

Estudio de idiomas en el dominio .ES

Carlos Castillo
Universidad de Chile
ccastill@dcc.uchile.cl

Enero 2003

Resumen

Se presentan resultados de un estudio realizado en Diciembre del año 2002 sobre aproximadamente 1 millón de páginas del dominio .ES (español). De las páginas de las cuales se logró estimar el idioma, un 6 % estaba en catalán, 62 % en castellano y 32 % en inglés.

1. Colección

Durante diciembre del año 2002, se realizó una recolección de páginas por el dominio .ES utilizando el crawler de WIRE [1]. Este crawler utiliza múltiples *threads* para bajar páginas y un post-procesamiento de las páginas que mantiene el formato lógico y el formato físico del HTML eliminando el código HTML que no es relevante desde el punto de vista de recuperación de información, esto permite ahorrar aproximadamente un 60 % de espacio de la colección.

El tamaño de la colección es el indicado en el Cuadro 1:

Número de sitios conocidos	19.179
Número de páginas conocidas	5.877.289
Número de páginas visitadas	1.465.530
Número de conexiones exitosas	1.454.305
Número de páginas con estado HTTP OK	1.001.712
Tamaño de páginas procesadas	7 Gigabytes

Cuadro 1: Tamaño de la colección

El crawler utiliza el algoritmo Pagerank [2] y la profundidad de las páginas para priorizar ciertas páginas frente a otras.

La profundidad de una página es 1 si está en la raíz de un sitio web y N es el mínimo de profundidad de las páginas que la apuntan es $N - 1$. Las páginas dinámicas son identificadas por la presencia de un carácter separador CGI: “?” en la URL.

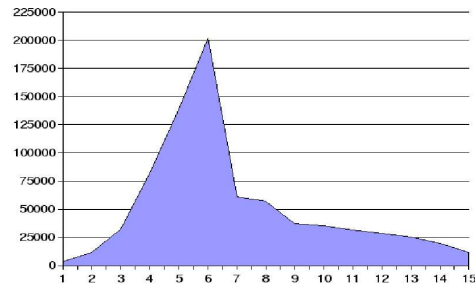


Figura 1: Páginas por profundidad en la colección.

La profundidad máxima explorada es de 5 enlaces para páginas dinámicas y 15 para páginas estáticas. El número de páginas crece rápidamente al aumentar la profundidad hasta el límite para páginas dinámicas, esto porque un mismo servidor puede generar infinitas páginas dinámicas, como se aprecia en el Cuadro 1.

2. Metodología

Se trabajó con los idiomas catalán, castellano e inglés, utilizando tres listas de palabras frecuentes indicadas en el Cuadro 5 en el anexo.

Para identificar el idioma de una página, se utilizó el siguiente algoritmo:

- Construir las listas de palabras frecuentes para idioma.
- Obtener el listado de palabras de la página, no incluyendo los *tags* HTML.
- Si la página tiene menos de 50 palabras, ignorar.
- Si la página no tiene más de 1 palabra de alguna lista, marcar como **indefinido**.
- Retornar como idioma de la página el de la lista de palabras frecuentes que más aparezcan en la página.

3. Resultados

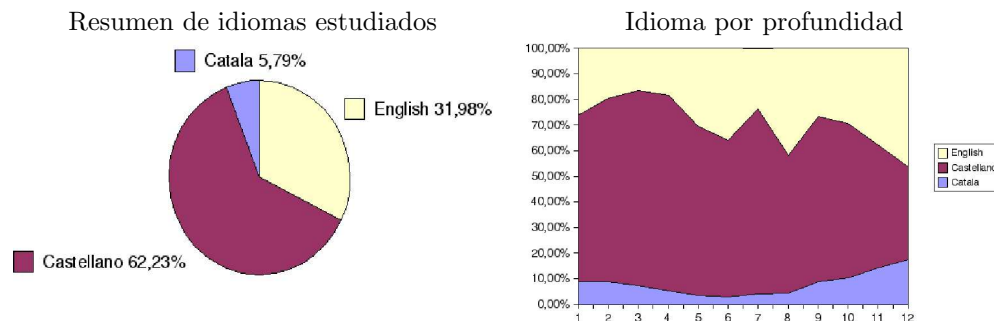


Figura 2: Resumen de la presencia de los idiomas estudiados.

