Regional and linguistic patterns in Google™ positioning

Isidro F. Aguillo, Ignacio García, Natalia Arroyo
CINDOC-CSIC

PARIS
4S et EASST Meeting
25 - 28 août 2004 / *August, 25-28, 2004*
École des Mines de Paris
60, boulevard Saint Michel, Paris 6ème

InternetLab

Topics and aims

Webspace is growing exponentially covering all sectors, regions and cultures

- "Digital divide" expanded: Growth is heterogeneous
- Language "slices" can be used for sampling social & cultural differences in web presence

Search engines are providing tools for automatic extraction of language filtered information

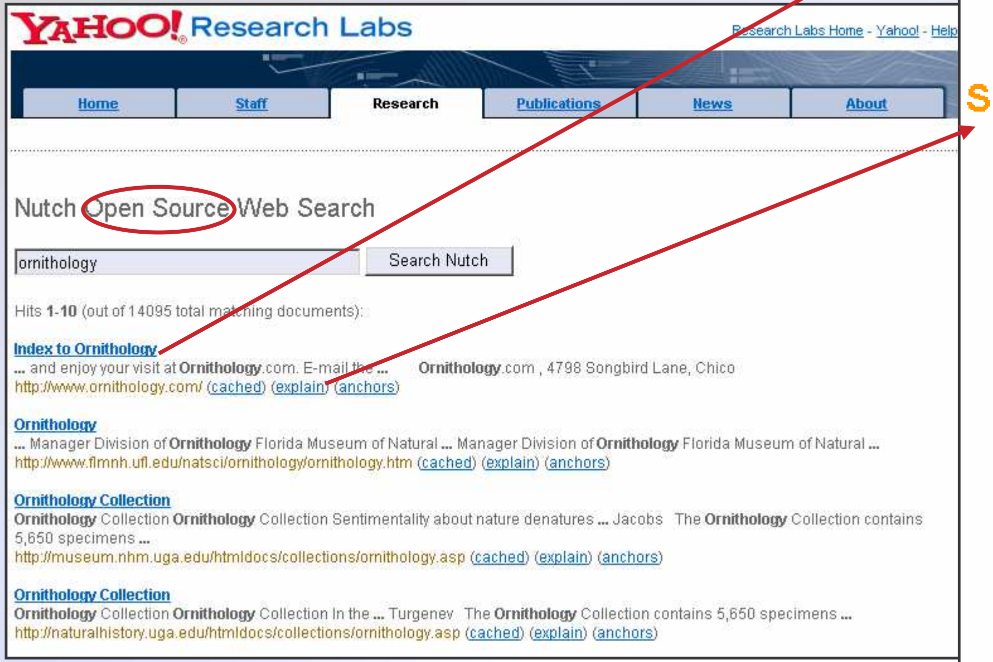
- Google Pagerank is useful as provides samples with the best positioned results
- The aim of this contribution is to show the differences in ranking performance of sites according to their language and type of institution involved

Google ranking

The order in which the results are provided by Google depends both on query matching and PageRank algorithms

Traditionally Google has the largest database and they are providing the details of the results in context, not the first sentences in the page

Web positioning strategies based in both adding keywords in key tags and attracting valuable links are being successful in Google



