FCCN
Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

LASIGE
Large-Scale Informatics Systems Laboratory

Information Search in Web Archives

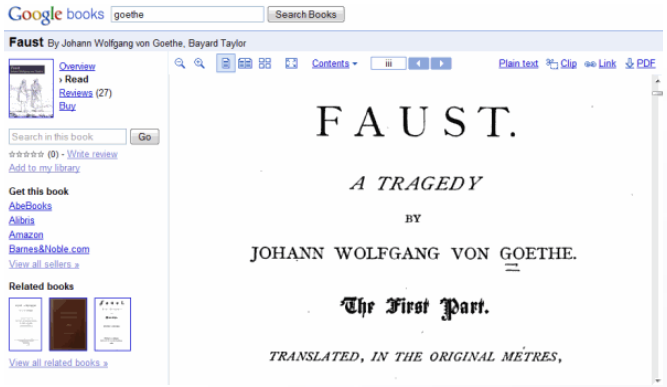
Miguel Costa
Portuguese Foundation for National Scientific Computing
LaSIGE @ Faculty of Sciences, University of Lisbon

1st International Workshop on Archiving Community Memories
September 6, 2013, Lisbon, Portugal

GOVERNO DE PORTUGAL
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

Our Memory is in Digital Form

E-books



Google books goethe Search Books

Faust By Johann Wolfgang von Goethe, Bayard Taylor

Overview
› Read
Reviews (27)
Buy

Search in this book Go
☆☆☆☆☆ (0) - Write review
Add to my library

Get this book
AbeBooks
Alibris
Amazon
Barnes&Noble.com
View all sellers »

Related books
View all related books »

Contents ▾ iii Plain text Clip Link PDF

FAUST.

A TRAGEDY

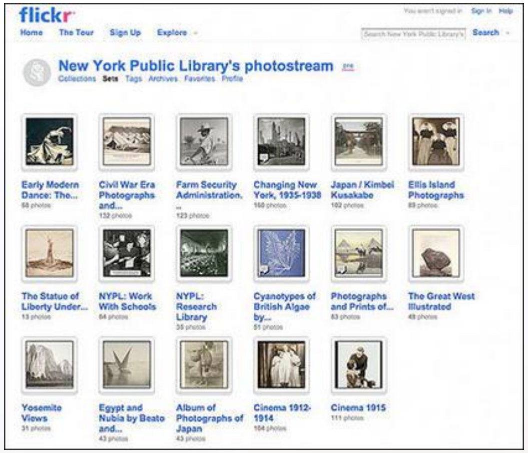
BY

JOHANN WOLFGANG VON GOETHE.

The First Part.

TRANSLATED, IN THE ORIGINAL METRES,

Web photo galleries

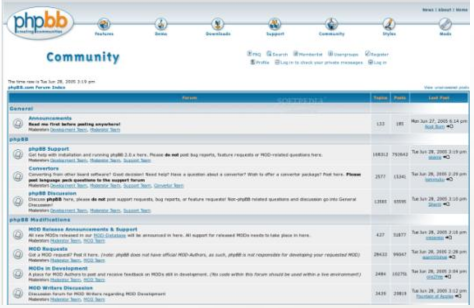


flickr
You aren't signed in Sign In Help
Home The Tour Sign Up Explore
Search New York Public Library's Search

New York Public Library's photostream
Collections Sets Tags Archives Favorites Profile

Early Modern Dance: The... 68 photos
Civil War Era Photographs and... 132 photos
Farm Security Administration. ... 123 photos
Changing New York, 1935-1938 160 photos
Japan / Kimbei Kusakabe 102 photos
Ellis Island Photographs 89 photos
The Statue of Liberty Under... 13 photos
NYPL: Work With Schools 84 photos
NYPL: Research Library 35 photos
Cyanotypes of British Algae by... 51 photos
Photographs and Prints of... 83 photos
The Great West Illustrated 48 photos
Yosemite Views 31 photos
Egypt and Nubia by Beato and... 43 photos
Album of Photographs of Japan 43 photos
Cinema 1912-1914 104 photos
Cinema 1915 111 photos

Forums



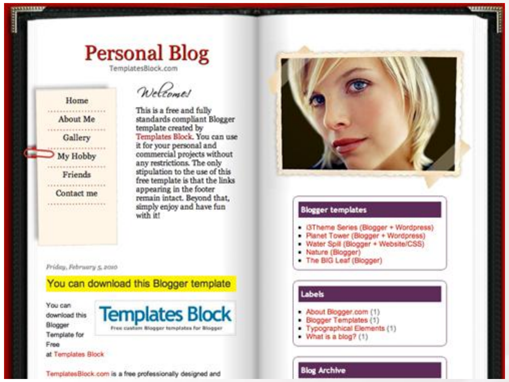
phpbb creating communities
News | About | Home
Features Demo Downloads Support Community Styles Mods
FAQ Search Memberlist Usergroups Register
Profile Log in to check your private messages Log in

Community

phpBB.com Forum Index

General
Announcements
Read me first before posting anywhere!
phpBB
phpBB Support
Convertors
phpBB Discussion
phpBB Modifications
MOD Release Announcements & Support
MOD Requests
MODs in Development
MOD Writers Discussion

Blogs



Personal Blog
TemplatesBlock.com

Home
About Me
Gallery
My Hobby
Friends
Contact me

Welcome!

This is a free and fully standards compliant Blogger template created by Templates Block. You can use it for your personal and commercial projects without any restrictions. The only stipulation to the use of this free template is that the links appearing in the footer remain intact. Beyond that, simply enjoy and have fun with it!

Friday, February 5, 2010

You can download this Blogger template

You can download this Blogger Template for Free at Templates Block

Templates Block
Free custom Blogger templates for Blogger

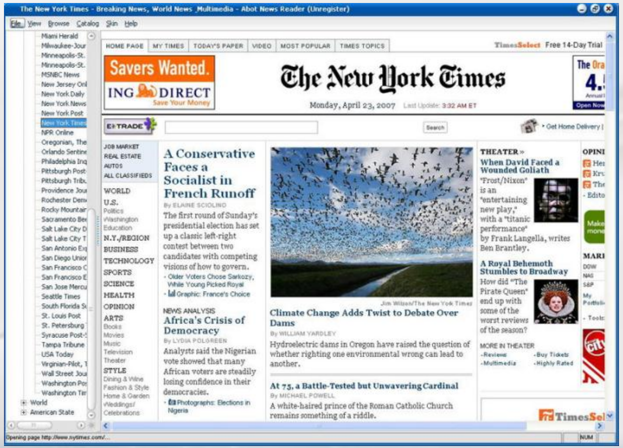
TemplatesBlock.com is a free professionally designed and

Blogger templates
- i3Theme Series (Blogger + Wordpress)
- Planet Tower (Blogger + Wordpress)
- Water Spill (Blogger + Website/CSS)
- Nature (Blogger)
- The BIG Leaf (Blogger)

Labels
- About Blogger.com (1)
- Blogger Templates (1)
- Typographical Elements (1)
- What is a blog? (1)

Blog Archive

Online newspapers



The New York Times - Breaking News, World News _Multimedia - Abot News Reader (Unregister)
File View Browse Catalog Skin Help
HOME PAGE MY TIMES TODAY'S PAPER VIDEO MOST POPULAR TIMES TOPICS
TimesSelect Free 14-Day Trial
Savers Wanted. ING DIRECT Save Your Money
E*TRADE Search Get Home Delivery
Opening page http://www.nytimes.com/...
NUM