La Figura 2 resume los resultados obtenidos, en cuanto a los idiomas encontrados. Nótese que se logró identificar el 73.27% de las páginas observadas; normalmente se trata de páginas breves o que contienen listas de palabras (ej.: listado de productos o de elementos que no incluyan conectivos) o que están en otros idiomas.

La tabla siguiente resume los resultados en cuanto a porcentajes sobre el total y sobre las páginas identificadas por el algoritmo.

Número de páginas con más de 50 palabras	778.373
Páginas en catalá	34.495 (4.24%)
Páginas en castellano	370.595 (45.59%)
Páginas en inglés	190.481 (23.43%)
Páginas no definidas	217.297 (26.73%)
Porcentaje sobre definidas de catalá	5.79%
Porcentaje sobre definidas de castellano	62.23%
Porcentaje sobre definidas de inglés	31.98%

Cuadro 2: Resultados obtenidos.

Es importante hacer notar que la Web es una colección muy heterogénea de documentos; por lo general la mayor cantidad de las páginas tiene escaso o nulo interés desde el punto de vista de la información que contiene; por ejemplo, en muchos casos se trata de archivos de foros de discusión o listas de correo en el cual cada página aporta sólo un párrafo de información.

Para esto, ordenamos las páginas por Pagerank y el resultado es el de la Figura 3.

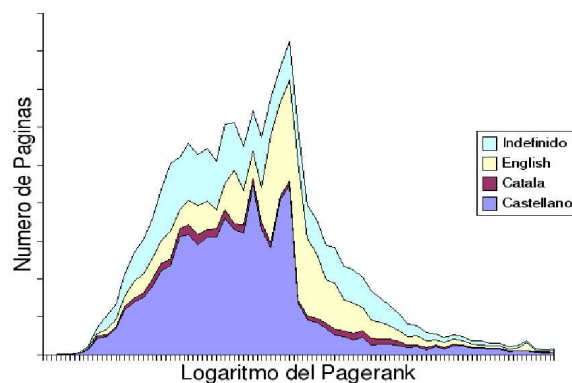


Figura 3: Idiomas estudiados por Pagerank. Las páginas de mayor calidad se encuentran al lado izquierdo.

4. Sitios con páginas en catalán

Los 15 sitios que se encontraron que tenían más páginas en catalán se encuentran en el Cuadro 3. Este cuadro continúa en el anexo en el Cuadro 5.

Sitio	Páginas en catalá	Porcentaje del sitio
cisne.sim.ucm.es	7995	28,12 %
ubucat.ubu.es	3171	10,53 %
www.edreams.es	2037	6,77 %
brumario.usal.es	1492	5,79 %
www.diba.es	1471	96,14 %
biblio.cesga.es	1198	5,19 %
www.ozu.es	1111	41,46 %
compras2.ozu.es	822	6,57 %
diba.es	820	85,06 %
diana.uca.es	659	2,57 %
www.vogue.es	526	43,76 %
www.ub.es	479	21,53 %
www.ua.es	458	10,36 %
elmundodeporte.elmundo.es	425	12,05 %
www.kodak.es	393	3,81 %

Cuadro 3: Los 15 sitios con más páginas en catalá y porcentaje de esas páginas sobre el total de páginas del sitio. Continúa en el anexo.

Entre estos sitios destacan algunos catálogos bibliográficos de universidades y servidores de archivo de listas de discusión por correo electrónico. Nótese que

si bien el crawler intenta realizar detección de duplicados, esta detección no permite identificar duplicados cercanos (ej.: una página que está en dos sitios, pero con pequeñas diferencias entre uno y otro).

5. Comentarios

El listado de sitios obtenidos podría usarse para construir un corpus de páginas web en catalán en el futuro, o un corpus bilingüe.

Se requieren nuevos estudios para intentar disminuir el número de páginas no definidas, aunque por las características de la Web es común que haya una abundante cantidad de contenido inclasificable desde el punto de vista documental.

Una revisión cuidadosa de una muestra de páginas indefinidas podría ayudar a identificar cuáles otros idiomas están presentes en el dominio .ES para incluirlos en futuros estudios.