Nutch Open Source Web Search

ornithology

Search Nutch

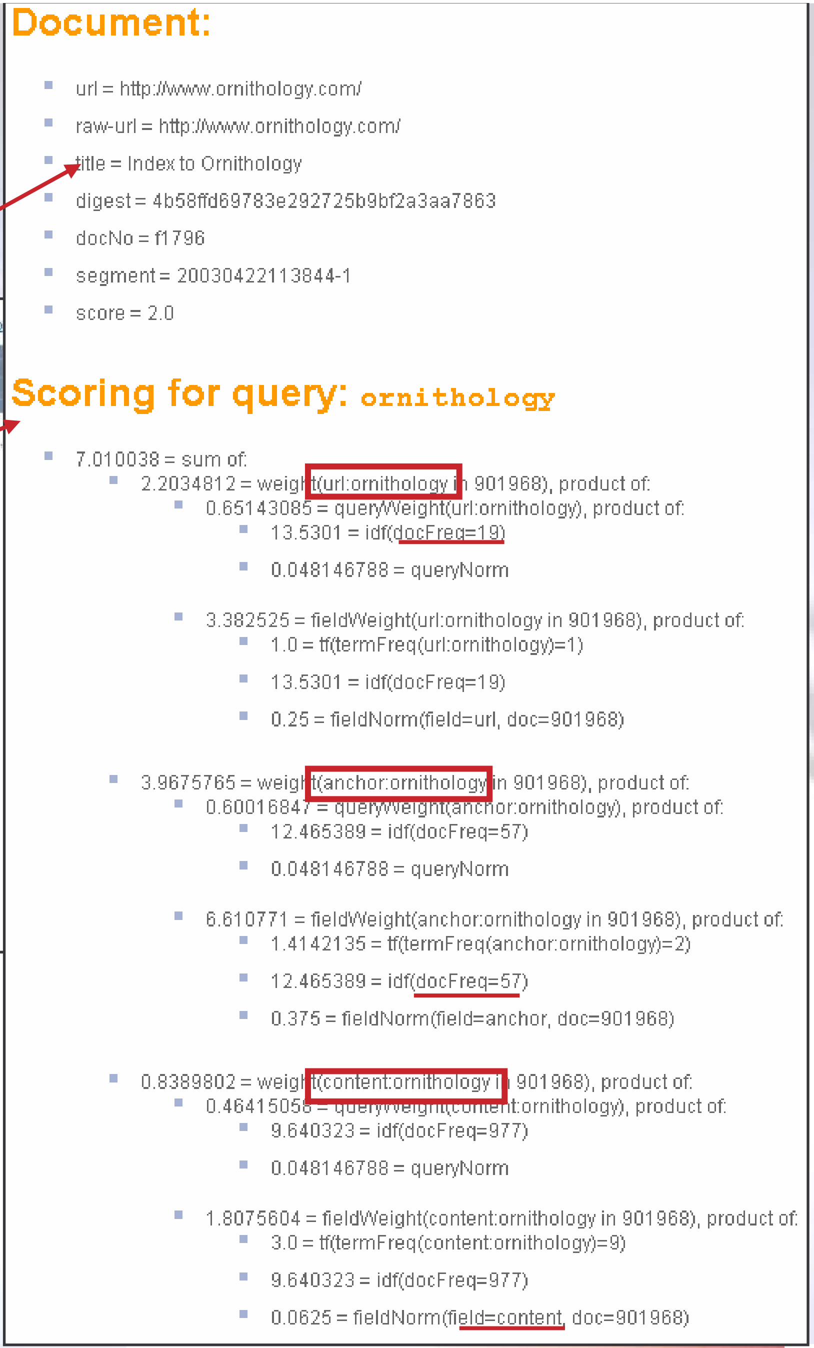
Hits **1-10** (out of 14095 total matching documents):

Index to Ornithology
... and enjoy your visit at **Ornithology**.com. E-mail the ... **Ornithology**.com , 4798 Songbird Lane, Chico
http://www.ornithology.com/ (cached) (explain) (anchors)

Ornithology
... Manager Division of **Ornithology** Florida Museum of Natural ... Manager Division of **Ornithology** Florida Museum of Natural ...
http://www.flmnh.ufl.edu/natsci/ornithology/ornithology.htm (cached) (explain) (anchors)

Ornithology Collection
Ornithology Collection **Ornithology** Collection Sentimentality about nature denatures ... Jacobs The **Ornithology** Collection contains 5,650 specimens ...
http://museum.nhm.uga.edu/htmldocs/collections/ornithology.asp (cached) (explain) (anchors)