Miami Herald
Milwaukee-Jour
Minneapolis-St.
Minneapolis-St.
MSNBC News
New Jersey Onl
New York Daily
New York News
New York Post
New York Times
NPR Online
Oregonian, The
Orlando Sentine
Philadelphia Inq
Pittsburgh Post
Pittsburgh Trib
Providence Jour
Rochester Dem
Rocky Mountain
Sacramento Be
Salt Lake City D
Salt Lake City T
San Antonio Ex
San Diego Unior
San Francisco C
San Francisco E
San Jose Mercu
Seattle Times
South Florida S
St. Louis Post
St. Petersburg
Syracuse Post-
Tampa Tribune
USA Today
Virginian-Pilot,
Wall Street Jou
Washington Po
Washington Tin
World
American State

The New York Times
Monday, April 23, 2007 Last Update: 3:32 AM ET

JOB MARKET
REAL ESTATE
AUTOS
ALL CLASSIFIEDS
WORLD
U.S.
Politics
Washington
Education
N.Y./REGION
BUSINESS
TECHNOLOGY
SPORTS
SCIENCE
HEALTH
OPINION
ARTS
Books
Movies
Music
Television
Theater
STYLE
Dining & Wine
Fashion & Style
Home & Garden
Weddings/Celebrations

A Conservative Faces a Socialist in French Runoff
By ELAINE SCIOLINO
The first round of Sunday's presidential election has set up a classic left-right contest between two candidates with competing visions of how to govern.
- Older Voters Chose Sarkozy, While Young Picked Royal
- Graphic: France's Choice

NEWS ANALYSIS
Africa's Crisis of Democracy
By LYDIA POLGREEN
Analysts said the Nigerian vote showed that many African voters are steadily losing confidence in their democracies.
- Photographs: Elections in Nigeria

Jim Wilson/The New York Times
Climate Change Adds Twist to Debate Over Dams
By WILLIAM YARDLEY
Hydroelectric dams in Oregon have raised the question of whether righting one environmental wrong can lead to another.

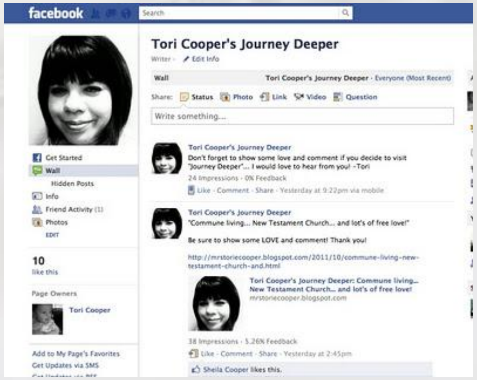
At 75, a Battle-Tested but Unwavering Cardinal
By MICHAEL POWELL
A white-haired prince of the Roman Catholic Church remains something of a riddle.

THEATER »
When David Faced a Wounded Goliath
"Frost/Nixon" is an "entertaining new play," with a "titanic performance" by Frank Langella, writes Ben Brantley.

A Royal Behemoth Stumbles to Broadway
How did "The Pirate Queen" end up with some of the worst reviews of the season?

MORE IN THEATER
-Reviews -Buy Tickets
-Multimedia -Highly Rated

Social networks



Tori Cooper's Journey Deeper
Writer · Edit Info

Wall Tori Cooper's Journey Deeper · Everyone (Most Recent)

Get Started
Wall
Hidden Posts
Info
Friend Activity (1)
Photos
EDIT

10
like this

Page Owners
Tori Cooper

Add to My Page's Favorites
Get Updates via SMS

Tori Cooper's Journey Deeper
Don't forget to show some love and comment if you decide to visit "Journey Deeper"... I would love to hear from you! -Tori
24 Impressions · 0% Feedback
Like · Comment · Share · Yesterday at 9:22pm via mobile

Tori Cooper's Journey Deeper
"Commune living... New Testament Church... and lot's of free love!"

Be sure to show some LOVE and comment! Thank you!

http://mrstoriecooper.blogspot.com/2011/10/commune-living-new-testament-church-and.html

Tori Cooper's Journey Deeper: Commune living... New Testament Church... and lot's of free love!
mrstoriecooper.blogspot.com

38 Impressions · 5.26% Feedback
Like · Comment · Share · Yesterday at 2:45pm
Sheila Cooper likes this.

The Web is Ephemeral

- 50 days - 50% of documents are changed

  (Cho and Garcia-Molina. 2000)

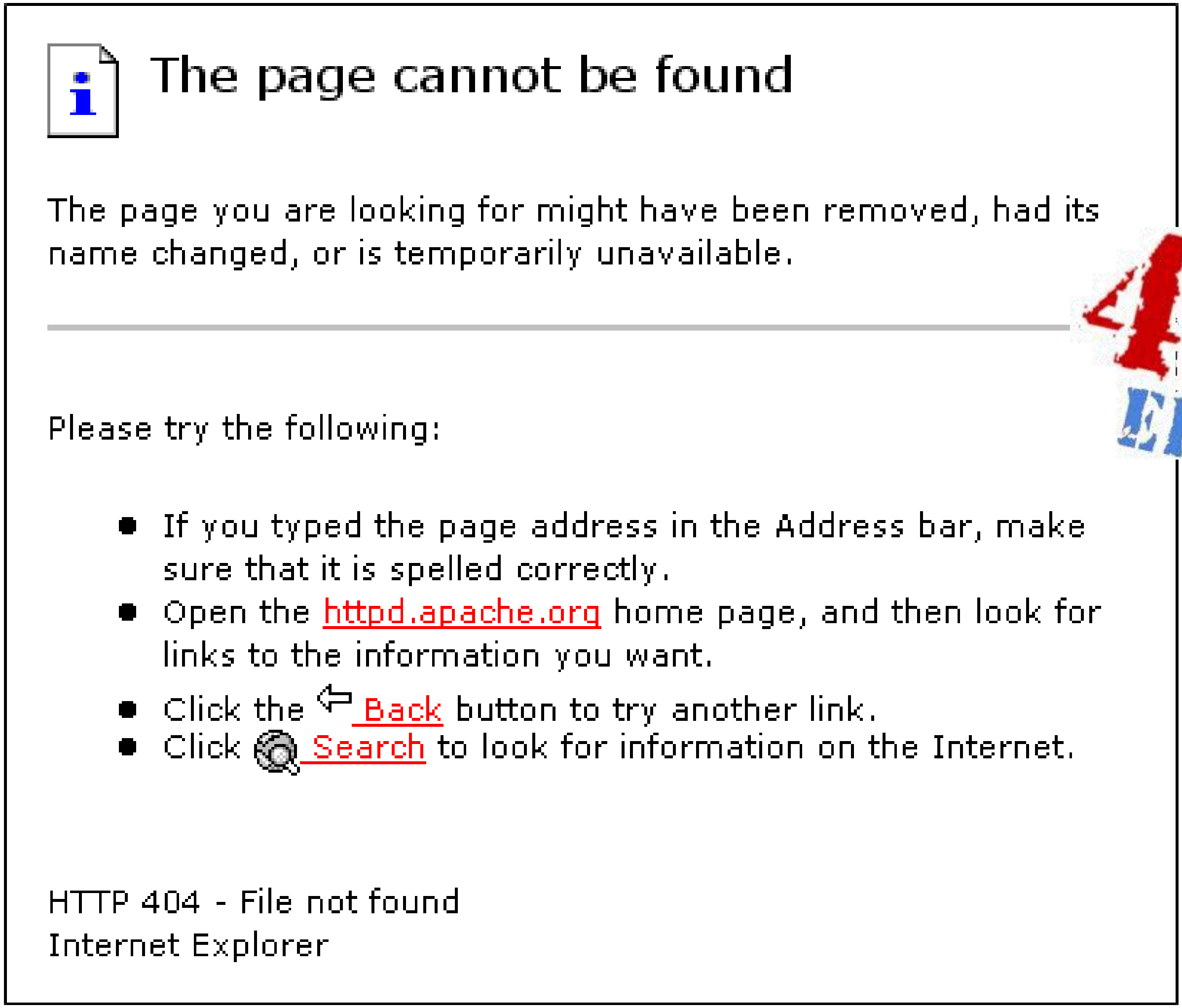
- 1 year - 80% of documents become inaccessible

  (Ntoulas, Cho and Olson. 2004)

- 27 months - 13% of web references disappear

  (http://webcitation.org/. 2007)

Will we face a Digital Dark Age?