Hay que destacar que por diversos motivos en muchos casos las empresas y organizaciones de un país utilizan los dominios de primer nivel genéricos .net, .com y .org para poner sus páginas, y no el dominio de su país; así mismo, algunas empresas internacionales mantienen *mirrors* o copias de sus sitios en cada país en el cuál tienen sucursales.

Un fichero con las estadísticas más relevantes está disponible, se puede solicitar enviando correo electrónico a `ccastill@dcc.uchile.cl`.

Referencias

- [1] BAEZA-YATES, R., AND CASTILLO, C. Balancing volume, quality and freshness in web crawling. In *International Conference on Hybrid Intelligent Systems* (2002), IOS Press.
- [2] PAGE, L., BRIN, S., MOTWAIN, R., AND WINOGRAD, T. The pagerank citation algorithm: bringing order to the web. In *7th World Wide Web Conference* (1998).

Anexos: Tablas de datos

La presente tabla de palabras fue facilitada por Steffan Bott de la Universitat Pompeu Fabras. Se han eliminado las interferencias, es decir, las palabras que son comunes en los tres idiomas.

Para el caso de las palabras que no pertenecen a los caracteres del `us-ascii`, se utilizó la palabra con acento como carácter y la palabra con acento en HTML, ejemplo: *és* y *Éeacute;s*.

Catalán		Castellano		Inglés	
els	amb	su	como	the	of
és	com	más	pero	and	that
més	però	le	sus	it	is
tot	aquest	¿	había	was	for
aquesta	són	cuando	todo	on	you
molt	havia	ya	muy	with	by
seu	fer	yo	qué	at	have
seva	quan	este	porque	are	not
li	també	sólo	también	this	but
ens	ho	años	fue	had	they
em	això	hasta	está	his	from
perquè	què	todos	él	she	which
altres	dir	desde	puede	we	an
encara	estat	ahora	tiempo	there	her
tots	fet	siempre	¡	were	one
bé	així	eso	tiene	do	been
anys	qual	uno	otra	all	their
després	tant	estaba	otro	would	will
només	vaig	aunque	ese	what	if
mateix	altra	mismo	día	when	said
qui	uns	así	bien	who	more
aquests	sempre	hace	vez	about	them
home	ell	algo	donde	some	could
dia	ben	esa	parte	him	into
altre	tota	hombre	mundo	its	then

Cuadro 4: Tabla de palabras comunes, sin interferencias

Sitio	Páginas en catalá	Porcentaje del sitio
www.comb.es	392	47,75 %
portall.lacaixa.es	321	34,33 %
www.xtec.es	254	64,47 %
www.gencat.es	244	45,35 %
buscador.terra.es	223	24,06 %
elvino.paginasamarillas.es	212	97,25 %
leslu.upc.es	212	6,97 %
www.upc.es	211	50,72 %
tienda.tiscali.es	208	5,51 %
listserv.rediris.es	205	1,67 %
empresas.lacaixa.es	205	25,37 %
fama.us.es	204	3,19 %
jabega.uma.es	197	3,74 %
www.uv.es	188	15,12 %
www.dooyoo.es	169	0,44 %
buscador.ozu.es	168	29,58 %
www.fib.upc.es	167	81,46 %
empreses.lacaixa.es	166	43,92 %
news.lycos.es	164	7,24 %
www.etsetb.upc.es	155	65,68 %
www.uib.es	149	39,52 %
categorias.ozu.es	148	67,89 %
www.pre.gva.es	136	88,31 %
cibernauta.grupocorreos.es	133	17,10 %
pie.xtec.es	126	67,02 %
www.teowin.es	126	36,95 %
cibernauta.nortecastilla.es	119	21,88 %
www.caixaterrassa.es	117	100,00 %
eps.udg.es	116	68,64 %
escher.upc.es	107	62,94 %
adv.millorsoft.es	107	93,04 %
www.lavanguardia.es	106	0,63 %
www.bib.ub.es	106	26,43 %
www.ictnet.es	101	15,19 %

Cuadro 5: Continuación, sitios con más de 100 páginas en catalá (excluyendo los 15 primeros).