Ornithology Collection
Ornithology Collection **Ornithology** Collection In the ... Turgenev The **Ornithology** Collection contains 5,650 specimens ...
http://naturalhistory.uga.edu/htmldocs/collections/ornithology.asp (cached) (explain) (anchors)



Document:

- url = http://www.ornithology.com/
- raw-url = http://www.ornithology.com/
- title = Index to Ornithology
- digest = 4b58ffd69783e292725b9bf2a3aa7863
- docNo = f1796
- segment = 20030422113844-1
- score = 2.0

Scoring for query: ornithology

- 7.010038 = sum of:
  - 2.2034812 = weight(url:ornithology in 901968), product of:
    - 0.65143085 = queryWeight(url:ornithology), product of:
      - 13.5301 = idf(docFreq=19)
      - 0.048146788 = queryNorm
    - 3.382525 = fieldWeight(url:ornithology in 901968), product of:
      - 1.0 = tf(termFreq(url:ornithology)=1)
      - 13.5301 = idf(docFreq=19)
      - 0.25 = fieldNorm(field=url, doc=901968)
  - 3.9675765 = weight(anchor:ornithology in 901968), product of:
    - 0.60016847 = queryWeight(anchor:ornithology), product of:
      - 12.465389 = idf(docFreq=57)
      - 0.048146788 = queryNorm
    - 6.610771 = fieldWeight(anchor:ornithology in 901968), product of:
      - 1.4142135 = tf(termFreq(anchor:ornithology)=2)
      - 12.465389 = idf(docFreq=57)
      - 0.375 = fieldNorm(field=anchor, doc=901968)
  - 0.8389802 = weight(content:ornithology in 901968), product of:
    - 0.46415058 = queryWeight(content:ornithology), product of:
      - 9.640323 = idf(docFreq=977)
      - 0.048146788 = queryNorm
    - 1.8075604 = fieldWeight(content:ornithology in 901968), product of:
      - 3.0 = tf(termFreq(content:ornithology)=9)
      - 9.640323 = idf(docFreq=977)
      - 0.0625 = fieldNorm(field=content, doc=901968)

Google algorithms are commercial secrets but a similar approach is used in “open” Nutch

PageRank

PageRank is a numeric value that represents how important a page is on the web

- PageRank algorithm is public (Brin & Paige, 1998) and a lot of mathematical research has been done based on it

The PR of a webpage is calculated taking into account all of its inbound links

$$PR(A) = (1-d) + d(PR(t1)/C(t1) + \ldots + PR(tn)/C(tn))$$

- In the equation 't1 - tn' are pages linking to page A, 'C' is the number of outbound links that a page has and 'd' is a damping factor, usually set to 0.85

Factors affecting PR includes number of links received, the PR of the webpages originating these links and the links provided by the page itself