The page cannot be found

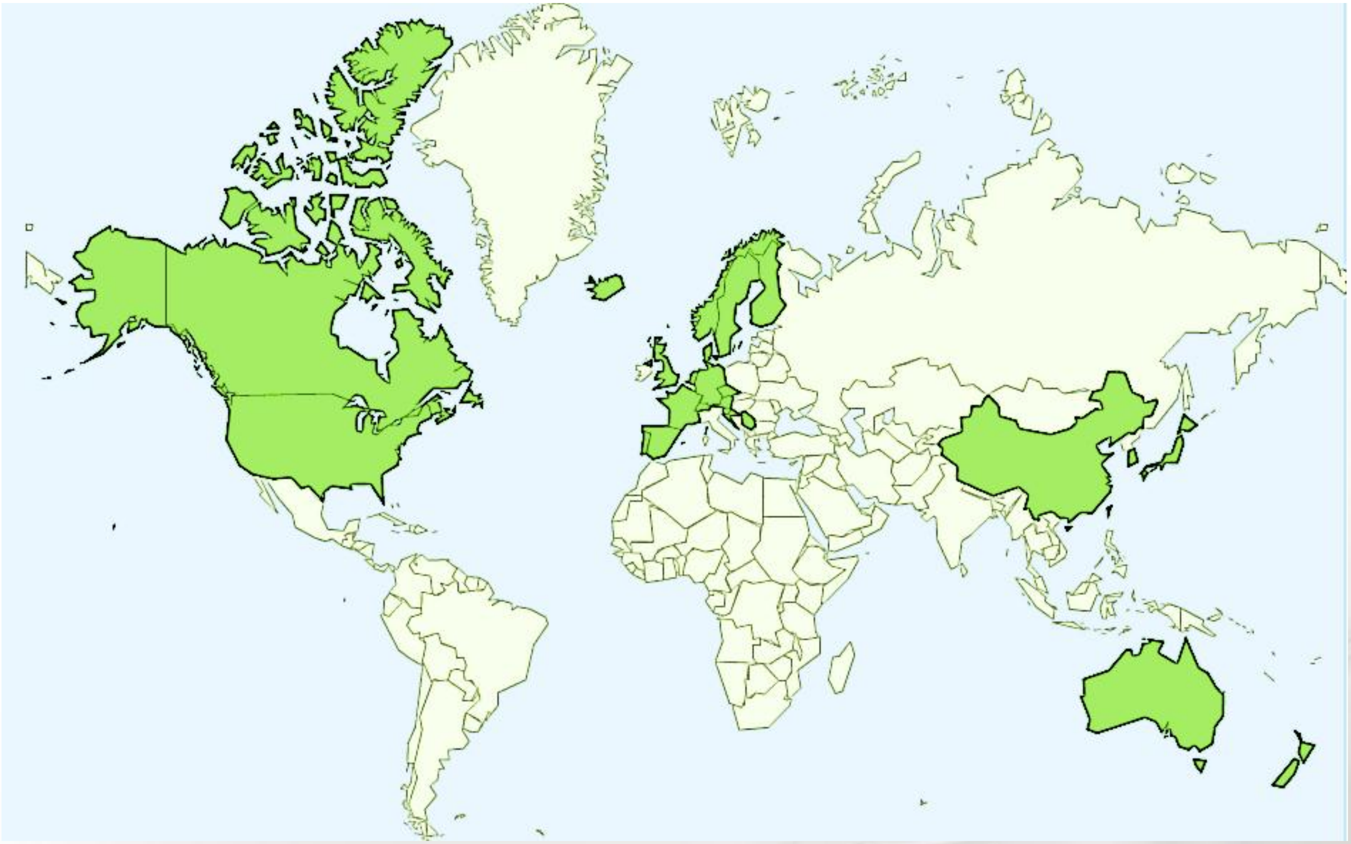
The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

Please try the following:

- If you typed the page address in the Address bar, make sure that it is spelled correctly.
- Open the httpd.apache.org home page, and then look for links to the information you want.
- Click the Back button to try another link.
- Click Search to look for information on the Internet.

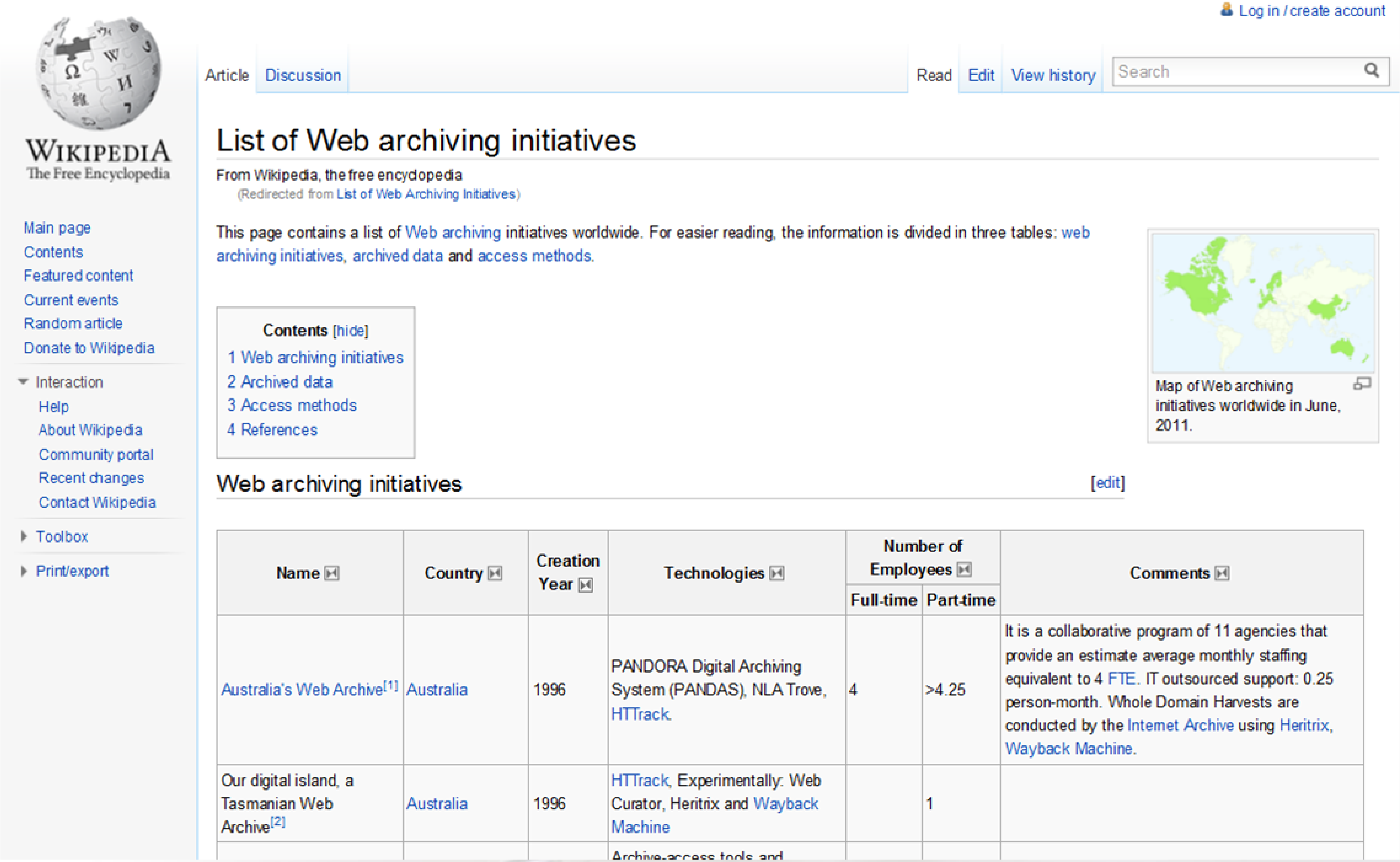
HTTP 404 - File not found
Internet Explorer

2010: Worldwide Web Archiving Initiatives



- +42 initiatives in 26 countries
- +180 billions of web contents since 1996 (6.6 PB)

Wikipedia Page with Updated List



List of Web archiving initiatives

From Wikipedia, the free encyclopedia
(Redirected from List of Web Archiving Initiatives)

This page contains a list of Web archiving initiatives worldwide. For easier reading, the information is divided in three tables: web archiving initiatives, archived data and access methods.

