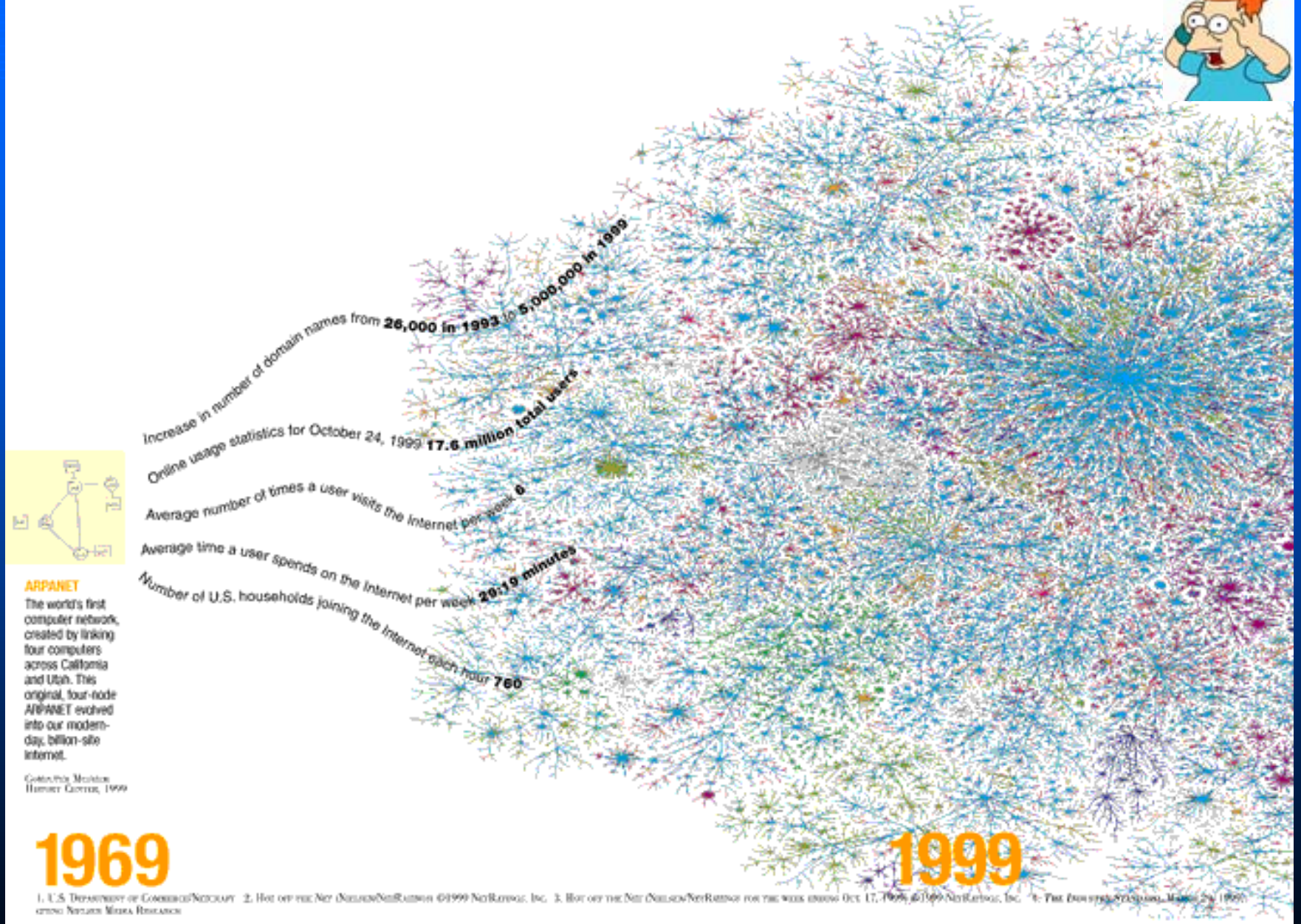
Visualización

Realimentación

Representación de documentos (índices)

Técnicas de agrupación

...frente a la búsqueda WWW



Buscadores web: Google

- web retrieval: búsqueda simple y eficiente + pagerank + ...
- Funciona muy bien para necesidades de información que se resuelven con uno o dos enlaces:
  - Páginas personales (e.g. “Julio Gonzalo”)
  - Portal de calidad sobre un tema (e.g. “Cámaras digitales”)
  - Portal de una empresa/institución/servicio web (e.g. “Instituto Cervantes”)
  - Página oficial de un evento (“Forum 2004”)

posicionamiento

- Pero cuando es necesario examinar varias páginas, reuniendo y relacionando información de varias fuentes, las listas ordenadas de Google no son suficiente:
  - Las listas ordenadas de enlaces no organizan conceptualmente los resultados de búsqueda.
  - Google no extrae ni relaciona *información* de los documentos.
  - Google nos “abandona” una vez que entramos en un portal, y los buscadores de portales son muy elementales.

Líneas de futuro inmediato

- Las listas de enlaces no organizan conceptualmente los resultados de búsqueda.

  → **1. Organización y visualización de resultados**

- Google no extrae ni relaciona *información* de los documentos.

  → **2. Buscadores "verticales"**

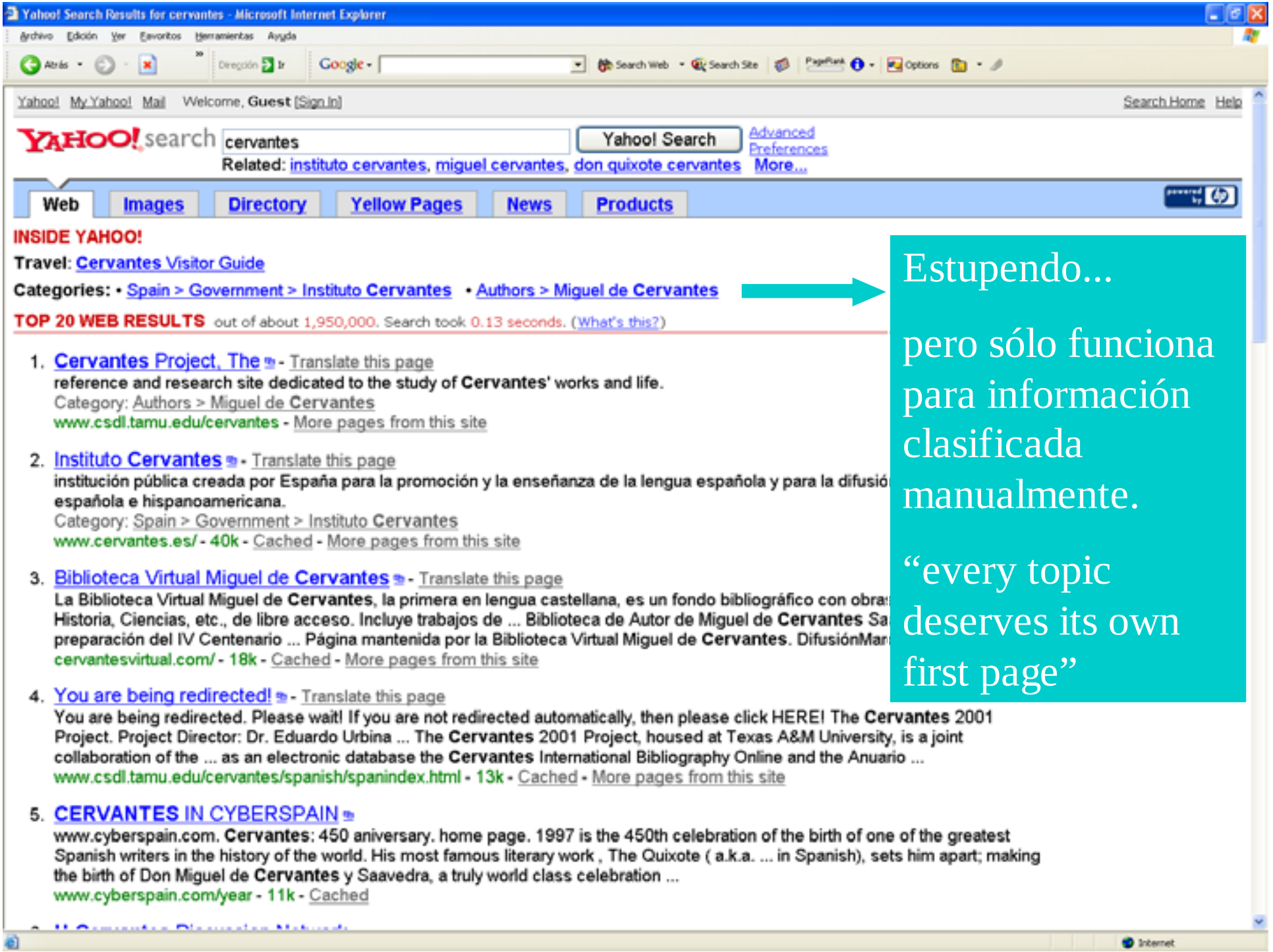
- Google nos "abandona" una vez que entramos en un portal

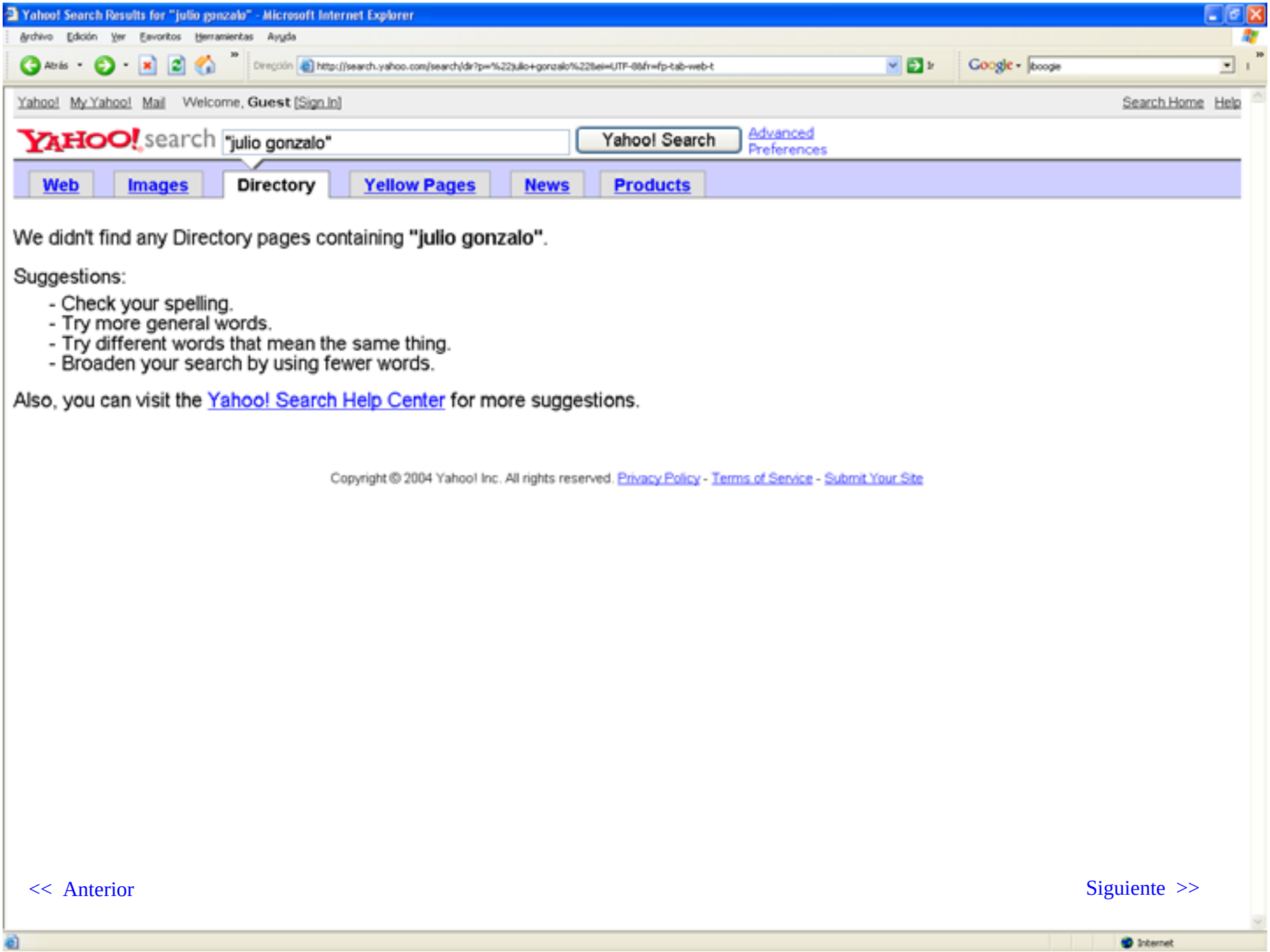
  → **3. Buscadores sobre portales**

1. Organización y visualización de los resultados de búsqueda

- Sugerencia de filtros: *Altavista, Teoma*
- Agrupación jerárquica: *Vivísimo, iBoogie*
- Agrupación jerárquica + Visualización: *Grokker*
- Resúmen multidocumento