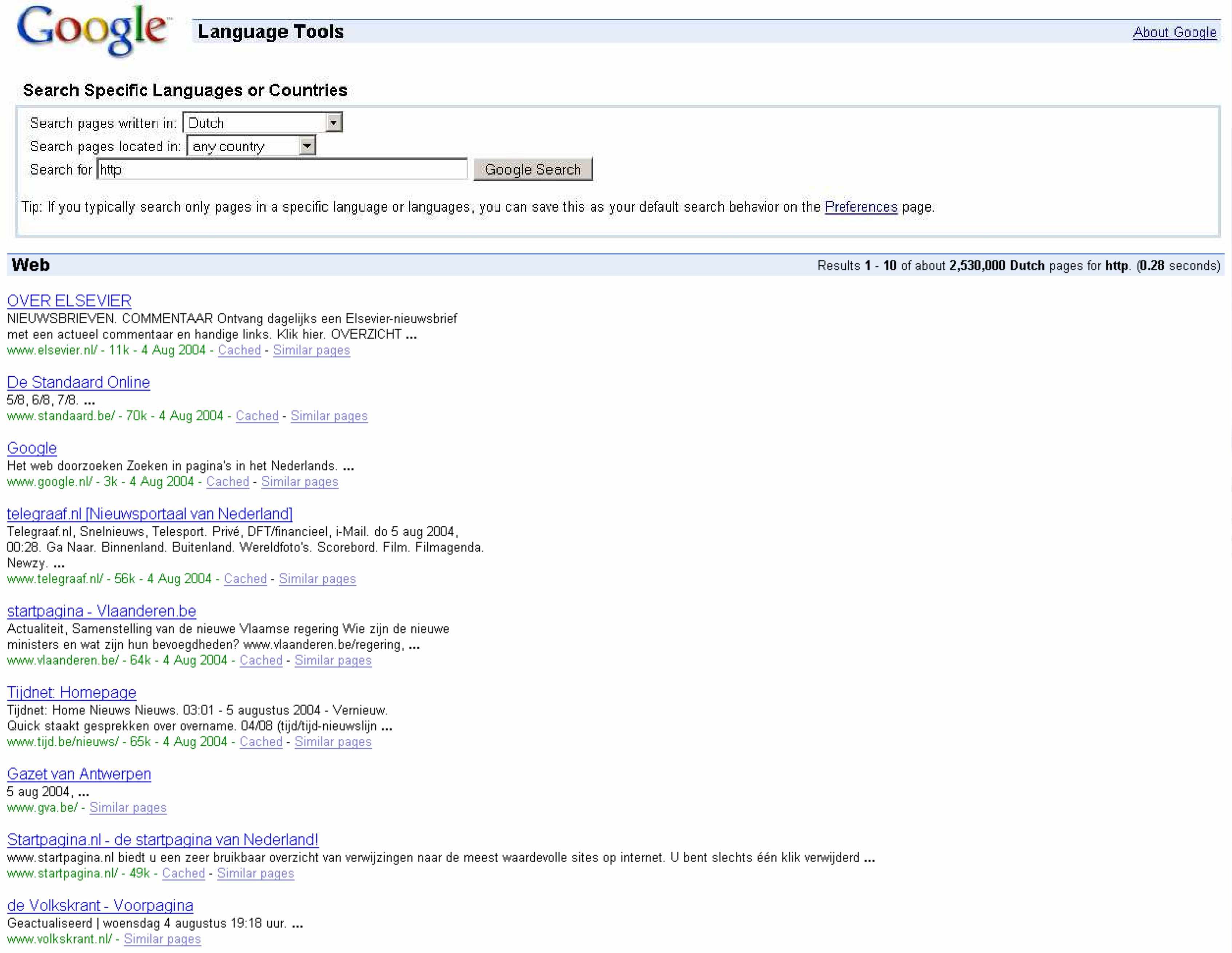
Methodology

In order to extract only PR ranking a “neutral” search term was used: http

Using Google language tools the first 100 results were obtained for the http request in

- French
- Dutch
- Portuguese
- Spanish and
- Spanish in the USA

Websites were classified according country, sector and language/es of the pages



Language Tools

Search Specific Languages or Countries

Search pages written in: Dutch

Search pages located in: any country

Search for http

Tip: If you typically search only pages in a specific language or languages, you can save this as your default search behavior on the Preferences page.

Web

Results **1** - **10** of about **2,530,000 Dutch** pages for **http**. (**0.28** seconds)

OVER ELSEVIER
NIEUWSBRIEVEN. COMMENTAAR Ontvang dagelijks een Elsevier-nieuwsbrief met een actueel commentaar en handige links. Klik hier. OVERZICHT **...**
www.elsevier.nl/ - 11k - 4 Aug 2004 - Cached - Similar pages

De Standaard Online
5/8, 6/8, 7/8. **...**
www.standaard.be/ - 70k - 4 Aug 2004 - Cached - Similar pages