Map of Web archiving initiatives worldwide in June, 2011.

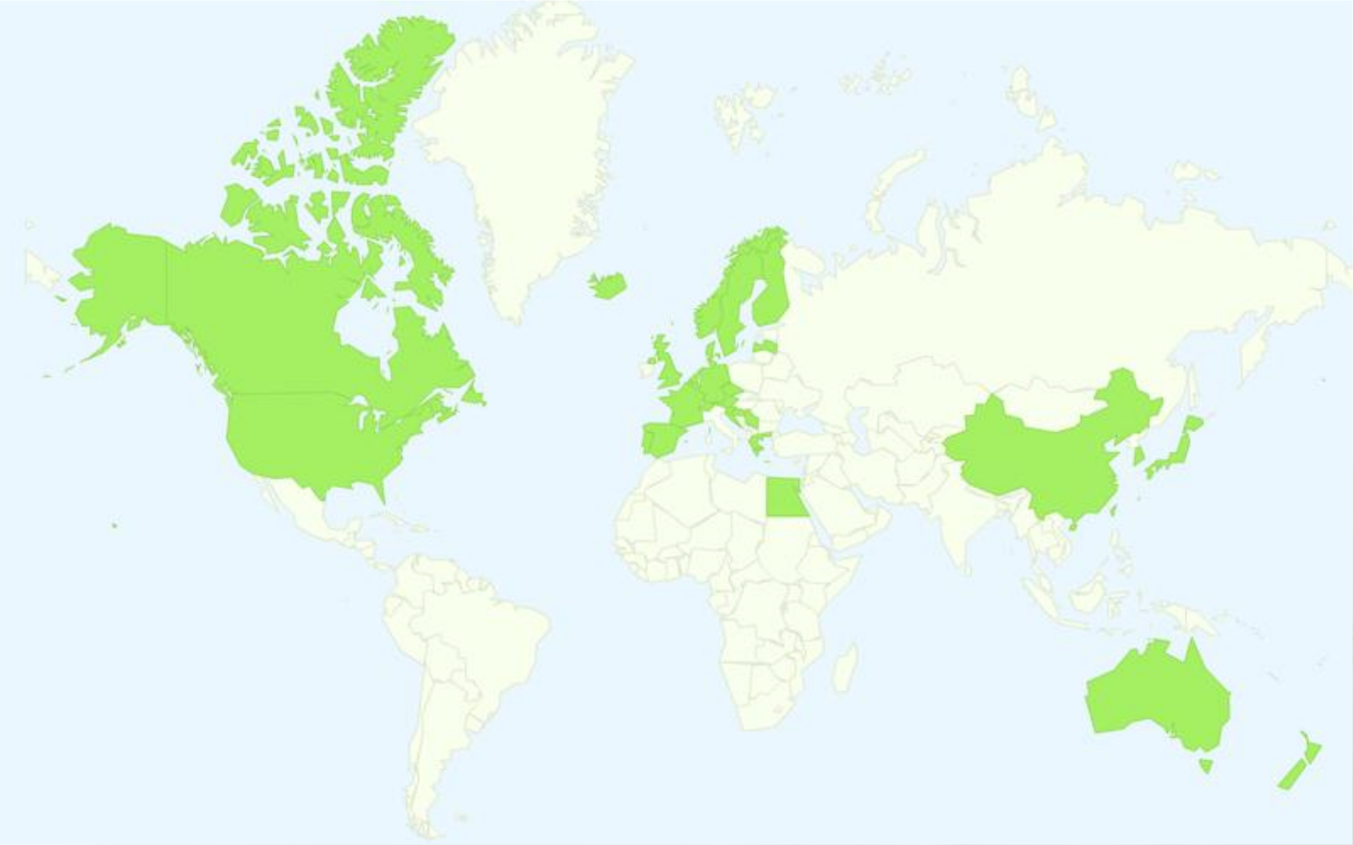
Web archiving initiatives

[edit]

Name	Country	Creation Year	Technologies	Number of Employees		Comments
				Full-time	Part-time	
Australia's Web Archive[1]	Australia	1996	PANDORA Digital Archiving System (PANDAS), NLA Trove, HTTrack.	4	>4.25	It is a collaborative program of 11 agencies that provide an estimate average monthly staffing equivalent to 4 FTE. IT outsourced support: 0.25 person-month. Whole Domain Harvests are conducted by the Internet Archive using Heritrix, Wayback Machine.
Our digital island, a Tasmanian Web Archive[2]	Australia	1996	HTTrack, Experimentally: Web Curator, Heritrix and Wayback Machine		1	
			Archive-access tools and			

http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

2013: Worldwide Web Archiving Initiatives



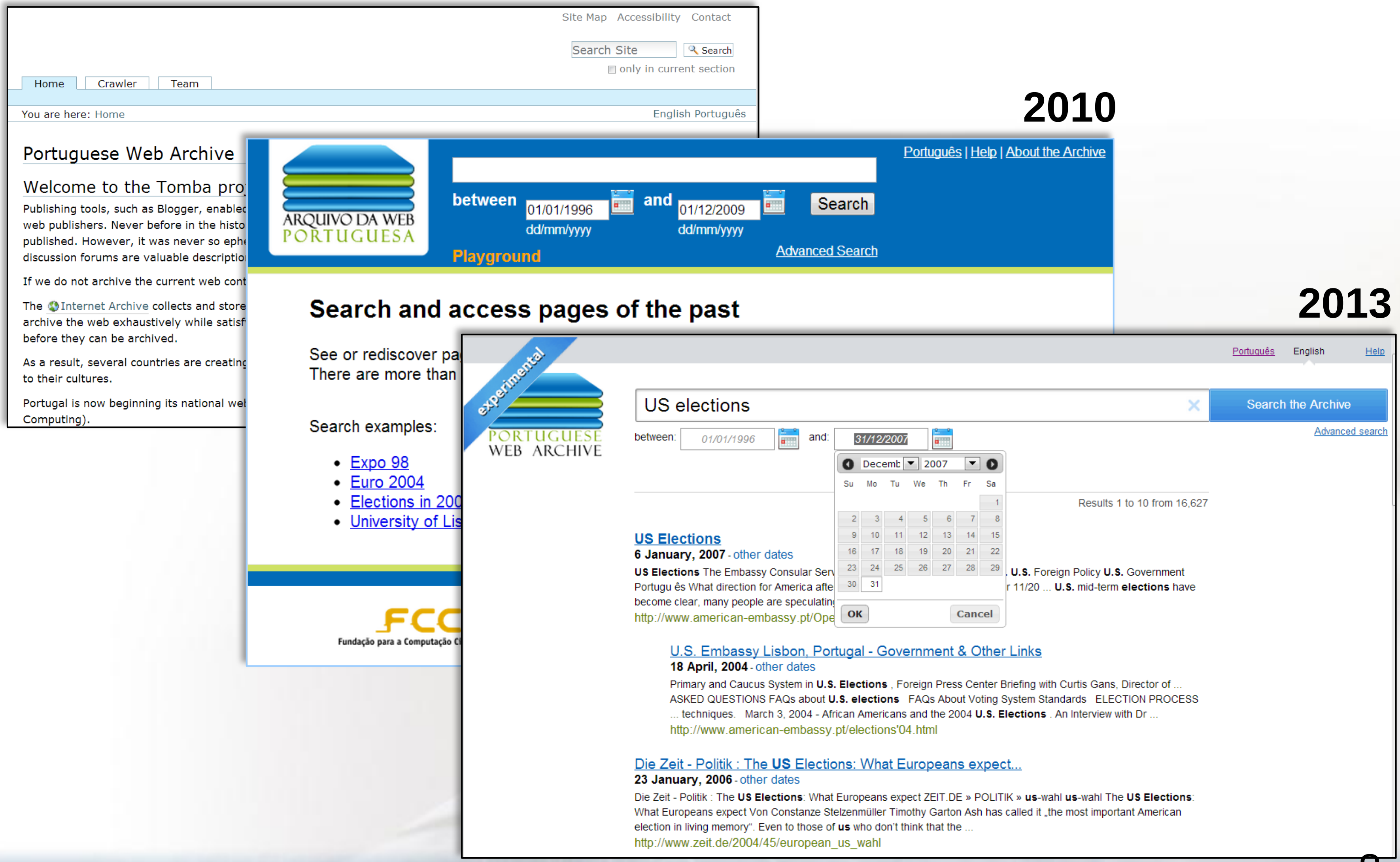
- +77 initiatives in 39 countries
- +294 billions of web contents since 1996 (8.5 PB)