Siguiente >>

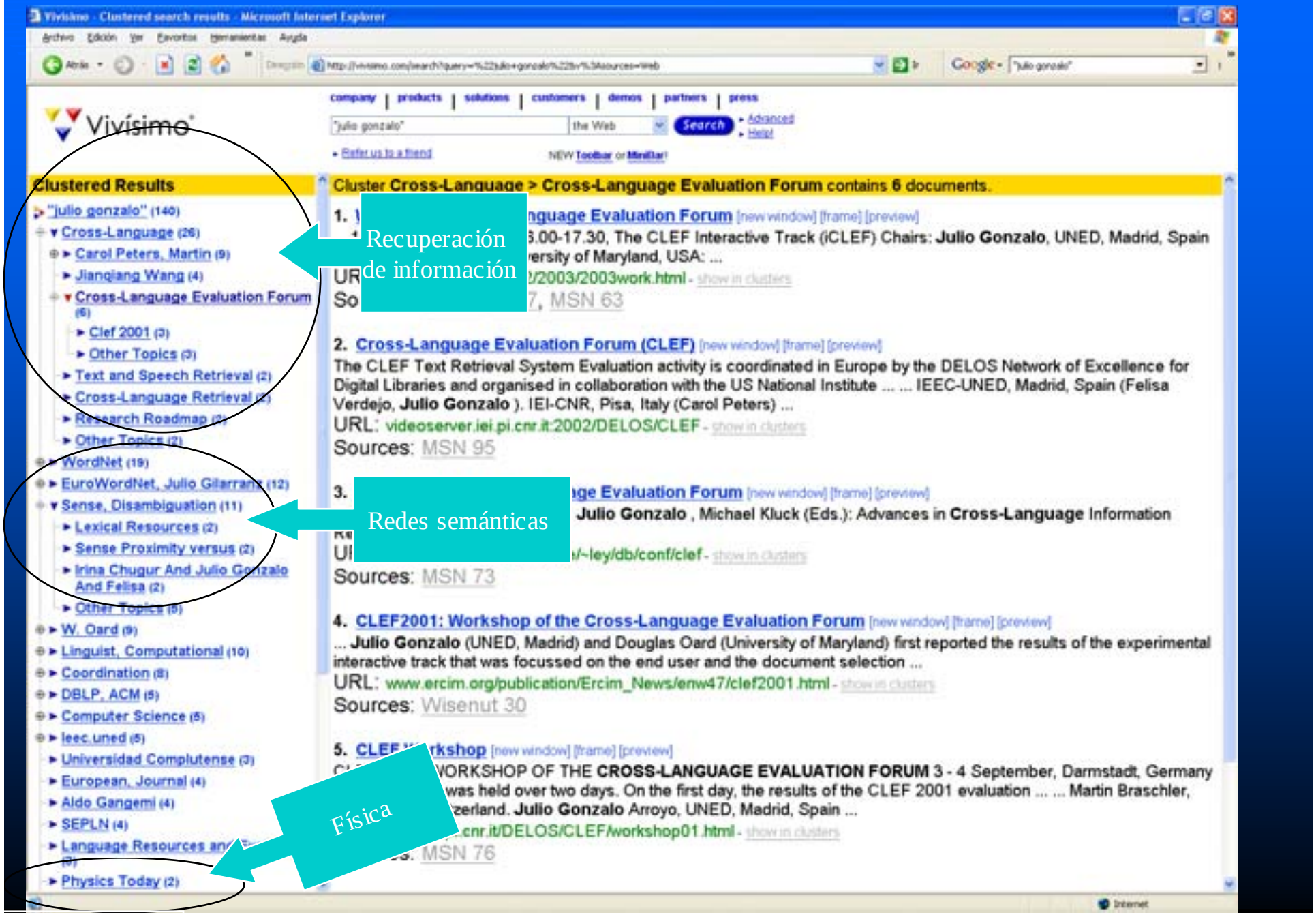

Yahoo! Search Results for cervantes - Microsoft Internet Explorer
Yahoo! My Yahoo! Mail Welcome, Guest [Sign In]
Search Home Help
YAHOO! search
cervantes
Yahoo! Search
Advanced
Preferences
Related: instituto cervantes, miguel cervantes, don quixote cervantes More...
Web
Images
Directory
Yellow Pages
News
Products
INSIDE YAHOO!
Travel: Cervantes Visitor Guide
Categories: • Spain > Government > Instituto Cervantes • Authors > Miguel de Cervantes
TOP 20 WEB RESULTS out of about 1,950,000. Search took 0.13 seconds. (What's this?)
1. Cervantes Project, The - Translate this page
reference and research site dedicated to the study of Cervantes' works and life.
Category: Authors > Miguel de Cervantes
www.csdl.tamu.edu/cervantes - More pages from this site
2. Instituto Cervantes - Translate this page
institución pública creada por España para la promoción y la enseñanza de la lengua española y para la difusió
española e hispanoamericana.
Category: Spain > Government > Instituto Cervantes
www.cervantes.es/ - 40k - Cached - More pages from this site
3. Biblioteca Virtual Miguel de Cervantes - Translate this page
La Biblioteca Virtual Miguel de Cervantes, la primera en lengua castellana, es un fondo bibliográfico con obra
Historia, Ciencias, etc., de libre acceso. Incluye trabajos de ... Biblioteca de Autor de Miguel de Cervantes Sa
preparación del IV Centenario ... Página mantenida por la Biblioteca Virtual Miguel de Cervantes. DifusiónMar
cervantesvirtual.com/ - 18k - Cached - More pages from this site
4. You are being redirected! - Translate this page
You are being redirected. Please wait! If you are not redirected automatically, then please click HERE! The Cervantes 2001 Project. Project Director: Dr. Eduardo Urbina ... The Cervantes 2001 Project, housed at Texas A&M University, is a joint collaboration of the ... as an electronic database the Cervantes International Bibliography Online and the Anuario ...
www.csdl.tamu.edu/cervantes/spanish/spanindex.html - 13k - Cached - More pages from this site
5. CERVANTES IN CYBERSPAIN
www.cyberspain.com. Cervantes: 450 aniversary. home page. 1997 is the 450th celebration of the birth of one of the greatest Spanish writers in the history of the world. His most famous literary work , The Quixote (a.k.a. ... in Spanish), sets him apart; making the birth of Don Miguel de Cervantes y Saavedra, a truly world class celebration ...
www.cyberspain.com/year - 11k - Cached
Estupendo...
pero sólo funciona para información clasificada manualmente.
“every topic deserves its own first page”


Yahoo! Search Results for "julio gonzalo" - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás
Dirección http://search.yahoo.com/search/dir?p=%22julio+gonzalo%22&ei=UTF-8&fr=fp-tab-web-t
Ir
Google
boogie
Yahoo! My Yahoo! Mail Welcome, Guest [Sign In]
Search Home Help
YAHOO! search
"julio gonzalo"
Yahoo! Search
Advanced
Preferences
Web
Images
Directory
Yellow Pages
News
Products
We didn't find any Directory pages containing "julio gonzalo".
Suggestions:
- Check your spelling.
- Try more general words.
- Try different words that mean the same thing.
- Broaden your search by using fewer words.
Also, you can visit the Yahoo! Search Help Center for more suggestions.
Copyright © 2004 Yahoo! Inc. All rights reserved. Privacy Policy - Terms of Service - Submit Your Site
<< Anterior
Siguiente >>
Internet

Búsqueda: "Julio Gonzalo":

- informático: 69 enlaces
- físico: 7 enlaces
- coronel: 1 enlace
- alcalde: 1enlace

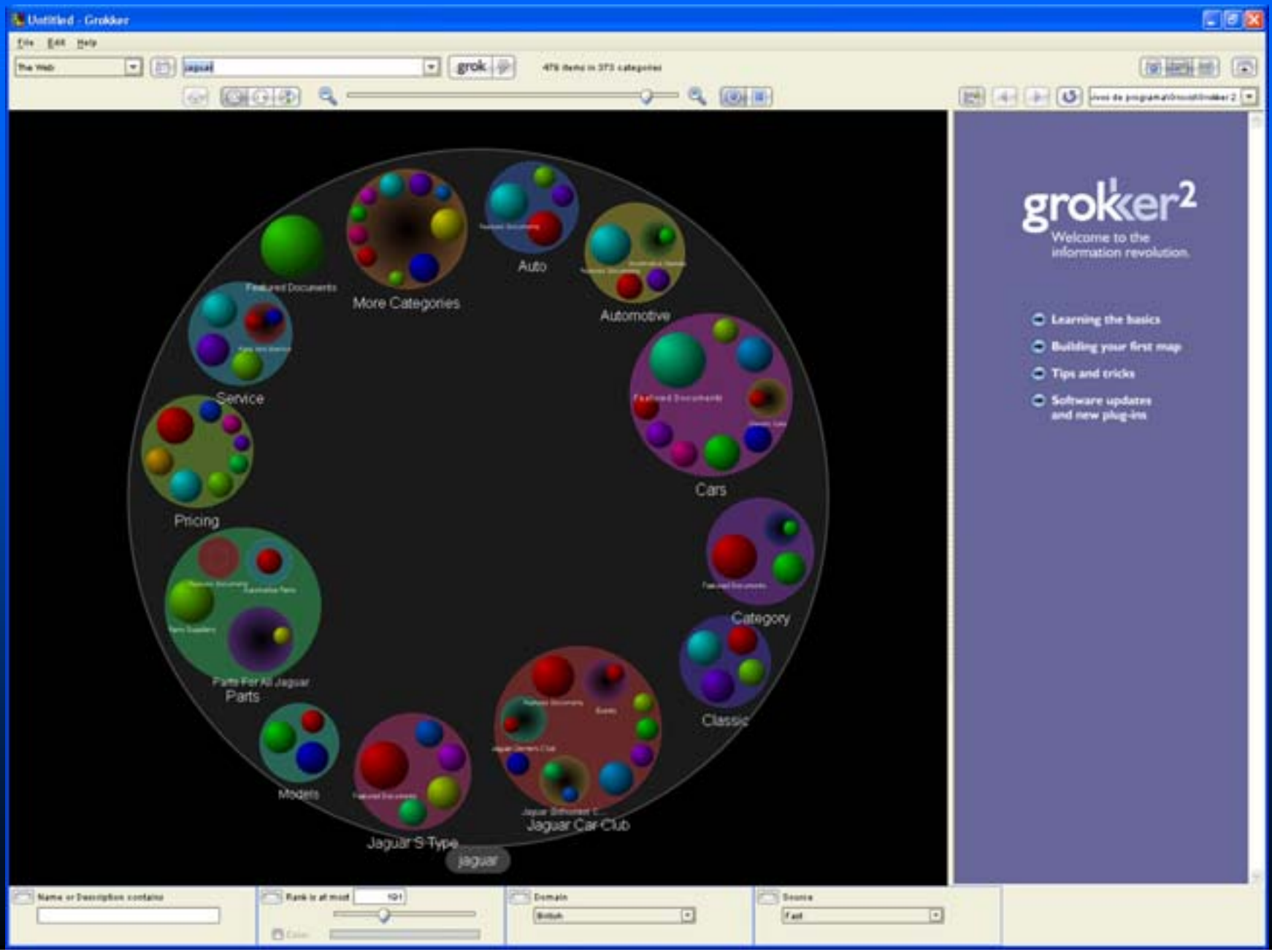
Agrupación jerárquica: *Vivísimo*

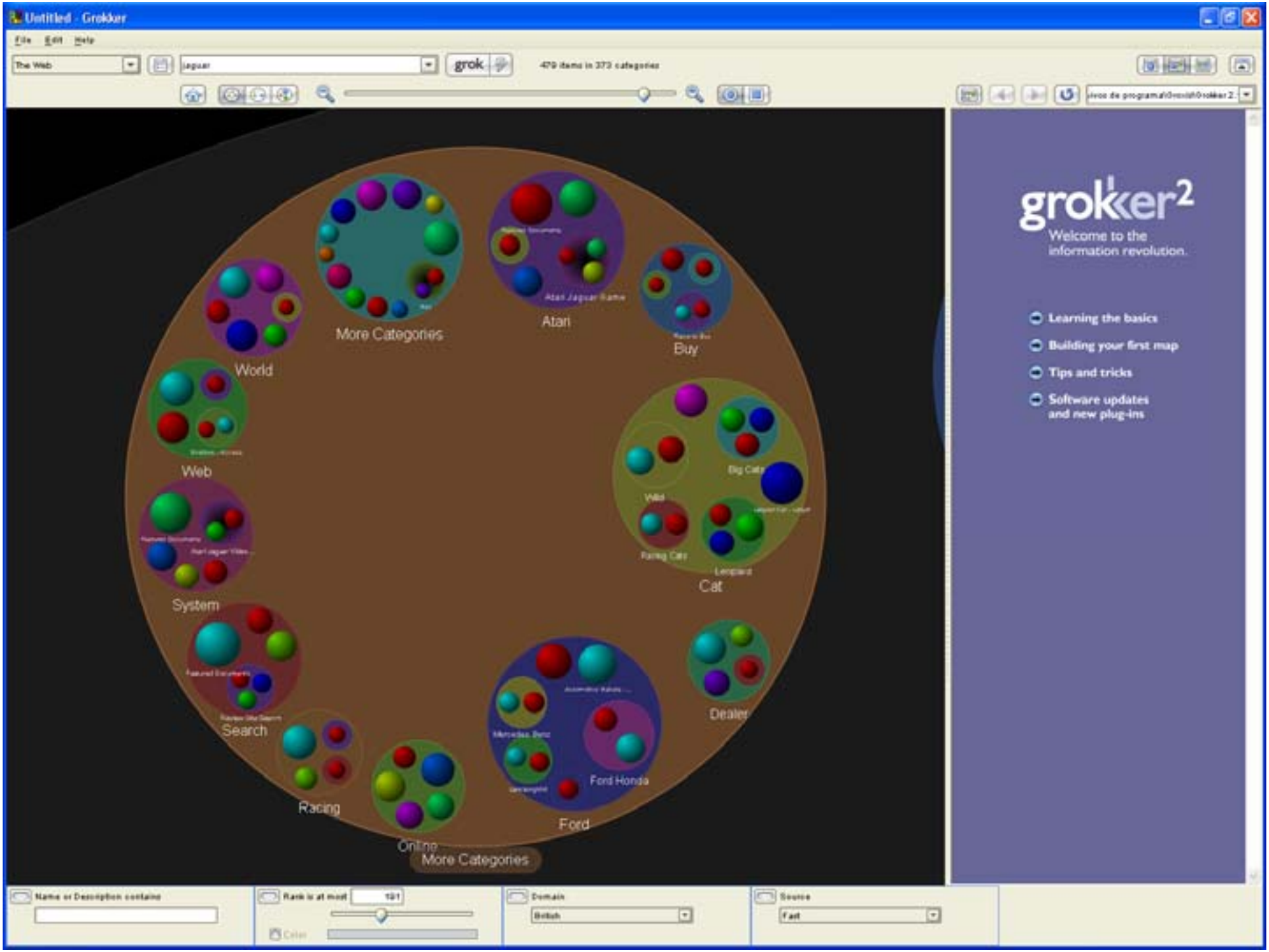


<< Anterior

Siguiente >>

Agrupación jerárquica + visualización: *Grokker*



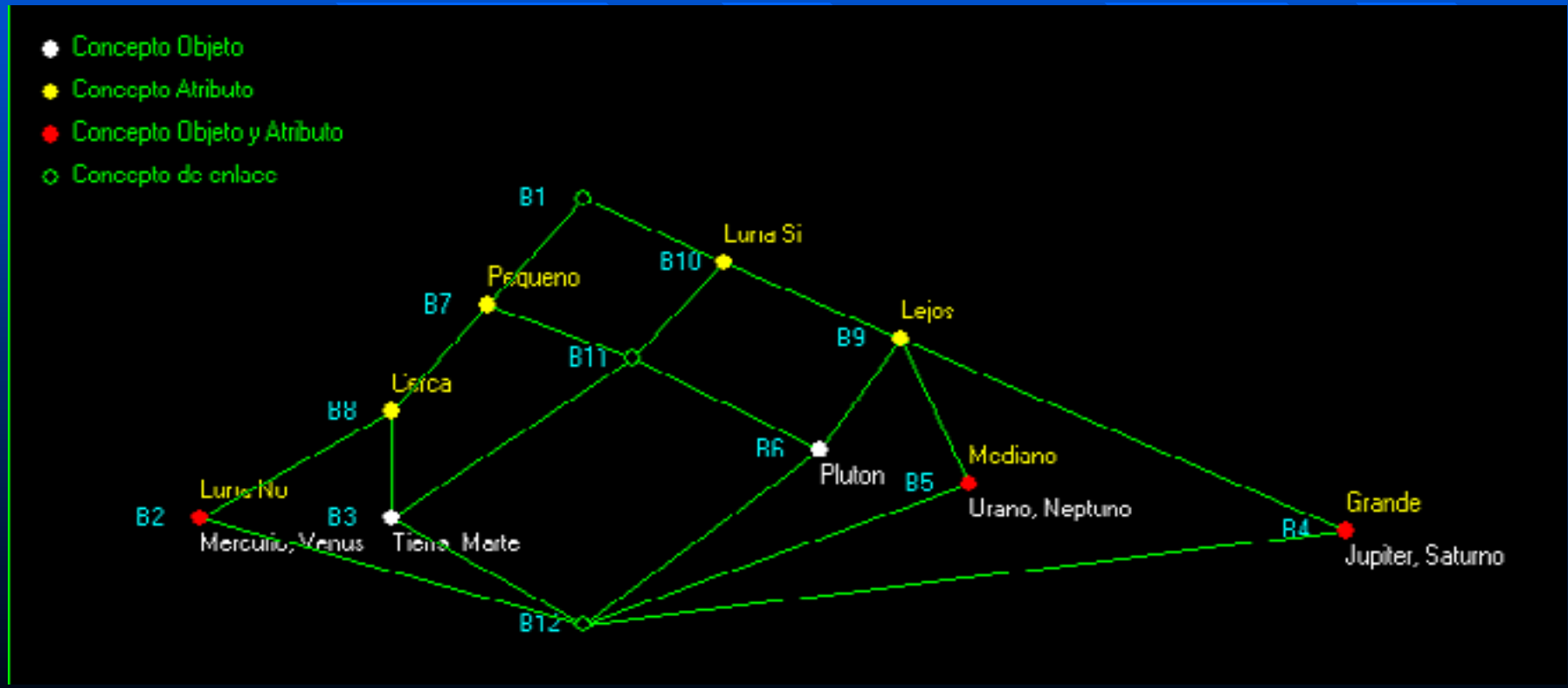

Untitled - Grokker
File Edit Help
The Web
jaguar
grok
479 items in 373 categories
More Categories
Atari
Buy
World
Web
System
Search
Racing
Online
Ford
Ford Honda
Dealer
Cat
More Categories
grokker2
Welcome to the information revolution.
Learning the basics
Building your first map
Tips and tricks
Software updates and new plug-ins
Name or Description contains
Rank is at most
Domain
Source