Google
Het web doorzoeken Zoeken in pagina's in het Nederlands. **...**
www.google.nl/ - 3k - 4 Aug 2004 - Cached - Similar pages

telegraaf.nl [Nieuwsportaal van Nederland]
Telegraaf.nl, Snelnieuws, Telesport. Privé, DFT/financieel, i-Mail. do 5 aug 2004, 00:28. Ga Naar. Binnenland. Buitenland. Wereldfoto's. Scorebord. Film. Filmagenda. Newzy. **...**
www.telegraaf.nl/ - 56k - 4 Aug 2004 - Cached - Similar pages

startpagina - Vlaanderen.be
Actualiteit, Samenstelling van de nieuwe Vlaamse regering Wie zijn de nieuwe ministers en wat zijn hun bevoegdheden? www.vlaanderen.be/regering, **...**
www.vlaanderen.be/ - 64k - 4 Aug 2004 - Cached - Similar pages

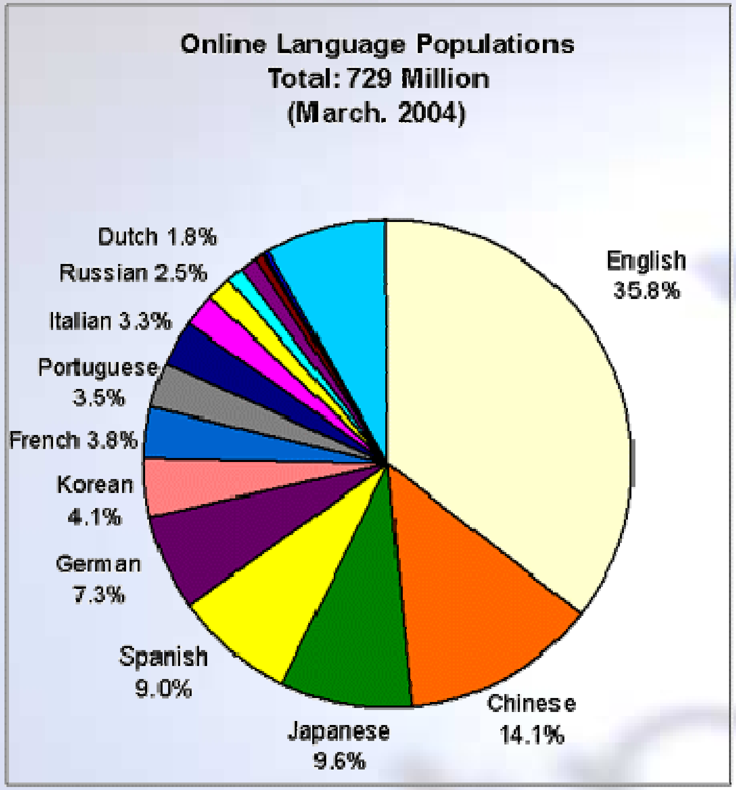
Tijdnet: Homepage
Tijdnet: Home Nieuws Nieuws. 03:01 - 5 augustus 2004 - Vernieuw. Quick staakt gesprekken over overname. 04/08 (tijd/tijd-nieuwslijn **...**
www.tijd.be/nieuws/ - 65k - 4 Aug 2004 - Cached - Similar pages