Portuguese Web Archive

2008

2010

2013



Site Map Accessibility Contact

Search Site 🔍 Search

□ only in current section

Home | Crawler | Team

You are here: Home

English Português

Portuguese Web Archive

Welcome to the Tomba pro

Publishing tools, such as Blogger, enabled
web publishers. Never before in the histo
published. However, it was never so ephe
discussion forums are valuable descriptio

If we do not archive the current web cont

The Internet Archive collects and store
archive the web exhaustively while satisf
before they can be archived.

As a result, several countries are creating
to their cultures.

Portugal is now beginning its national we
Computing).

ARQUIVO DA WEB PORTUGUESA

Português | Help | About the Archive

between 01/01/1996 dd/mm/yyyy and 01/12/2009 dd/mm/yyyy Search

Playground

Advanced Search

Search and access pages of the past

See or rediscover pa
There are more than

Search examples:

- Expo 98
- Euro 2004
- Elections in 200
- University of Lis

FCC Fundação para a Computação C

experimental

PORTUGUESE WEB ARCHIVE

Português English Help

US elections ✕ Search the Archive

Advanced search

between: 01/01/1996 and: 31/12/2007

Decemb 2007

Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

OK Cancel

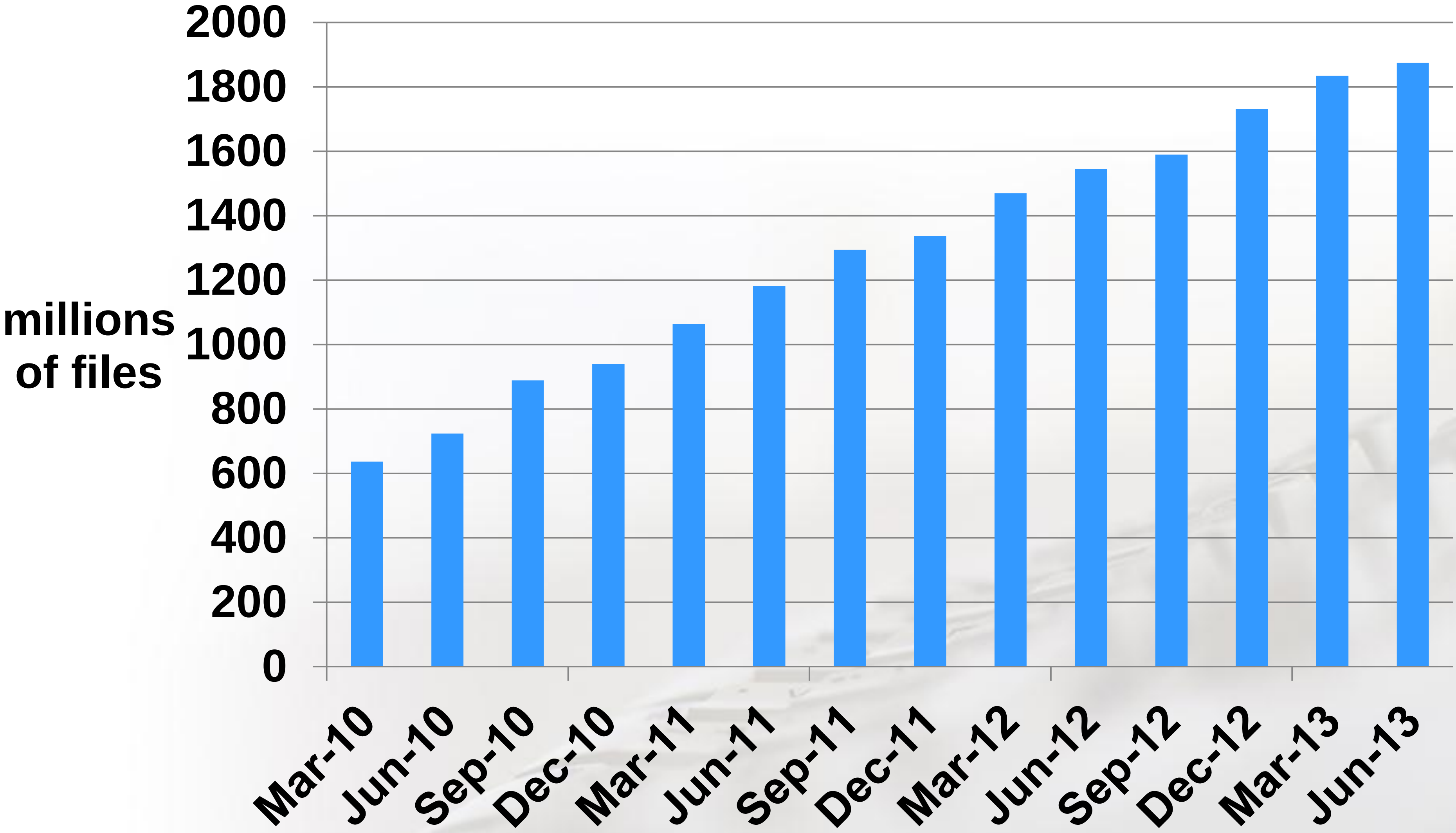
Results 1 to 10 from 16,627

US Elections
6 January, 2007 - other dates
US Elections The Embassy Consular Serv **U.S.** Foreign Policy **U.S.** Government
Portugu ês What direction for America afte r 11/20 ... **U.S.** mid-term **elections** have
become clear, many people are speculatin
http://www.american-embassy.pt/Ope

U.S. Embassy Lisbon, Portugal - Government & Other Links
18 April, 2004 - other dates
Primary and Caucus System in **U.S. Elections** , Foreign Press Center Briefing with Curtis Gans, Director of ...
ASKED QUESTIONS FAQs about **U.S. elections** FAQs About Voting System Standards ELECTION PROCESS
... techniques. March 3, 2004 - African Americans and the 2004 **U.S. Elections** . An Interview with Dr ...
http://www.american-embassy.pt/elections'04.html