Análisis Formal de Conceptos

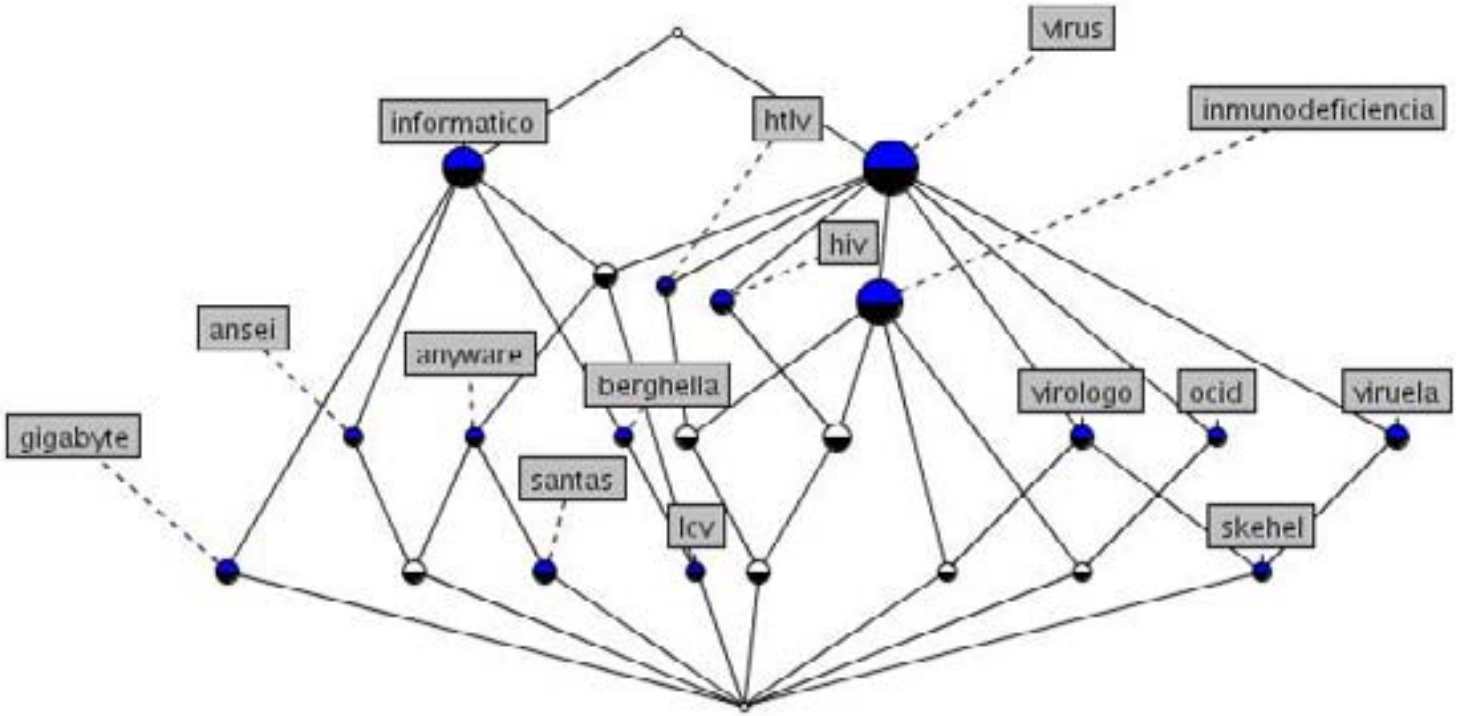
Planetas	Pequeno	Mediano	Grande	Cerca	Lejos	Luna Nu	Luna Si
Mercurio	X			X		X	
Venus	X			X		X	
Tierra	X			X			X
Marte	X			X			X
Jupiter			X		X		X
Saturno			X		X		X
Urano		X			X		X
Neptuno		X			X		X
Pluton	X				X		X

Matriz objetos/atributos

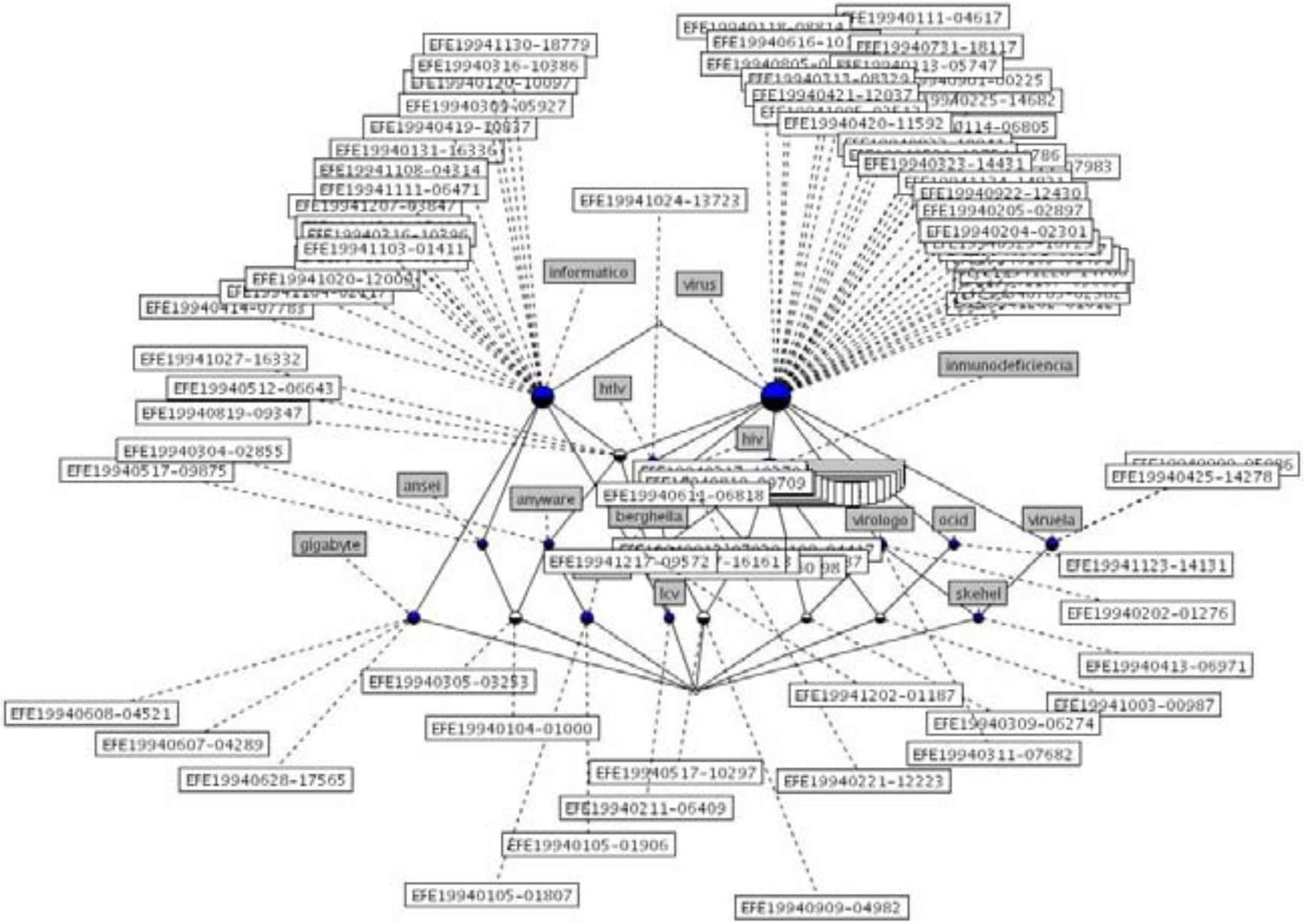
Retículo conceptual



UNED JBraindead: exploración mediante retículos



(consulta: virus informáticos)


EFE19940111-04617
EFE19940731-18117
EFE19940225-14682
EFE19940323-14431
EFE19940922-12430
EFE19940205-02897
EFE19940204-02301
EFE19941130-18779
EFE19940316-10386
EFE19940131-16336
EFE19941108-04314
EFE19941111-06471
EFE19941103-01411
EFE19941020-12000
EFE19941024-13723
informatico
virus
inmunodeficiencia
EFE19941027-16332
EFE19940512-06643
EFE19940819-09347
EFE19940304-02855
EFE19940517-09875
EFE19940425-14278
ansel
amyware
EFE19940611-06818
berghella
virologo
ocid
viruela
gigabyte
EFE19941123-14131
skehel
EFE19940202-01276
EFE19940413-06971
EFE19940305-03253
EFE19941202-01187
EFE19941003-00987
EFE19940309-06274
EFE19940608-04521
EFE19940607-04289
EFE19940628-17565
EFE19940104-01000
EFE19940517-10297
EFE19940311-07682
EFE19940221-12223
EFE19940211-06409
EFE19940105-01906
EFE19940105-01807
EFE19940909-04982

2. Buscadores verticales

Necesidad
De información

Colección de
documentos

consulta

extracción

datos

2. Buscadores verticales

- Agrupación: Google News
- Extracción: Citeseer (publicaciones científicas)
- TrypAdvisor (viajes)
- Froogle (compras)

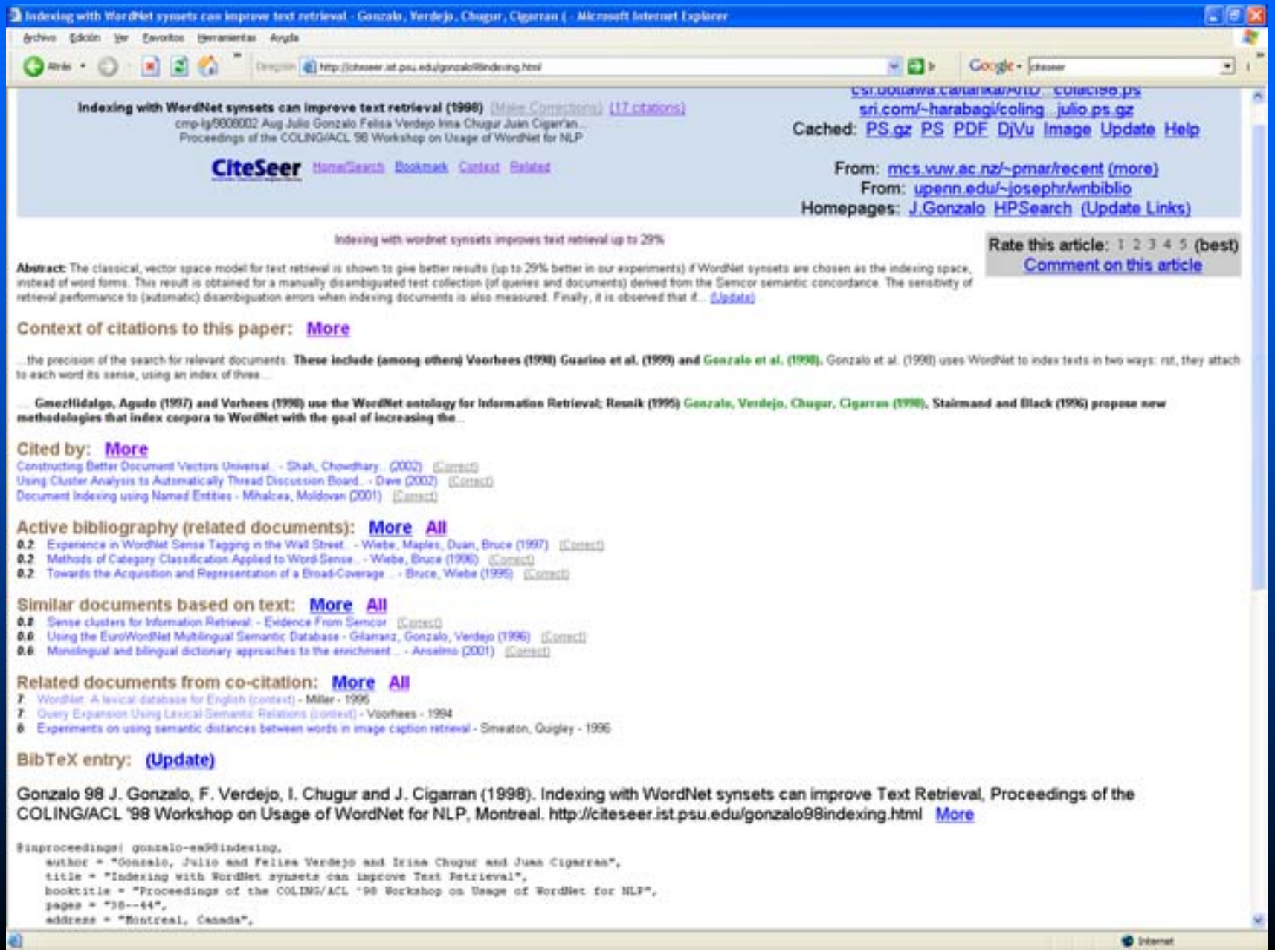
Filtrado y agrupación de resultados en dominio específico

<< Anterior

Siguiente >>

Extracción de información: compras

Extracción + minería: Citeseer



El complementario de los buscadores verticales: la *Web Semántica*

Necesidad
De información

Web semántica

pregunta

QL

Agente de búsqueda

RDF

datos, *gestiones*

ontologías

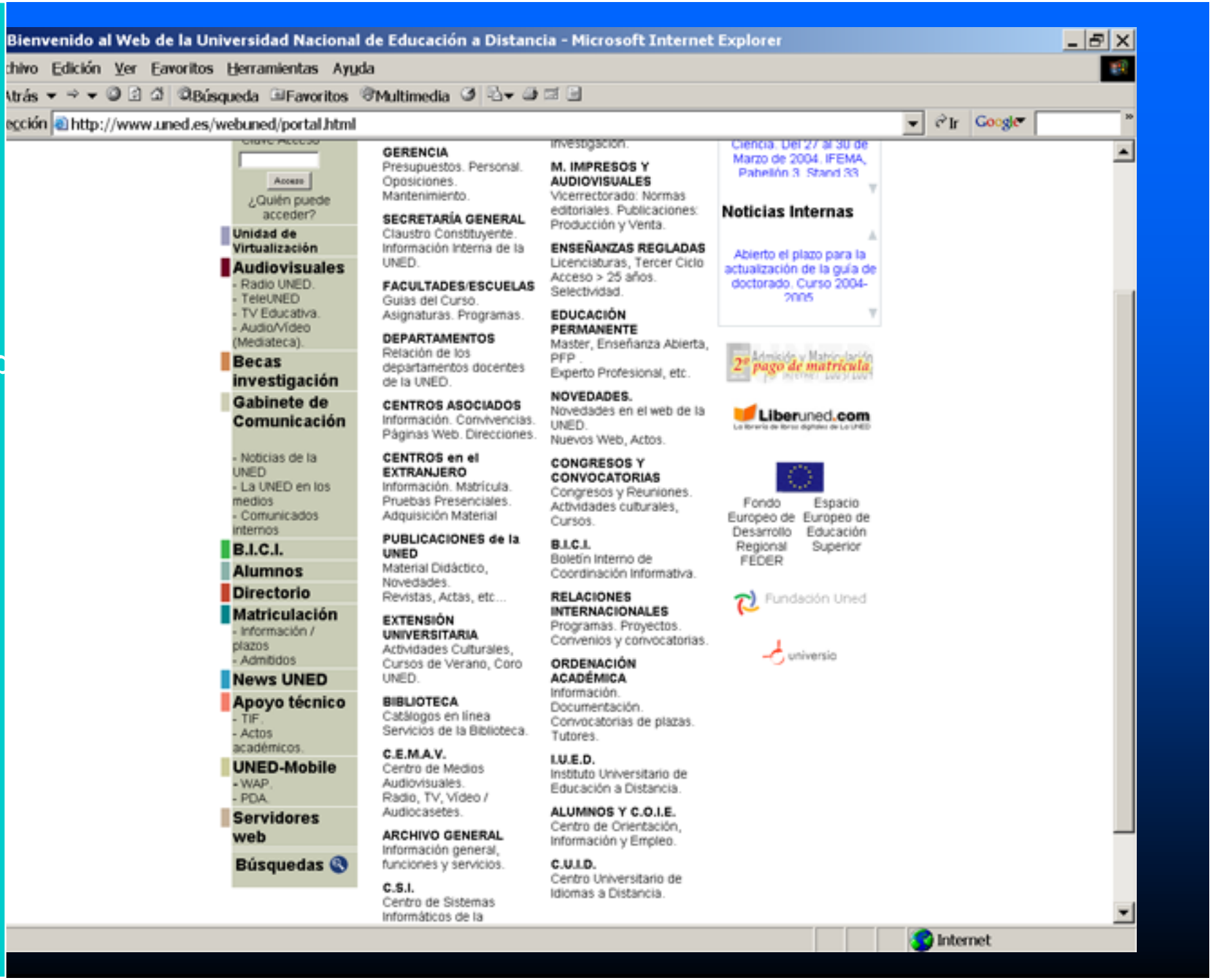
3. Buscadores sobre portales

- Ejemplo de búsqueda/navegación en portal
- Asistentes virtuales de búsqueda
- Visualizadores: Hyperbrowsing
- Uso de metadatos
- Cha-Cha (visualizacion de caminos)

Ejemplo: busco impreso solicitud reintegro ADSL en uned.es

- Primero veremos cómo se llega navegando.
- Después cómo se puede llegar buscando.
- Ambas opciones son catastróficas, a pesar de que se trata de una web premiada y un buscador por encima de la media.

¡No aparece el vicerrectorado de nuevas tecnologías!