Gazet van Antwerpen
5 aug 2004, **...**
www.gva.be/ - Similar pages

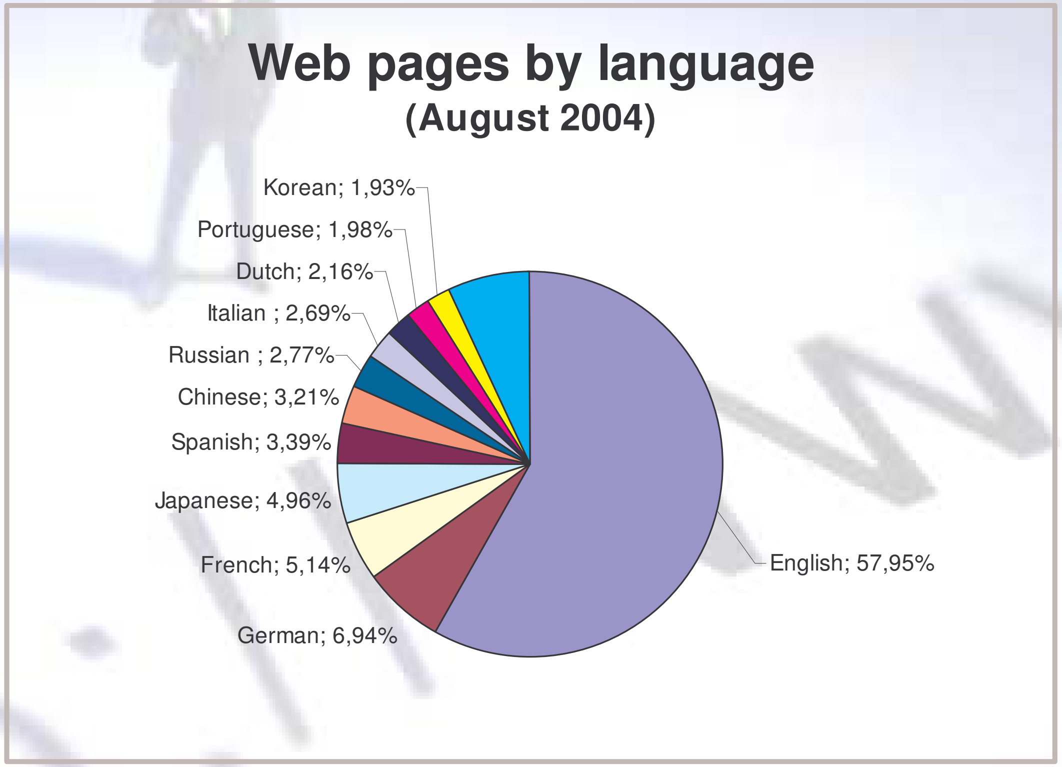
Startpagina.nl - de startpagina van Nederland!
www.startpagina.nl biedt u een zeer bruikbaar overzicht van verwijzingen naar de meest waardevolle sites op internet. U bent slechts één klik verwijderd **...**
www.startpagina.nl/ - 49k - Cached - Similar pages

de Volkskrant - Voorpagina
Geactualiseerd | woensdag 4 augustus 19:18 uur. **...**
www.volkskrant.nl/ - Similar pages

User and page languages



Global Reach (global-reach.biz/globstats)