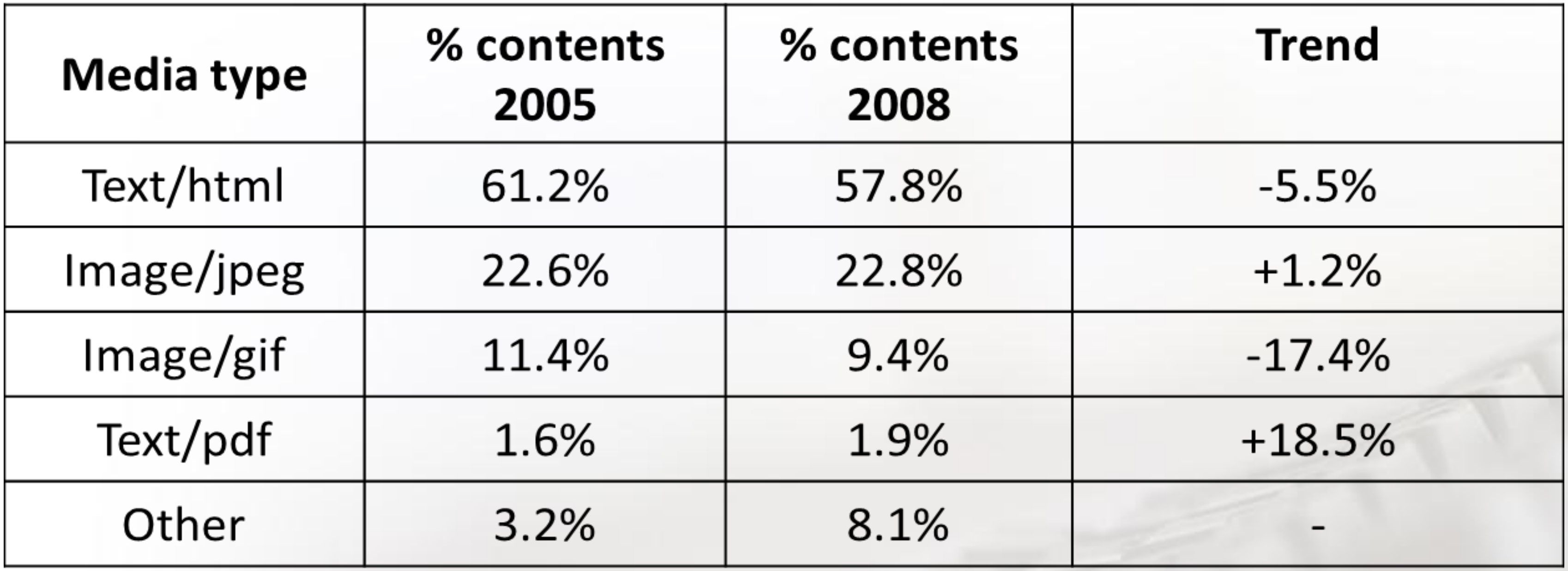
Die Zeit - Politik : The **US** Elections: What Europeans expect...
23 January, 2006 - other dates
Die Zeit - Politik : The **US Elections**: What Europeans expect ZEIT.DE » POLITIK » **us**-wahl **us**-wahl The **US Elections**:
What Europeans expect Von Constanze Stelzenmüller Timothy Garton Ash has called it „the most important American
election in living memory". Even to those of **us** who don't think that the ...
http://www.zeit.de/2004/45/european_us_wahl

1.8 Billion Archived Files (52 TB)



What can we do with all this data?

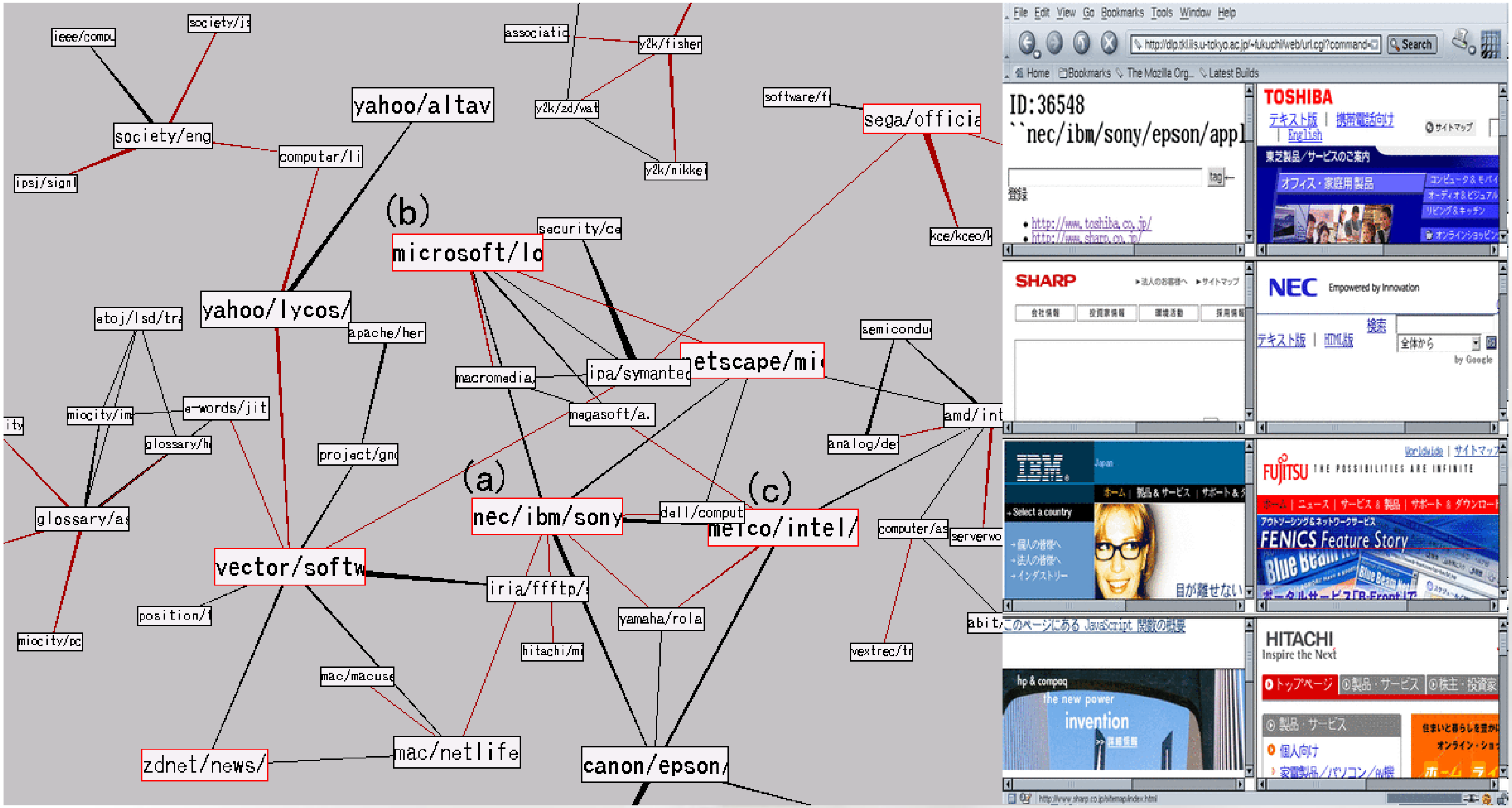
Portuguese WA: Web Characterization



Media type	% contents 2005	% contents 2008	Trend
Text/html	61.2%	57.8%	-5.5%
Image/jpeg	22.6%	22.8%	+1.2%
Image/gif	11.4%	9.4%	-17.4%
Text/pdf	1.6%	1.9%	+18.5%
Other	3.2%	8.1%	-