<< Anterior

Siguiente >>

Había que buscar en el BICI...

página principal - Microsoft Internet Explorer

chivo Edición Ver Favoritos Herramientas Ayuda

Atrás Búsqueda Favoritos Multimedia

ección http://www.uned.es/bici/ Ir Google

Universidad Nacional de Educación a Distancia

Audiovisuales | Becas | Noticias | Alumnos | Directorio | Matriculación | News UNED

Boletín Interno de Coordinación Informativa

Sección de Información

El Boletín Interno de Coordinación Informativa (B.I.C.I.) es una publicación institucional, semanal, de difusión gratuita, en la que se incluyen noticias generadas por los distintos órganos de la Universidad para conocimiento de la Comunidad Universitaria.
Podrá consultarse a través de esta página Web o directamente en las Unidades de Atención al Público y en los Centros Asociados de la UNED.
Este Boletín se publica todos los lunes lectivos del curso académico.

Nº de visitas: 38527

Curso 2003/2004

Curso 2002/2003

Curso 2001/2002

Curso 2000/2001

Curso 1999/2000

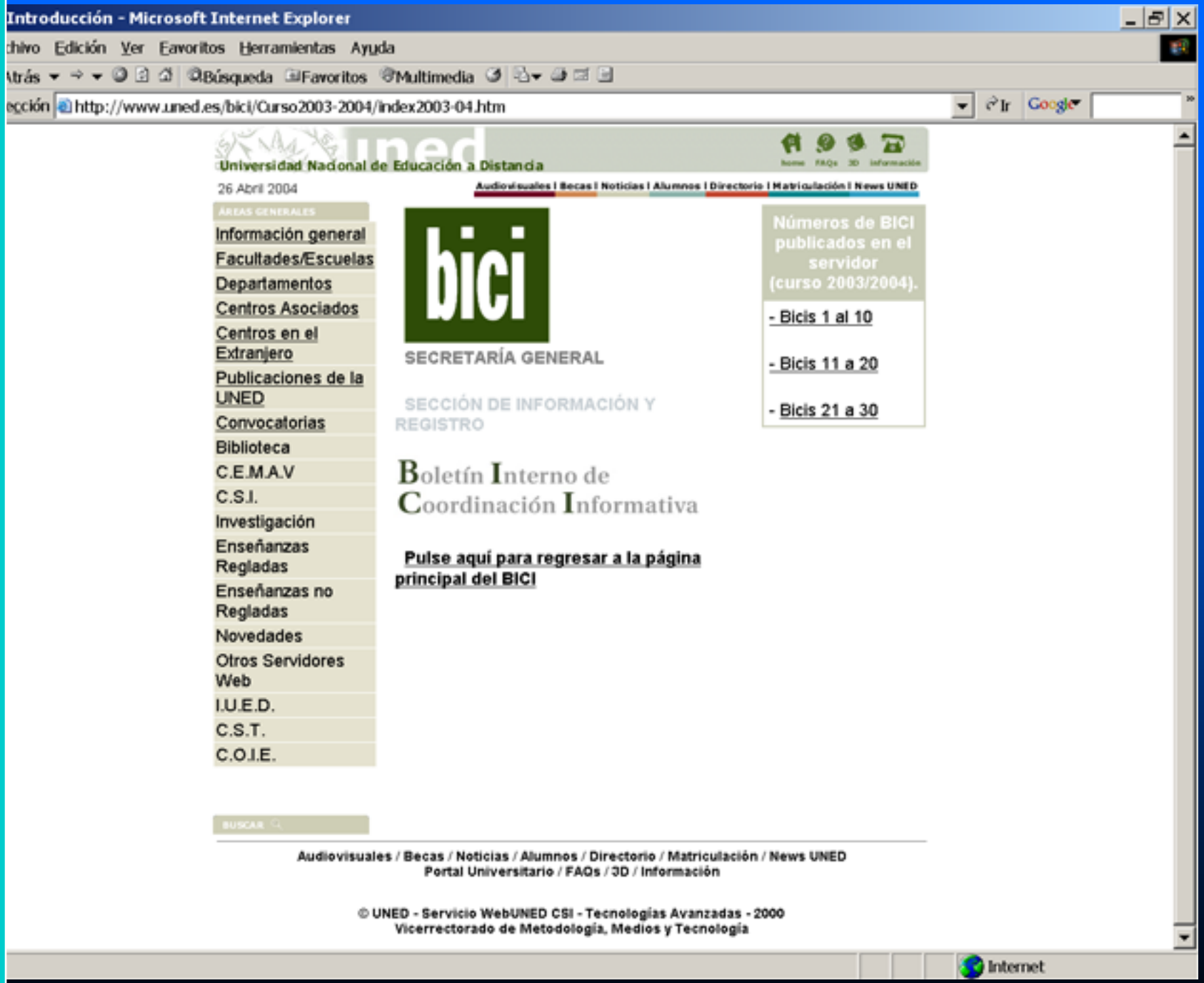
Curso 1998/1999

Curso 1997/1998

Reglamento del Bici

Internet

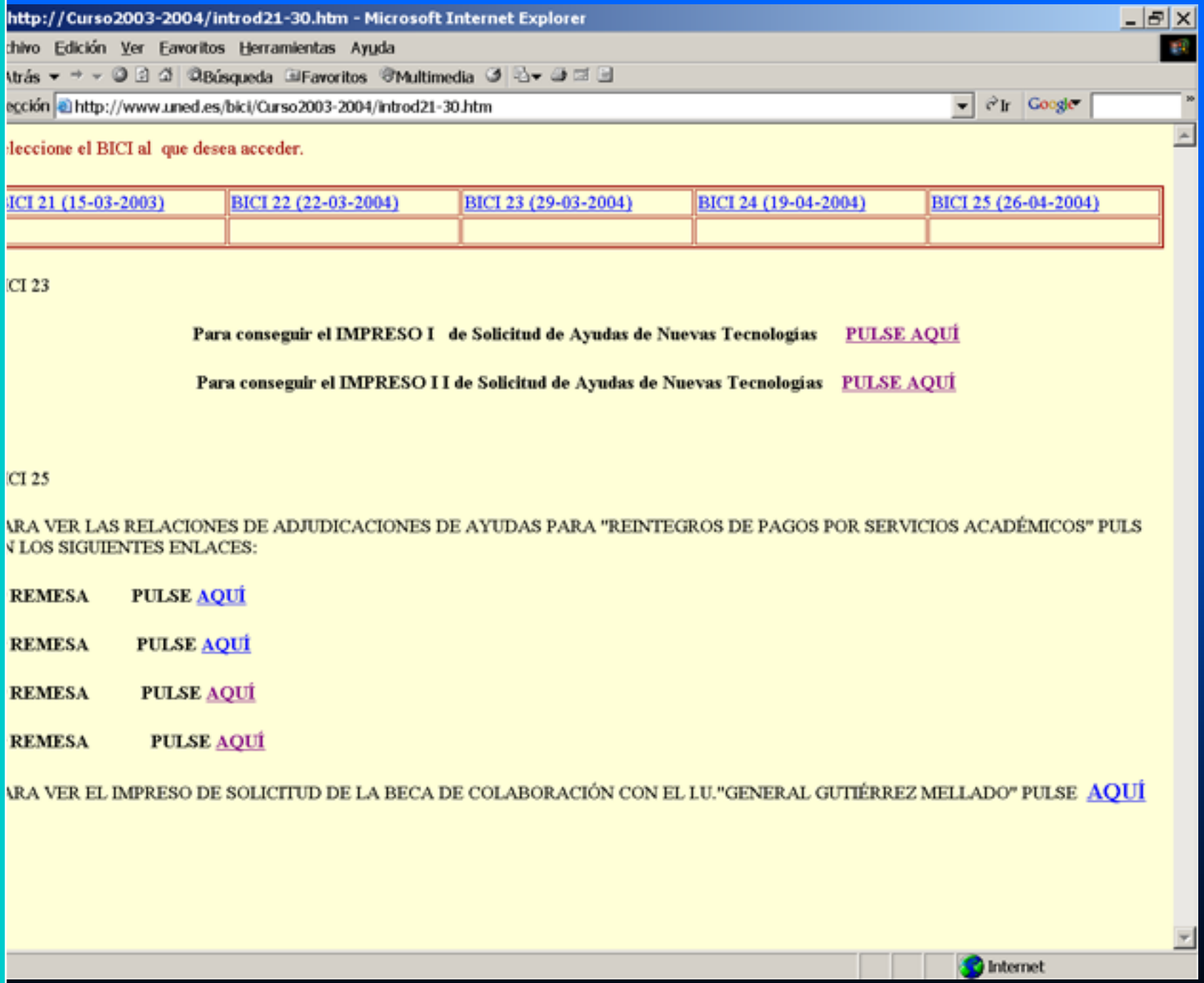
Curso 2003/2004...



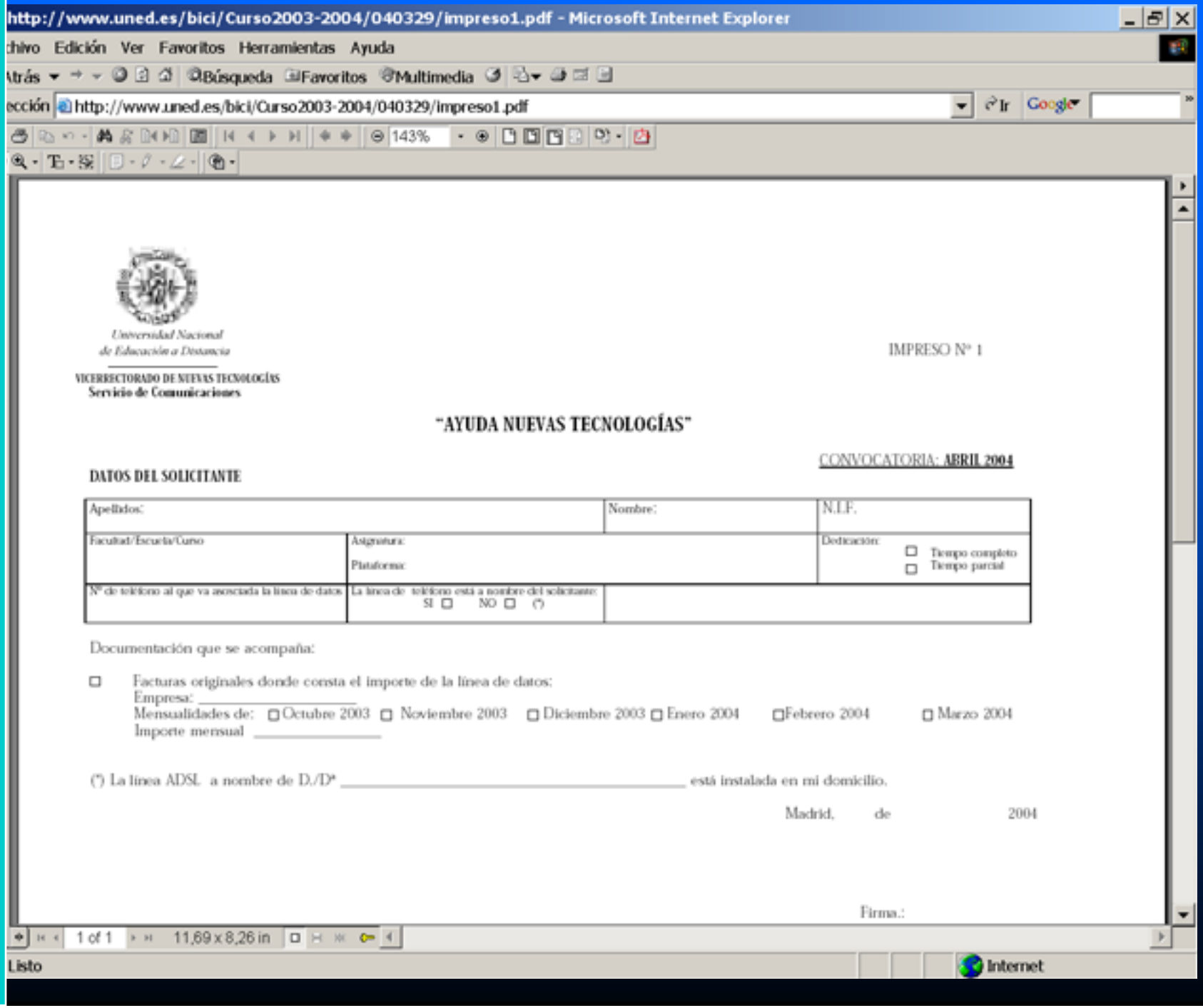
<< Anterior

Siguiente >>

BICIs 21 a 30...


http://Curso2003-2004/introd21-30.htm - Microsoft Internet Explorer
chivo Edición Ver Favoritos Herramientas Ayuda
Atrás Búsqueda Favoritos Multimedia
ección http://www.uned.es/bici/Curso2003-2004/introd21-30.htm
Ir Google
leccione el BICI al que desea acceder.
ICI 21 (15-03-2003) BICI 22 (22-03-2004) BICI 23 (29-03-2004) BICI 24 (19-04-2004) BICI 25 (26-04-2004)
ICI 23
Para conseguir el IMPRESO I de Solicitud de Ayudas de Nuevas Tecnologías PULSE AQUÍ
Para conseguir el IMPRESO I I de Solicitud de Ayudas de Nuevas Tecnologías PULSE AQUÍ
ICI 25
ARA VER LAS RELACIONES DE ADJUDICACIONES DE AYUDAS PARA "REINTEGROS DE PAGOS POR SERVICIOS ACADÉMICOS" PULS
N LOS SIGUIENTES ENLACES:
REMESA PULSE AQUÍ
REMESA PULSE AQUÍ
REMESA PULSE AQUÍ
REMESA PULSE AQUÍ
ARA VER EL IMPRESO DE SOLICITUD DE LA BECA DE COLABORACIÓN CON EL LU."GENERAL GUTIÉRREZ MELLADO" PULSE AQUÍ
Internet

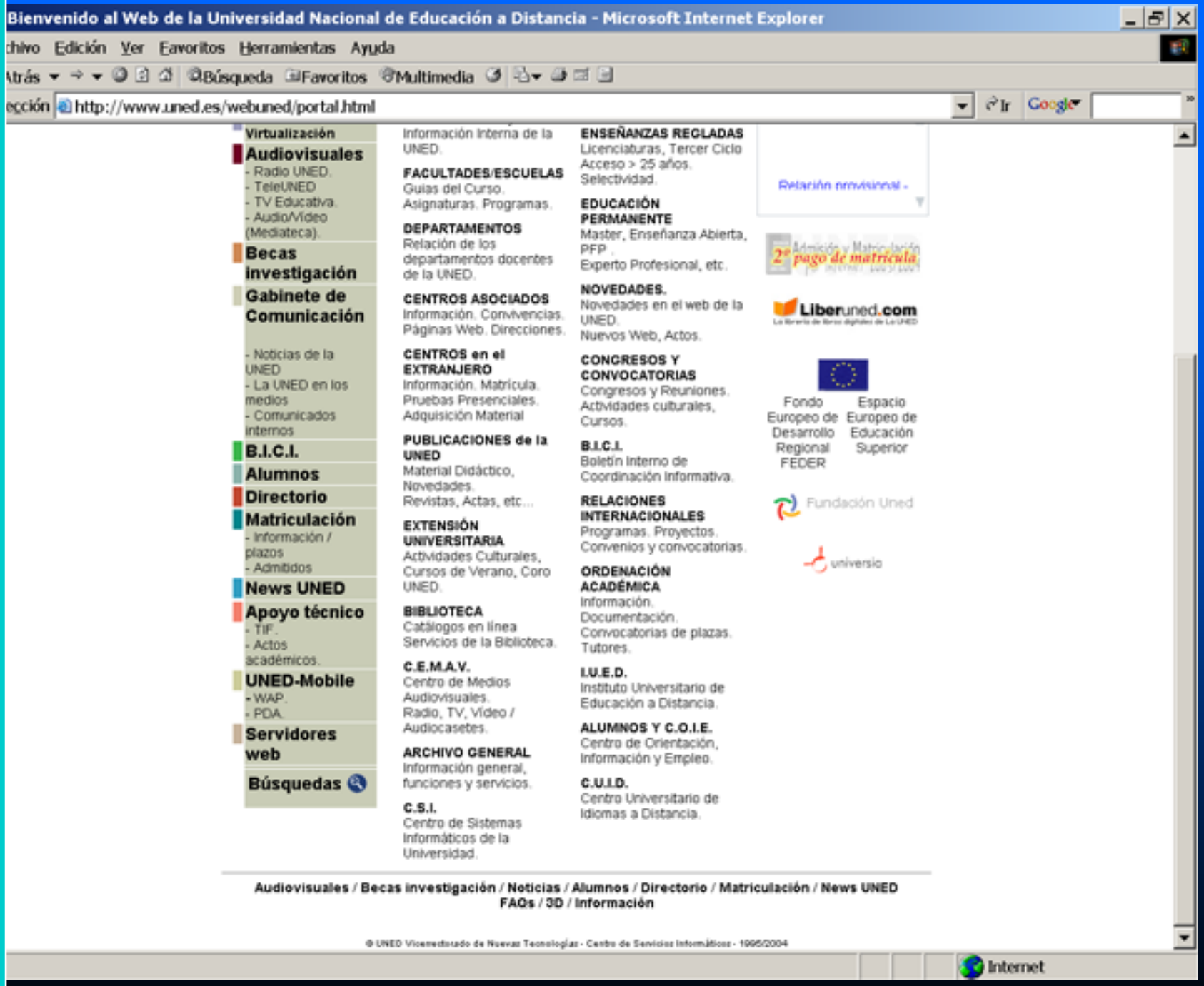
Al fin!