Aguillo et al., estimations using Yahoo and Google

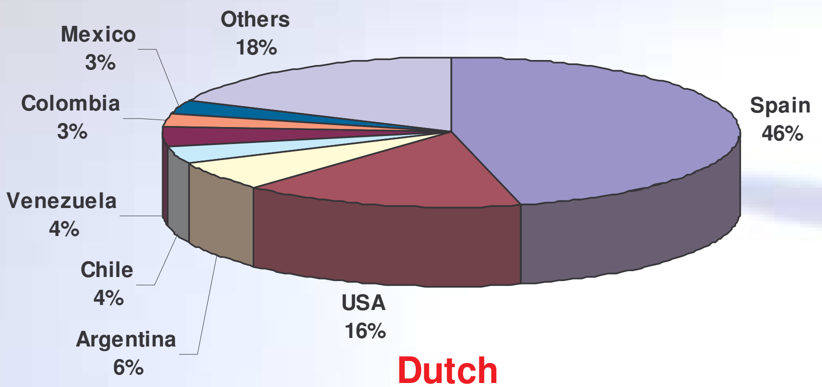
Classification

First 100 results classified according to type of institution or contents

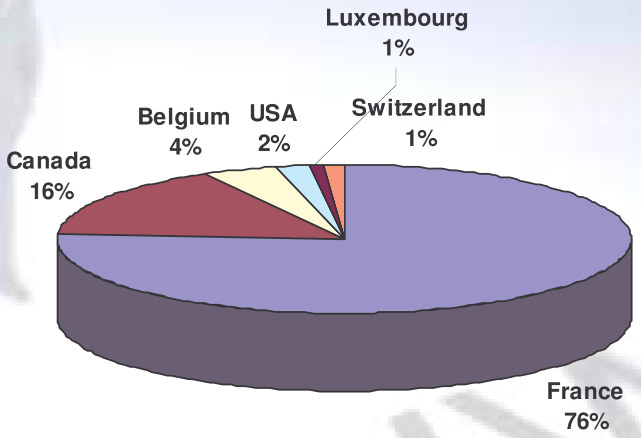
Private companies or firms	COMP
Cultural topics and issues	CULT
International groups	INT
Newspapers, Radio & TV	NEWS
Governmental	GOV
NGOs	NGO
Portals, directories or super sites	PORT
Universities	UNIV