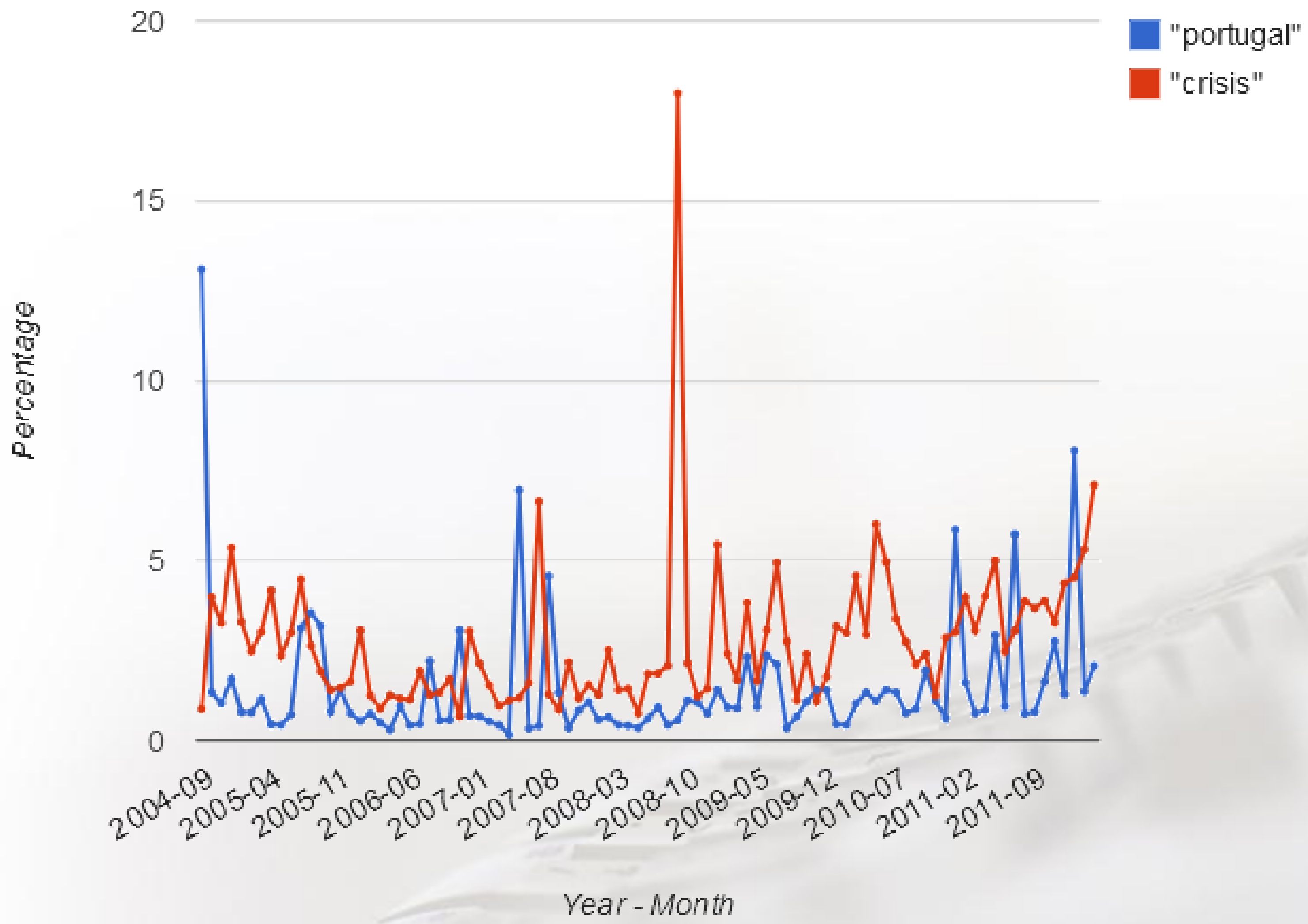
Trends in Web Characteristics, *7th Latin American Web Congress* 2009.

Japanese WA: Evolution of Web Communities



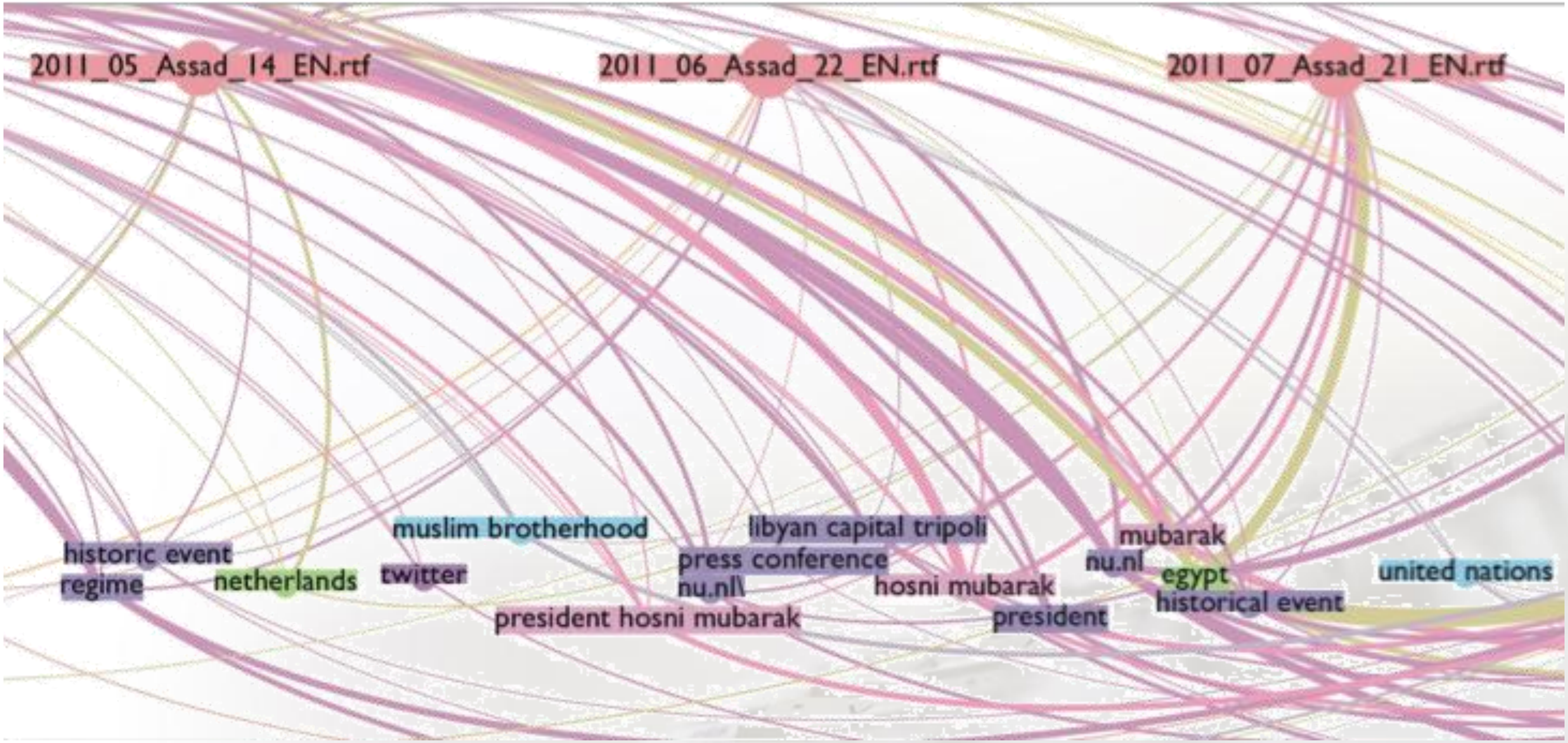
Extracting Evolution of Web Communities from a Series of Web Archives, *14th ACM Conference on Hypertext and Hypermedia* 2003.