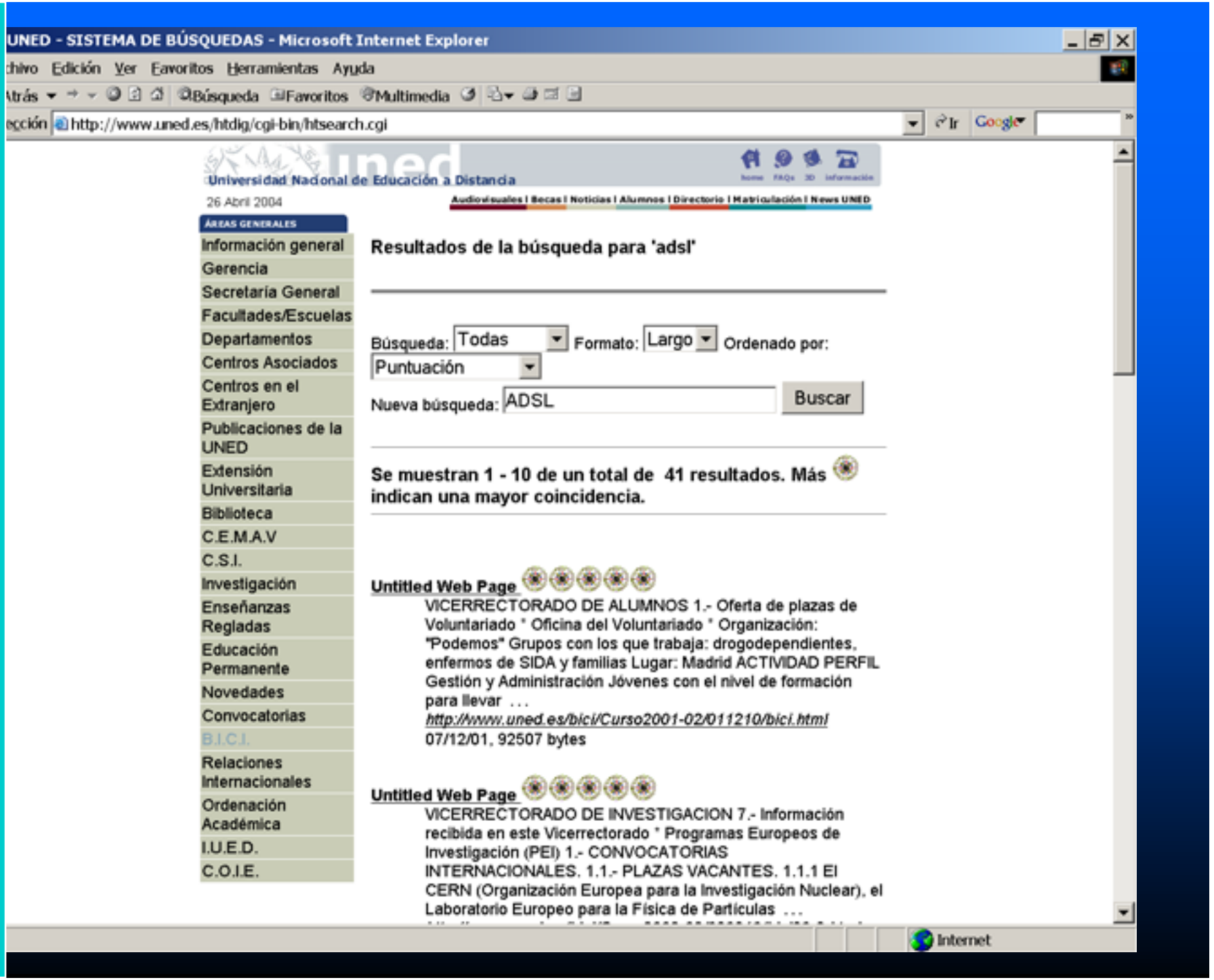
<< Anterior

Siguiente >>

Hay una opción de búsqueda



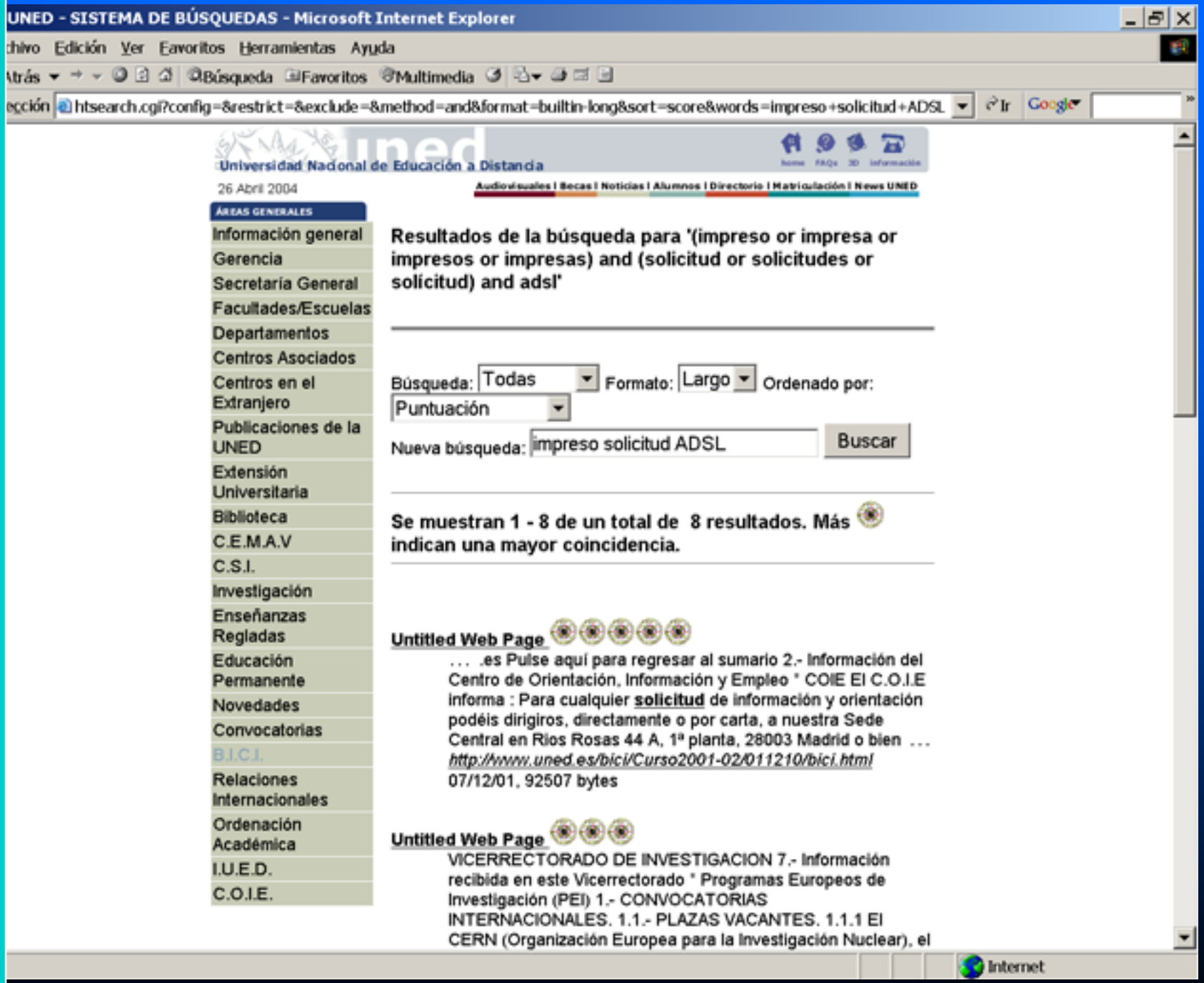
Busco "ADSL"



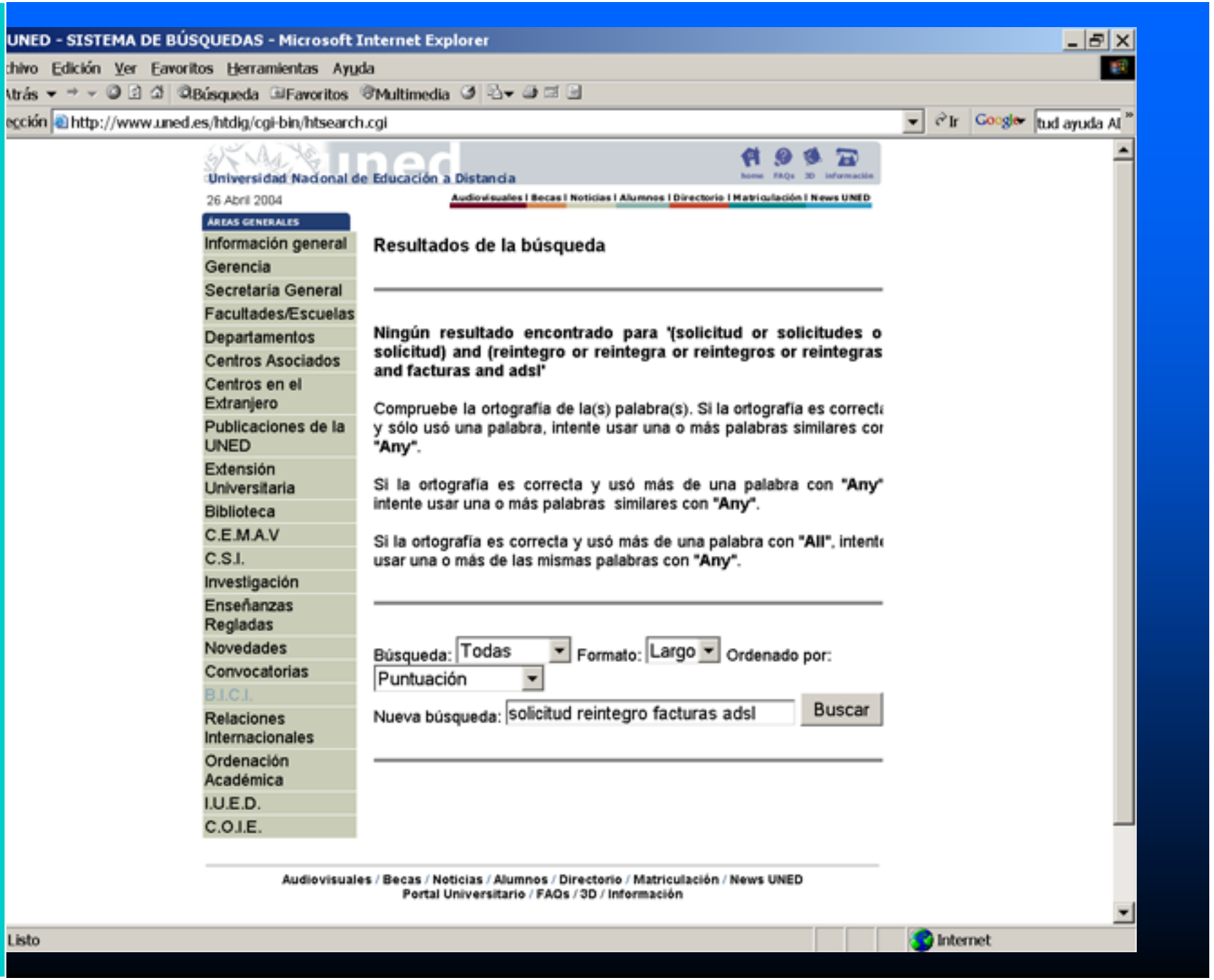
<< Anterior

Siguiente >>

Busco “impreso solicitud ADSL”



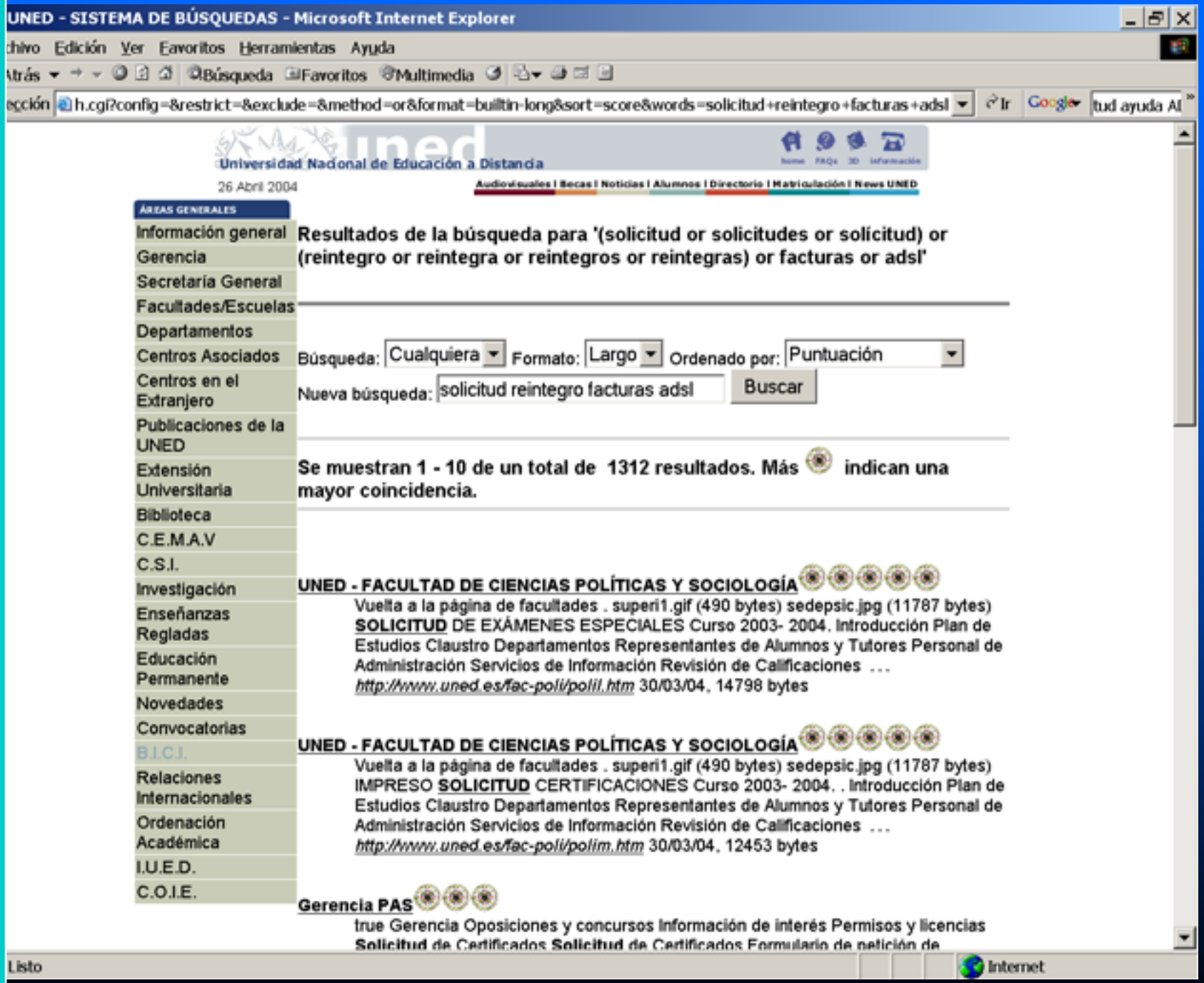
Busco “solicitud reintegro facturas ADSL”