Results: Countries

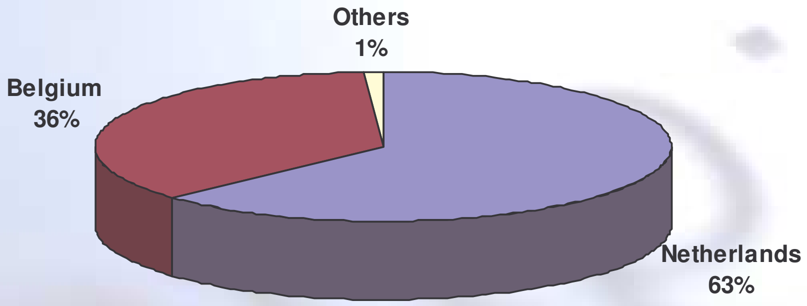
Spanish



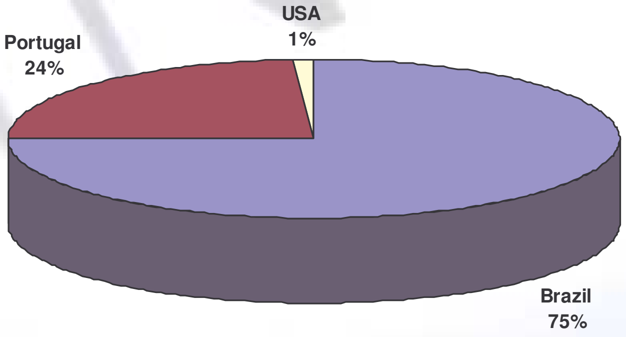
French



Dutch



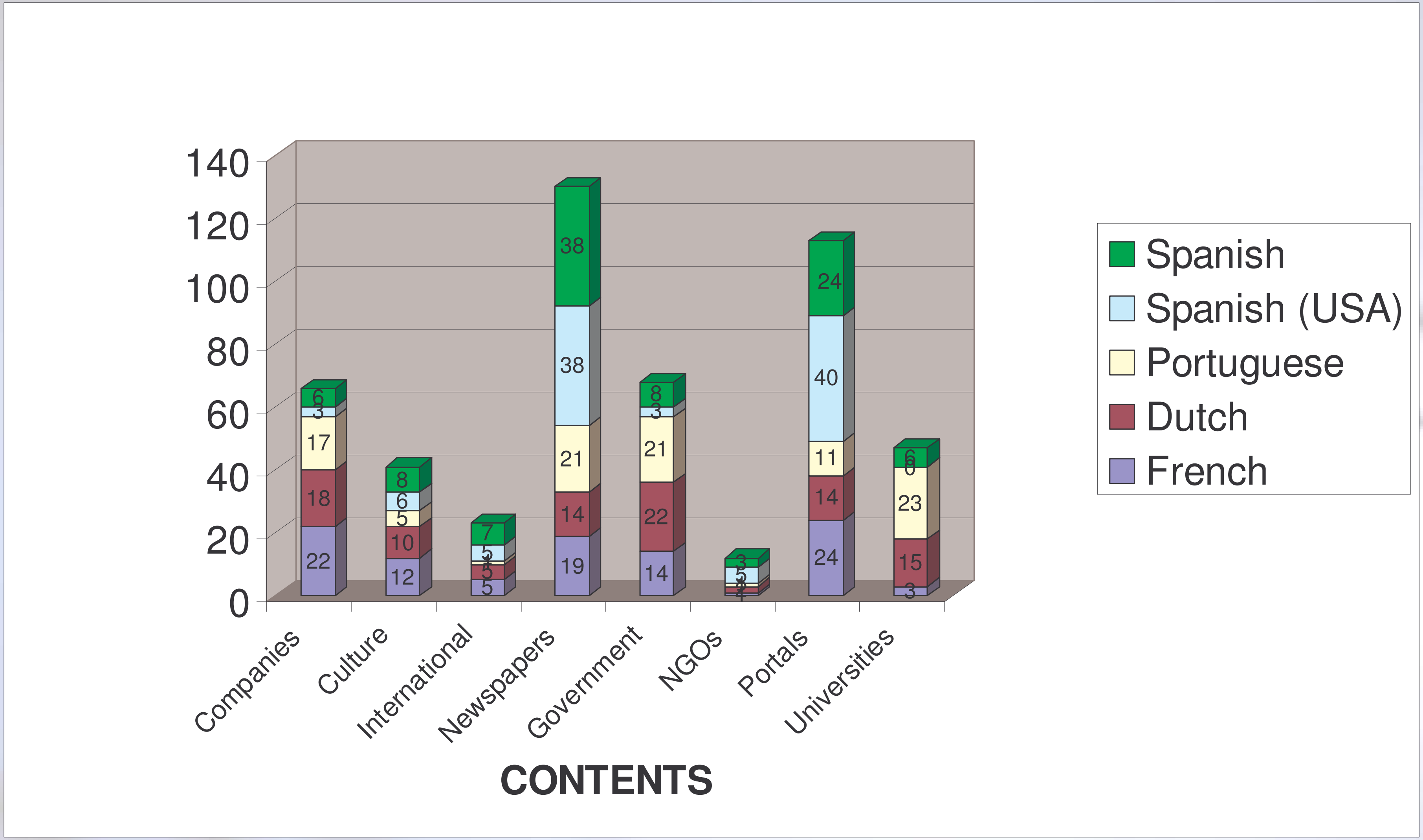
Portuguese



Results: Languages

	Portuguese	Spanish	Spanish (USA)	French	Dutch
Main Language	80	70	77	54	35
Main Language+English	9	17	10	23	39
English	1	0	0	2	8
Several languages	10	13	13	21	10

Results: Institutional type



Conclusions (I)

Google PageRank can be used for retrieving rankings of websites with high visibility

Language tools in Google allow the filtering of results according to this and other combined criteria

Automatic methods are valid till 1.000 first results

Conclusions (II)

Languages selected are spoken in several countries

- For the best ranked websites usually one country is specially well represented: Brazil and France (3/4 in their respective languages) or Netherlands (2/3)
- Spain is leader in Spanish but with important contributions by other countries

Increasing number of websites are multilingual

- Portuguese and Spanish sites use mainly their own language with no alternative texts in other languages
- French and Dutch sites are clearly more multilingual with English as standard offer

Contents of the most valued websites are “popular”

- News are the key information in all the languages specially if offered by well known newspapers radio or TV companies
- Main companies are well represented as governmental offices meaning it is important web presence for institutional visibility
- Brazilian universities can be seen as an important part of the Portuguese core
- Super-sites like big portals or directories with mostly local contents appear in the first positions in every language

More information

WISER

Web Indicators for Science, Technology & Innovation Research

www.webindicators.org

Questions?