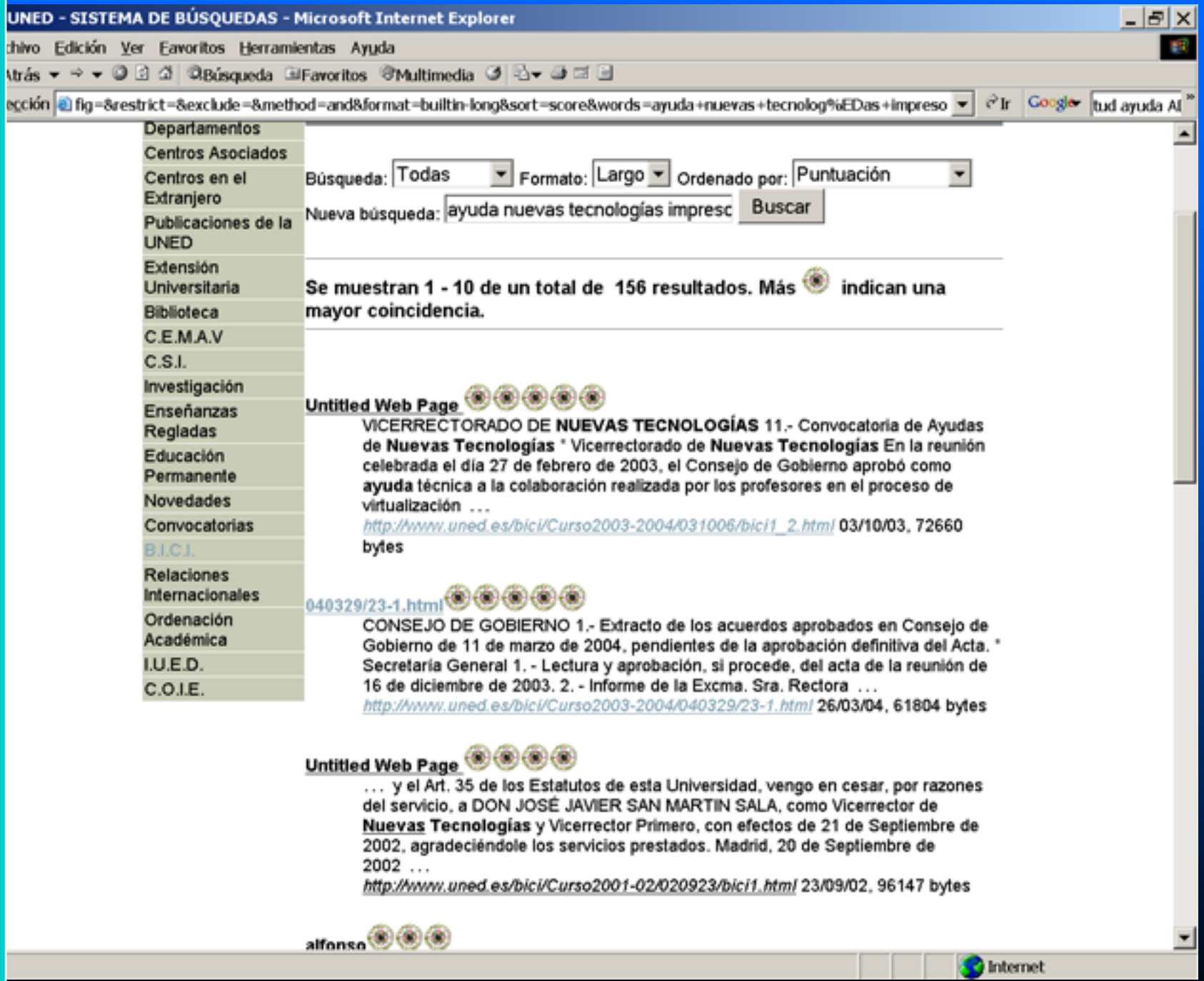
<< Anterior

Siguiente >>

Cambio “todas” por “cualquiera”

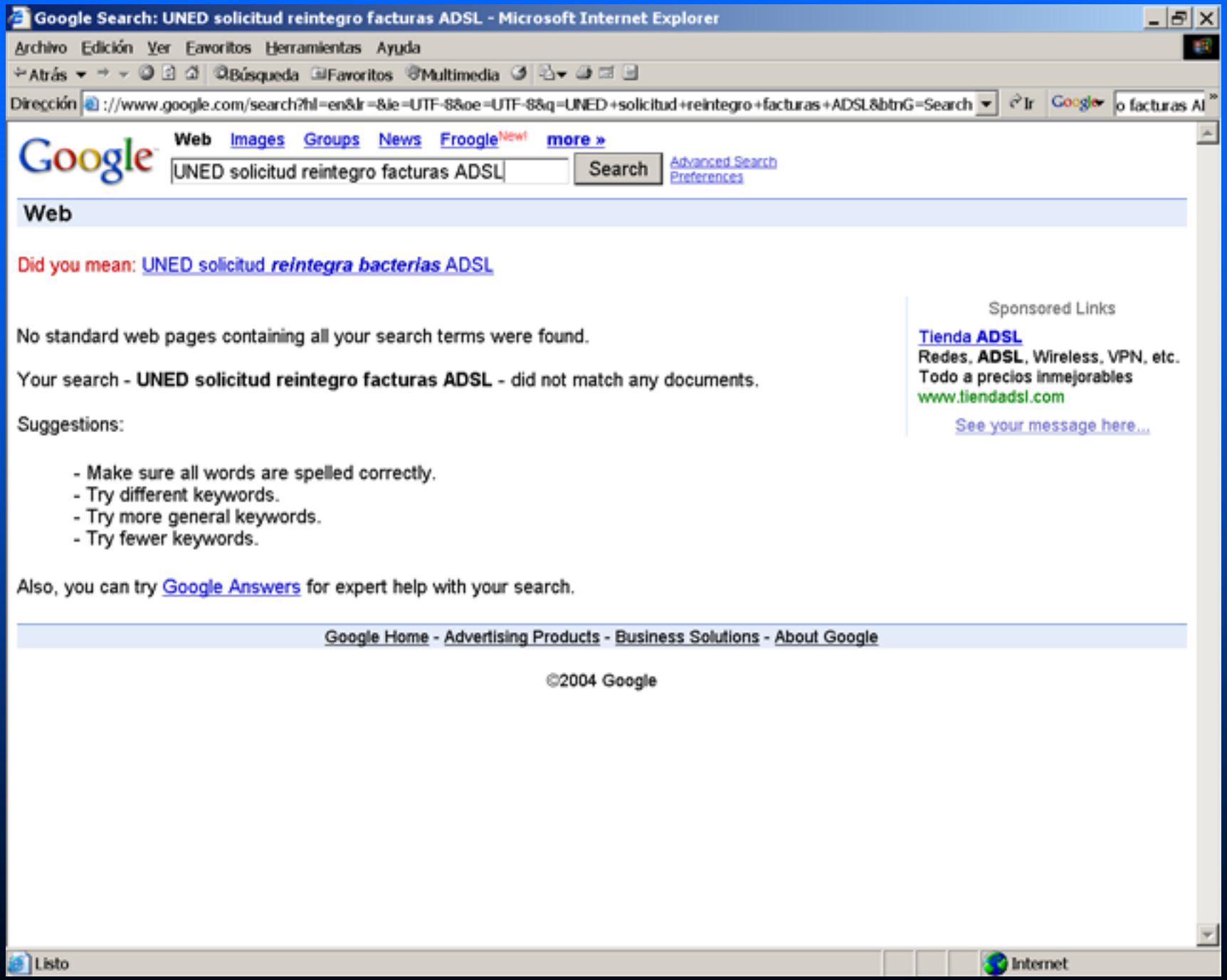


Uso las palabras clave del documento (lo conozco)



<< Anterior

Siguiente >>


Google Search: UNED solicitud reintegro facturas ADSL - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Búsqueda Favoritos Multimedia
Dirección ://www.google.com/search?hl=en&lr=&ie=UTF-8&oe=UTF-8&q=UNED+solicitud+reintegro+facturas+ADSL&btnG=Search
Ir
Google
o facturas Al
Google
Web Images Groups News Froogle New! more »
UNED solicitud reintegro facturas ADSL
Search
Advanced Search
Preferences
Web
Did you mean: UNED solicitud reintegra bacterias ADSL
No standard web pages containing all your search terms were found.
Your search - UNED solicitud reintegro facturas ADSL - did not match any documents.
Suggestions:
- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.
Also, you can try Google Answers for expert help with your search.
Sponsored Links
Tienda ADSL
Redes, ADSL, Wireless, VPN, etc.
Todo a precios inmejorables
www.tiendadsl.com
See your message here...
Google Home - Advertising Products - Business Solutions - About Google
©2004 Google
Listo
Internet

Buena impresión si la respuesta está en el FAQ...

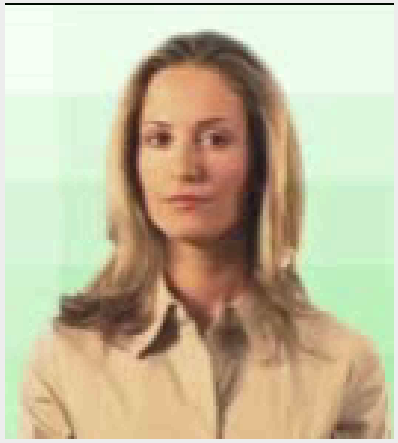
Asistentes virtuales: CajaMadrid

Pero ¿cómo hacerse entender cuando se equivoca?

Caja Madrid - Asistente Virtual - Micros...

> Ayuda > Cerrar

Pregunte a su
Asistente Virtual



creado por Avanza - Verity Response

> Aceptar

¿Puedo cambiar el plazo de amortización de mi hipoteca?

Sí, podrá hacer una subrogación de hipoteca cuando lo desee. Consúltenos en nuestras oficinas o llámenos, y le resolveremos todas las dudas que se le planteen. Le convendrá llevarla a cabo cuando el beneficio obtenido con el cambio

Siga de cerca su cartera de valores.
Consulte las **Cotizaciones** y el **Gráfico Interactivo**

Valore la última respuesta: Mala Aceptable Buena > Enviar

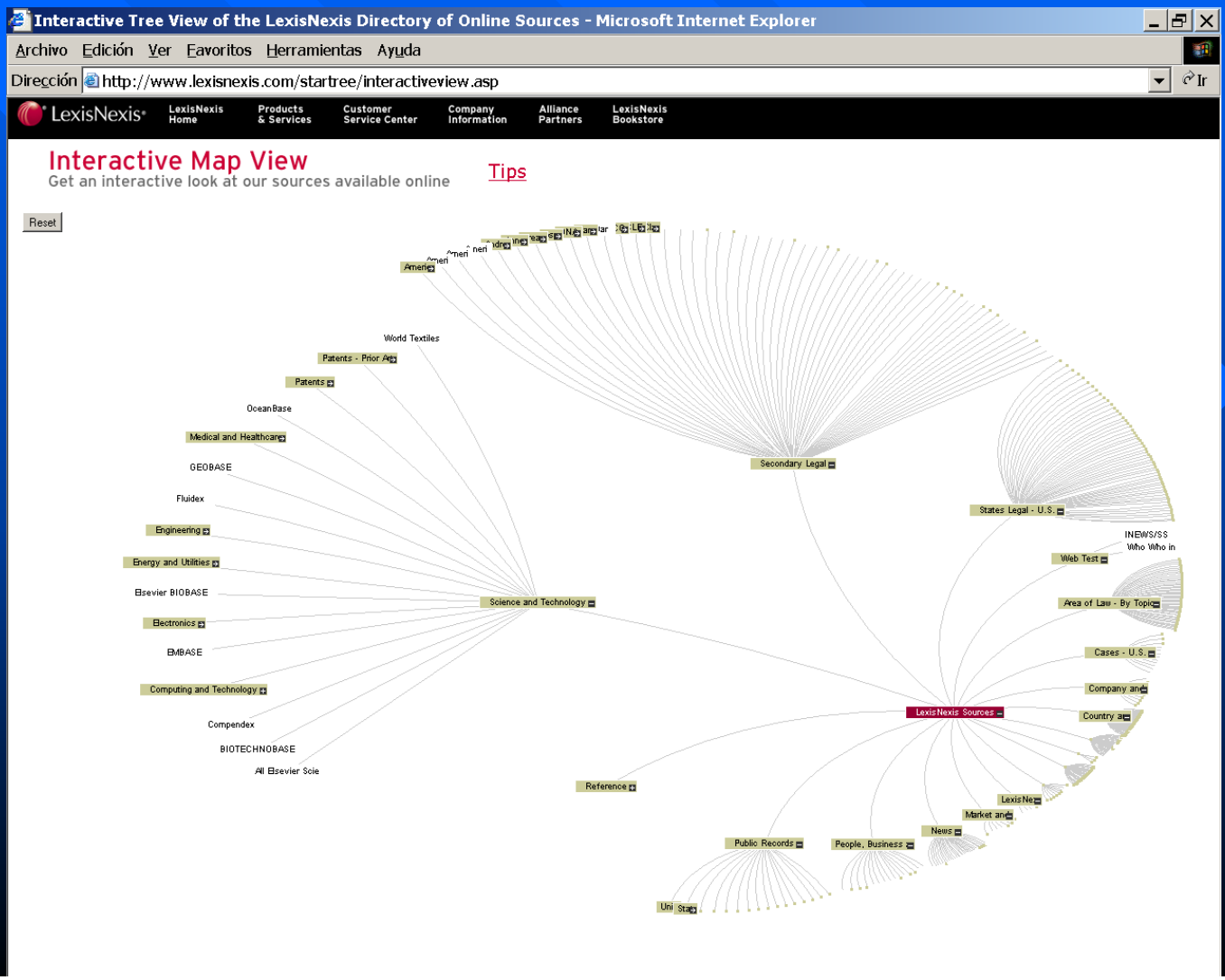
Condiciones de uso

<< Anterior

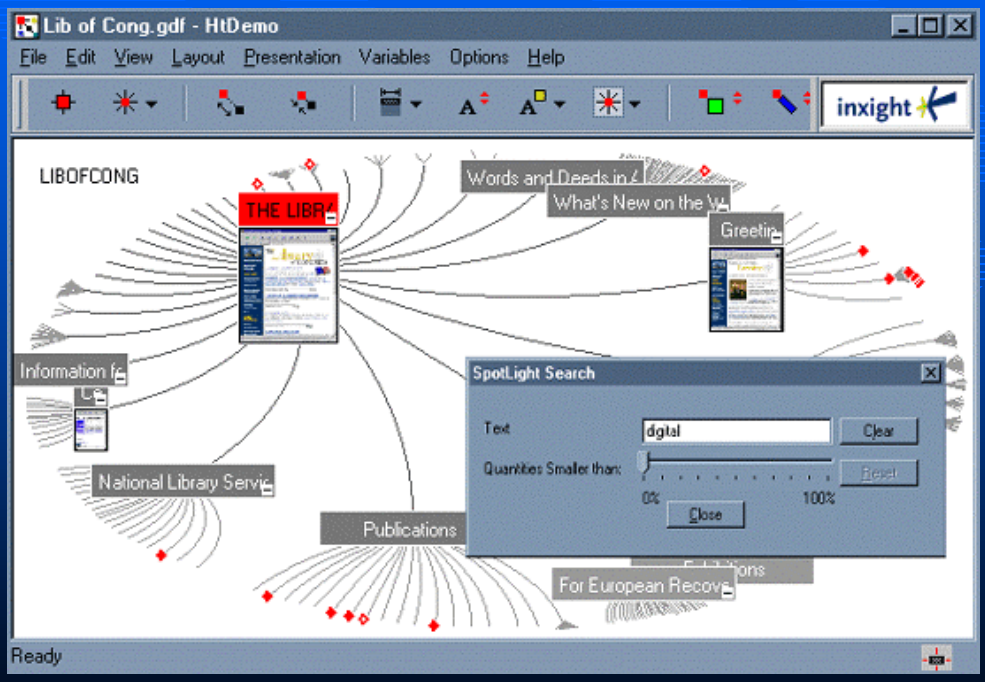
Siguiente >>

Los asistentes virtuales le dan a la bebida virtual...

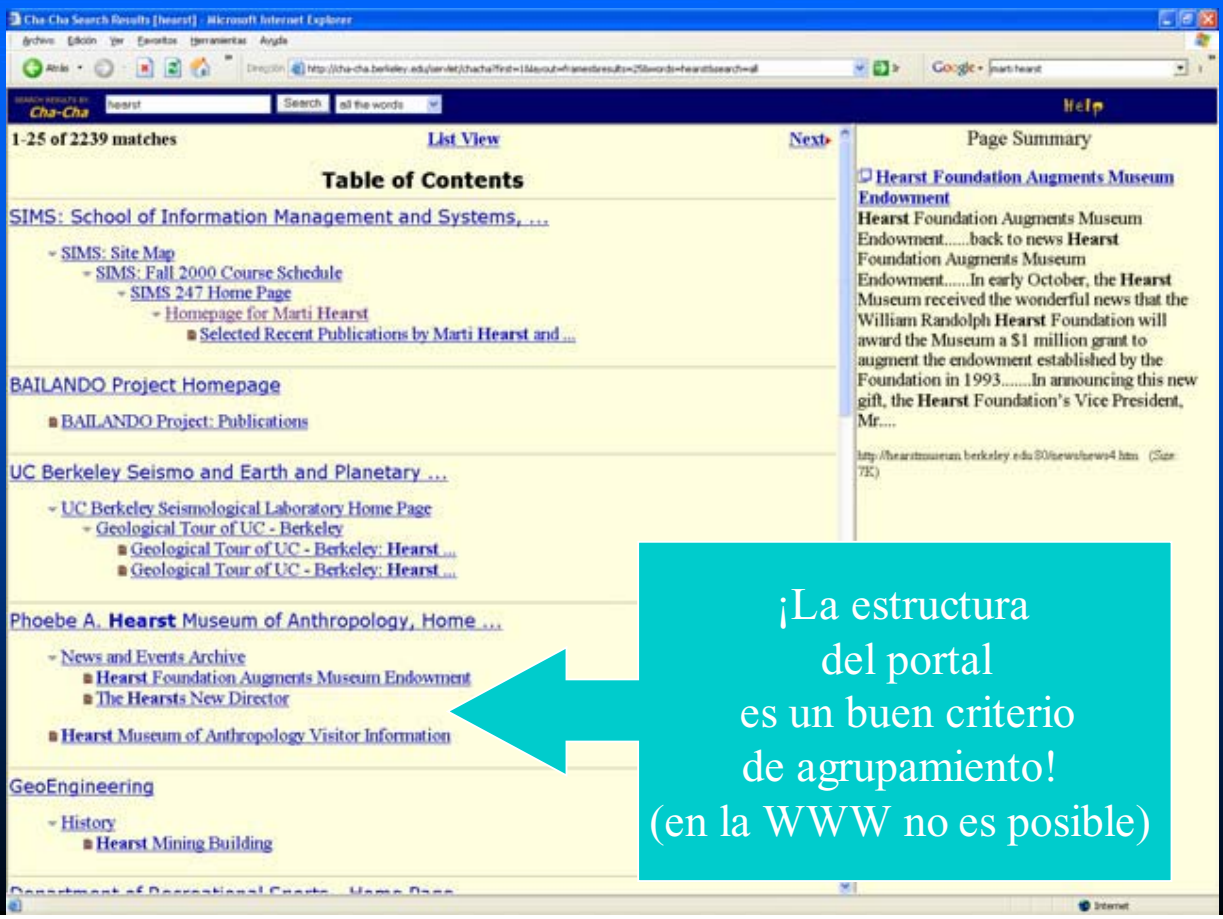
Visualización del sitio web: árbol hiperbólico



árbol hiperbólico + búsqueda



Páginas en contexto: *Cha-Cha*



Volvamos al problema del vocabulario

- Consulta: “reintegro ADSL”
- Documento: “ayuda nuevas tecnologias. Impreso número 1. Convocatoria Abril 2004”
- Ni Google ni el buscador UNED lo encuentran!

UNED Website Term Browser

Expansión y traducción de la consulta

	Tratados	de	Prohibición	de	Pruebas	Nucleares
Expansión	acuerdo capitulación concertación convenio cuidar, pacto manejar procesar		embargo entredicho interdicción interdicto proscripción		cata, catadura degustación ensayo escandallo experimento gustación muestreo, tanteo	nuclear
Traducción	accord discourse handle manage pact process treat treatise treaty		ban interdiction prohibition proscription		demonstrate establish, exhibit experiment experimentation fall, fitting indicate, point present, proof prove, run sample, sampling shew,show, taste test, trial, try	nuclear

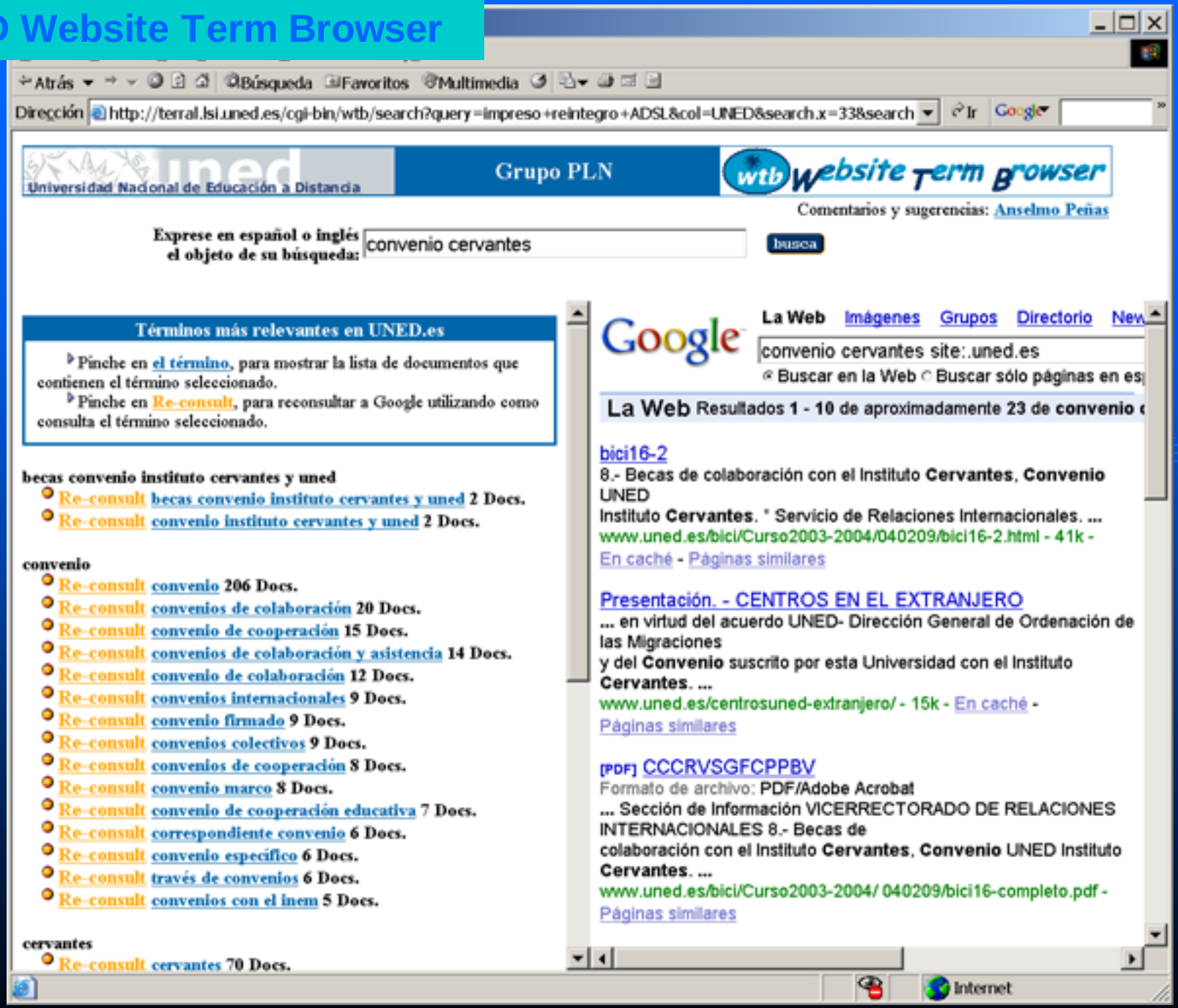
Nuclear fitting interdiction manage? Nuclear taste proscription process?

UNED Website Term Browser

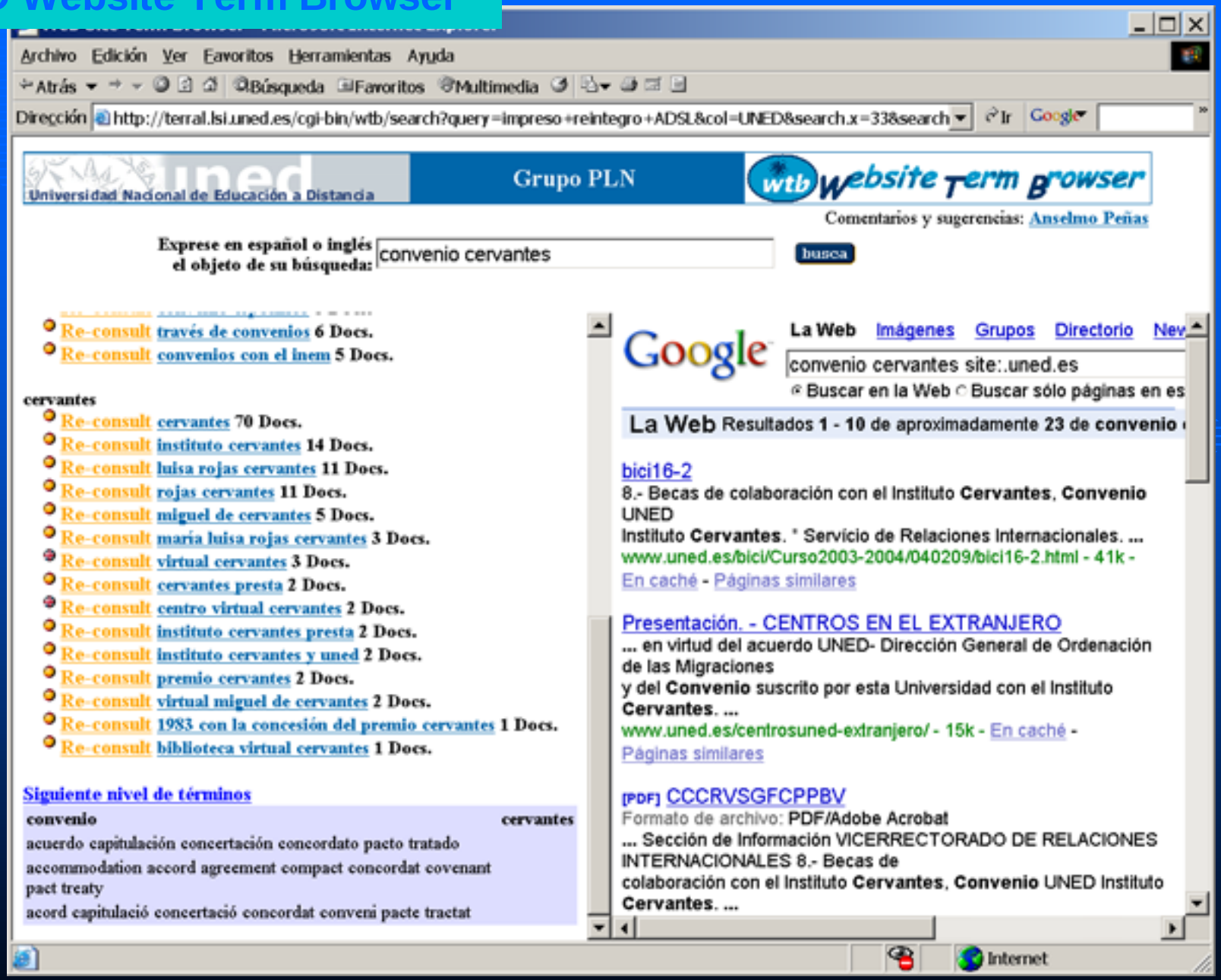
<< Anterior

Siguiente >>

UNED Website Term Browser



UNED Website Term Browser



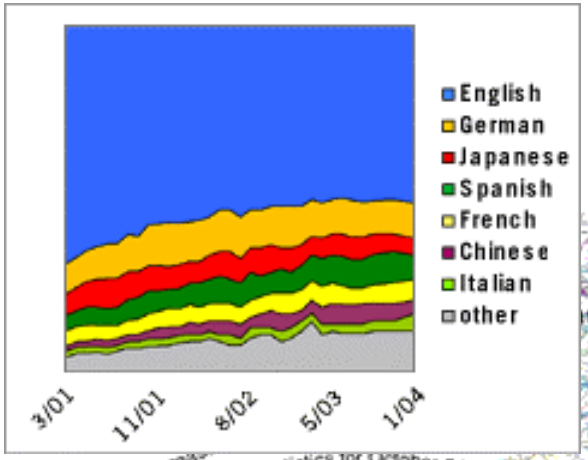
<< Anterior

Siguiente >>

¿Y a medio plazo?

- Eliminación de barreras idiomáticas y terminológicas
- Técnicas genéricas de localización y síntesis de *información*

Un detalle que se nos escapó sobre la WWW...



Исследований

att förstå

高等学校

يا ليلي يا عيني

Online usage statistics for October 24, 1999: 17.6 million total users

Average number of times a user visits the Internet per week: 6

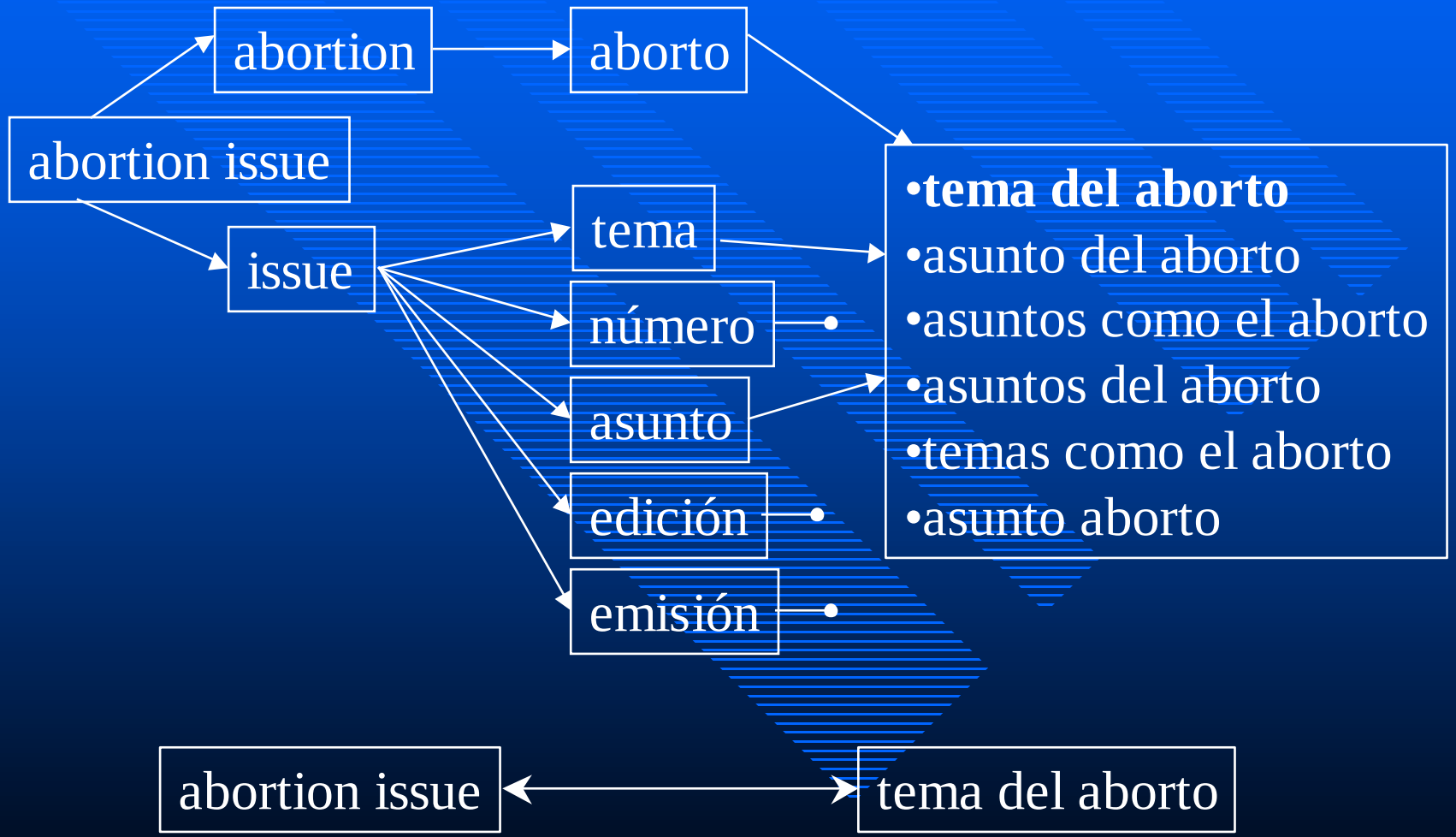
Average time a user spends on the Internet per week: 29:19 minutes

Number of U.S. households joining the Internet each hour: 760

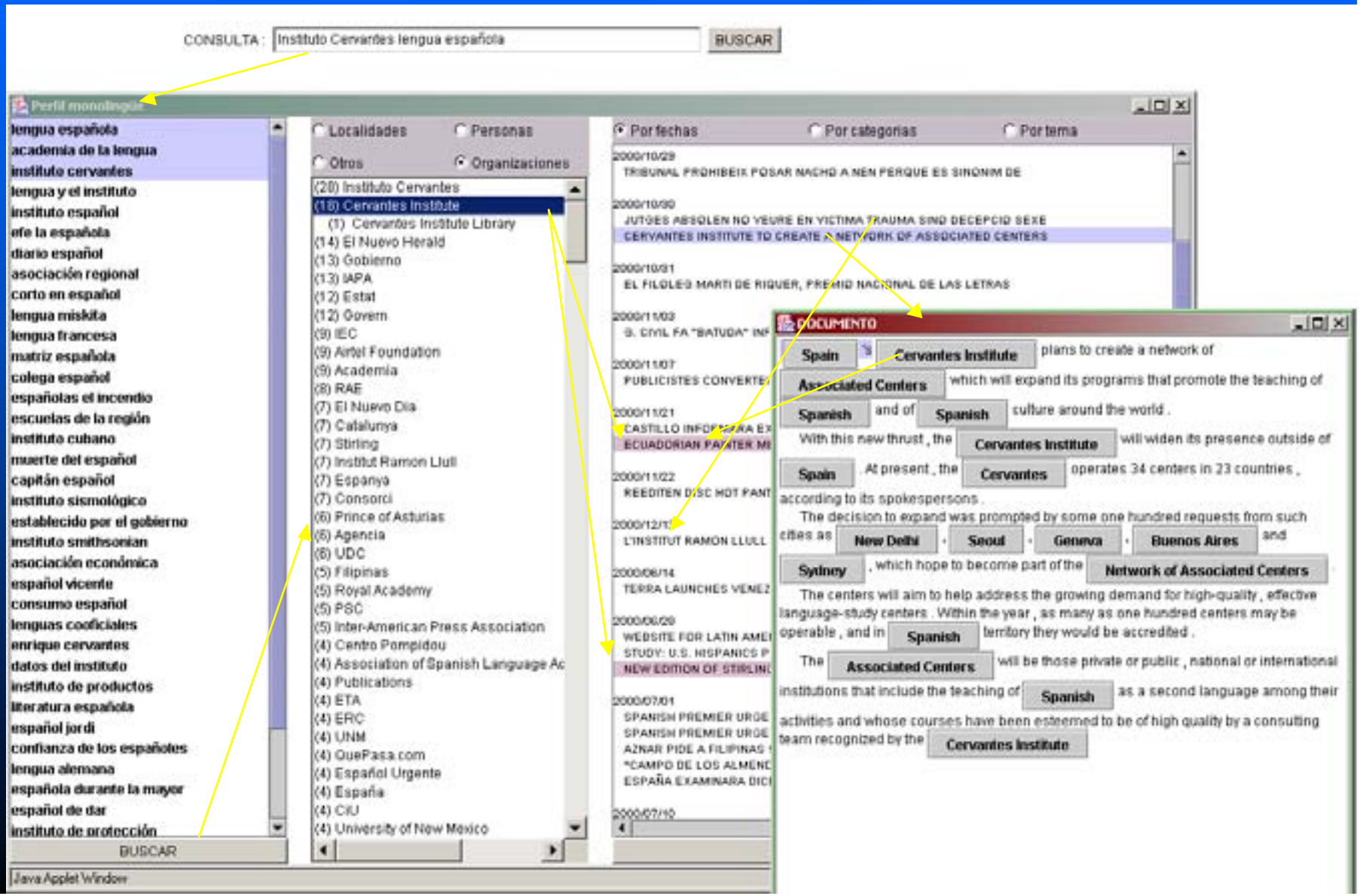
ARPANET
The world's first computer network, created by linking four computers across California and Utah. This original, four-node ARPANET evolved into our modern-day, billion-site Internet.

1969

Algoritmo de alineación



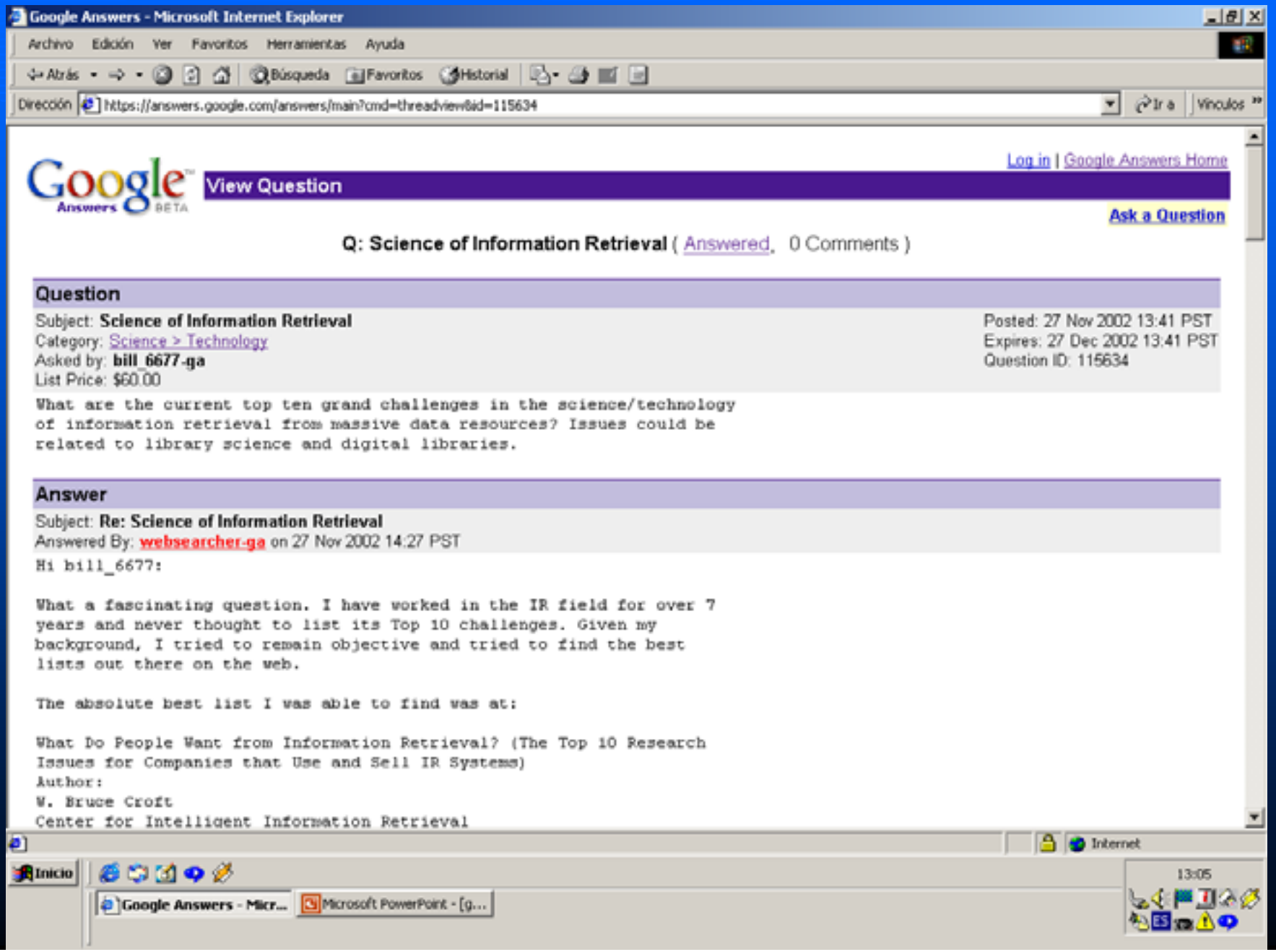
HERMES (UNED/UPC/UB/UPV): búsqueda y exploración multilingüe por entidades nombradas

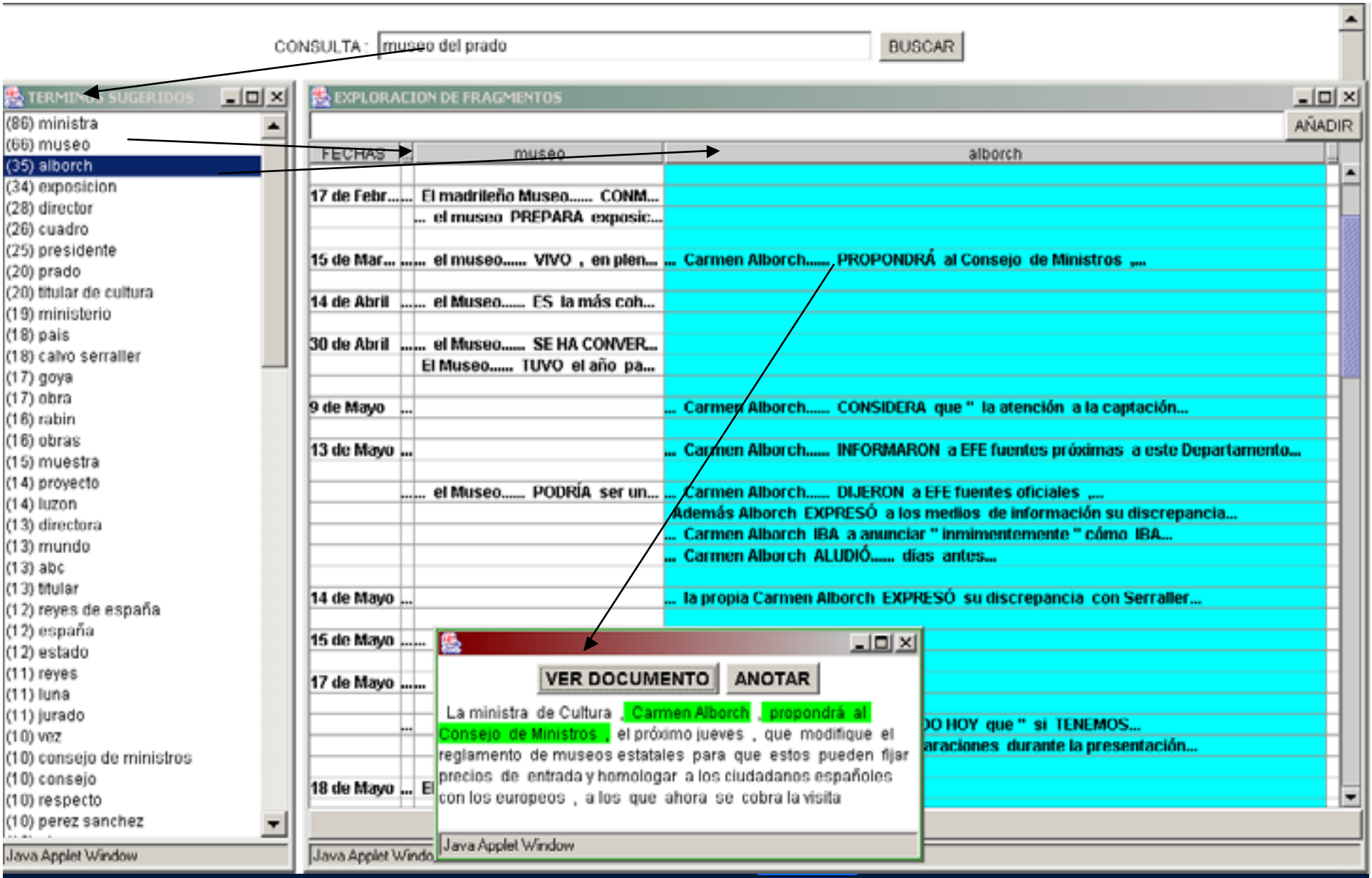


<< Anterior

Siguiente >>

Síntesis de información con la ayuda de Google.. y alguien más

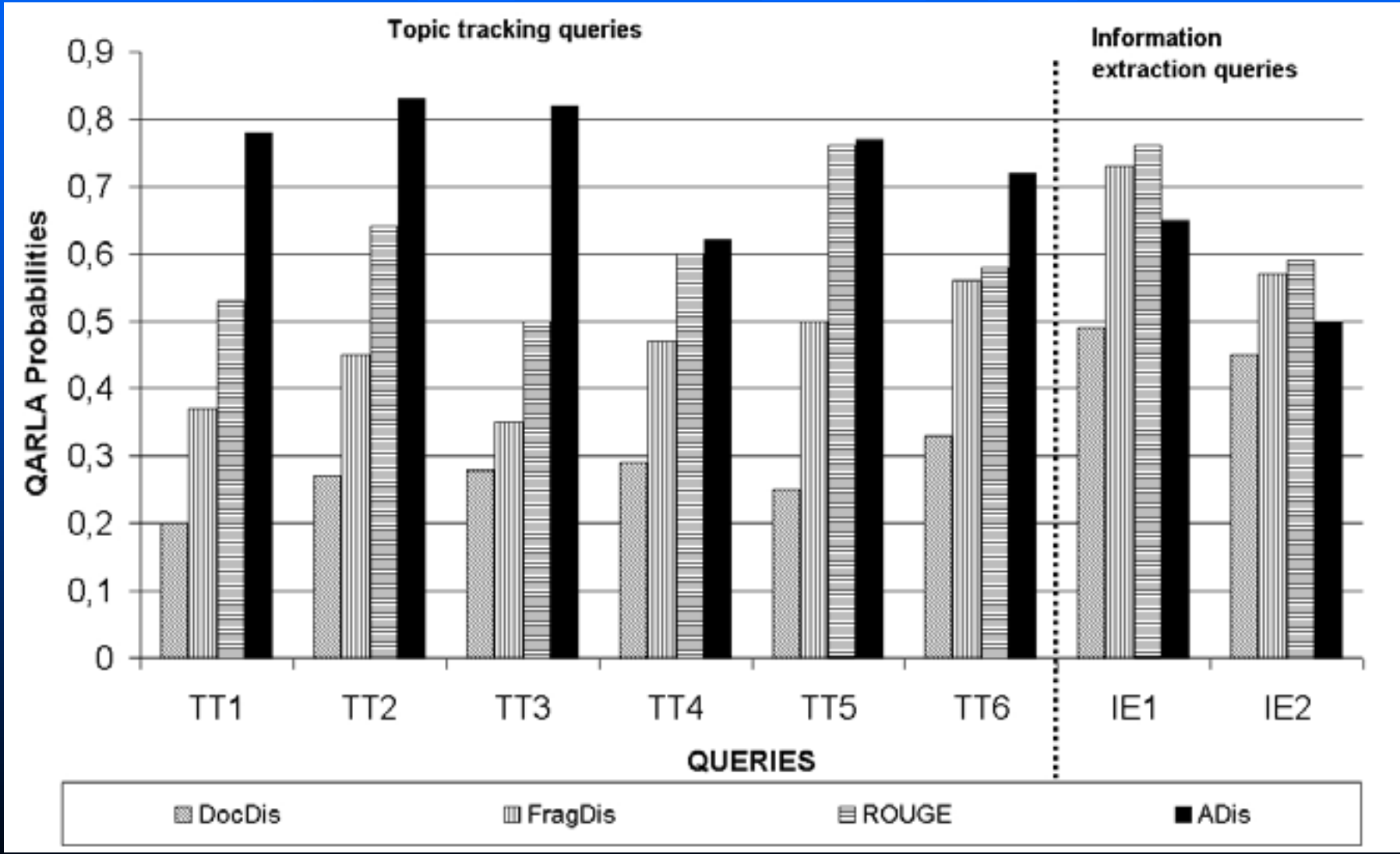




UNED PRiSMA

PRISMA: asistente para **síntesis** de información

¿Por qué agentes?



Implicaciones para posicionamiento

- El modelo de Google seguirá siendo válido para consultas de “dame un enlace”... Pero ¿son esas las más interesantes en temas culturales? **En otro tipo de consultas, la agrupación y los buscadores verticales harán menos crucial el posicionamiento.**
- Prepararse para (o crear!) **buscadores verticales** sobre temas culturales/educativos/científicos...**¿web semántica?**
- La pregunta más básica: ¿Dispone su sitio web de un **buscador apropiado para portales**?

Más información

- www.searchenginewatch.com
- www.sigir.org
- http://nlp.uned.es

<< Anterior