UK WA: Word Frequency Analysis



http://www.webarchive.org.uk/ukwa/ngram/

WebART: Co-word Analysis



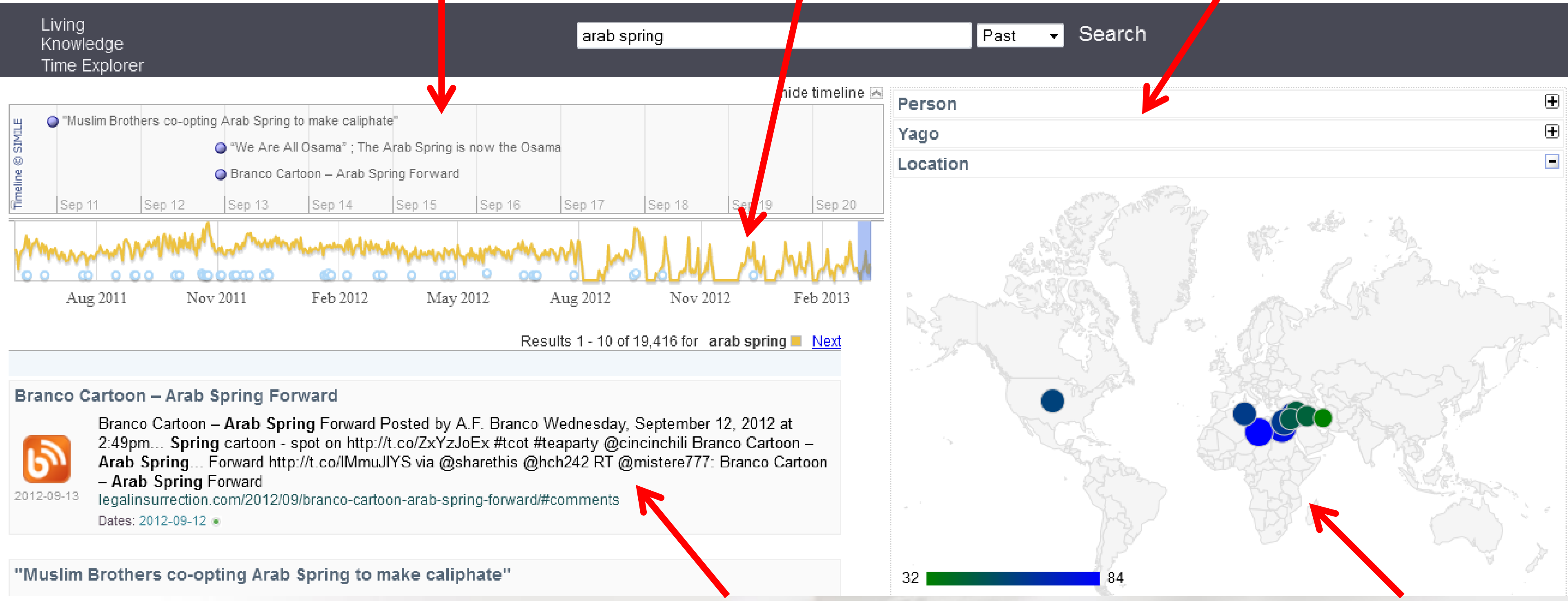
http://www.webarchiving.nl/

Living Knowledge (Yahoo!): News Analysis

timeline

frequency graph

entity selection



Results 1 - 10 of 19,416 for **arab spring** 🟨 Next

Branco Cartoon – Arab Spring Forward

Branco Cartoon – **Arab Spring** Forward Posted by A.F. Branco Wednesday, September 12, 2012 at 2:49pm... **Spring** cartoon - spot on http://t.co/ZxYzJoEx #tcot #teaparty @cincinchili Branco Cartoon – **Arab Spring**... Forward http://t.co/lMmuJIYS via @sharethis @hch242 RT @mistere777: Branco Cartoon – **Arab Spring** Forward

2012-09-13

legalinsurrection.com/2012/09/branco-cartoon-arab-spring-forward/#comments

Dates: 2012-09-12

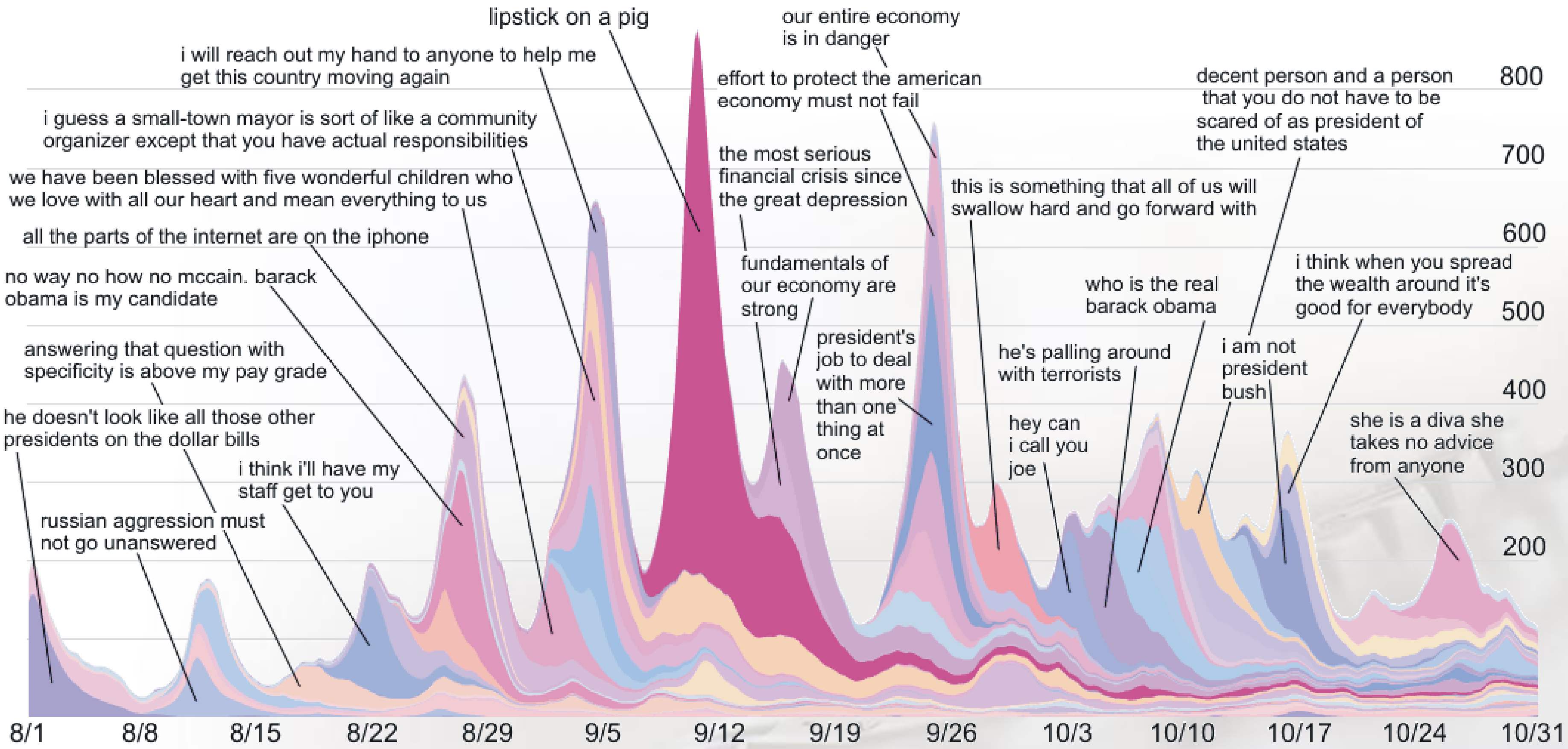
"Muslim Brothers co-opting Arab Spring to make caliphate"

results

geomapping

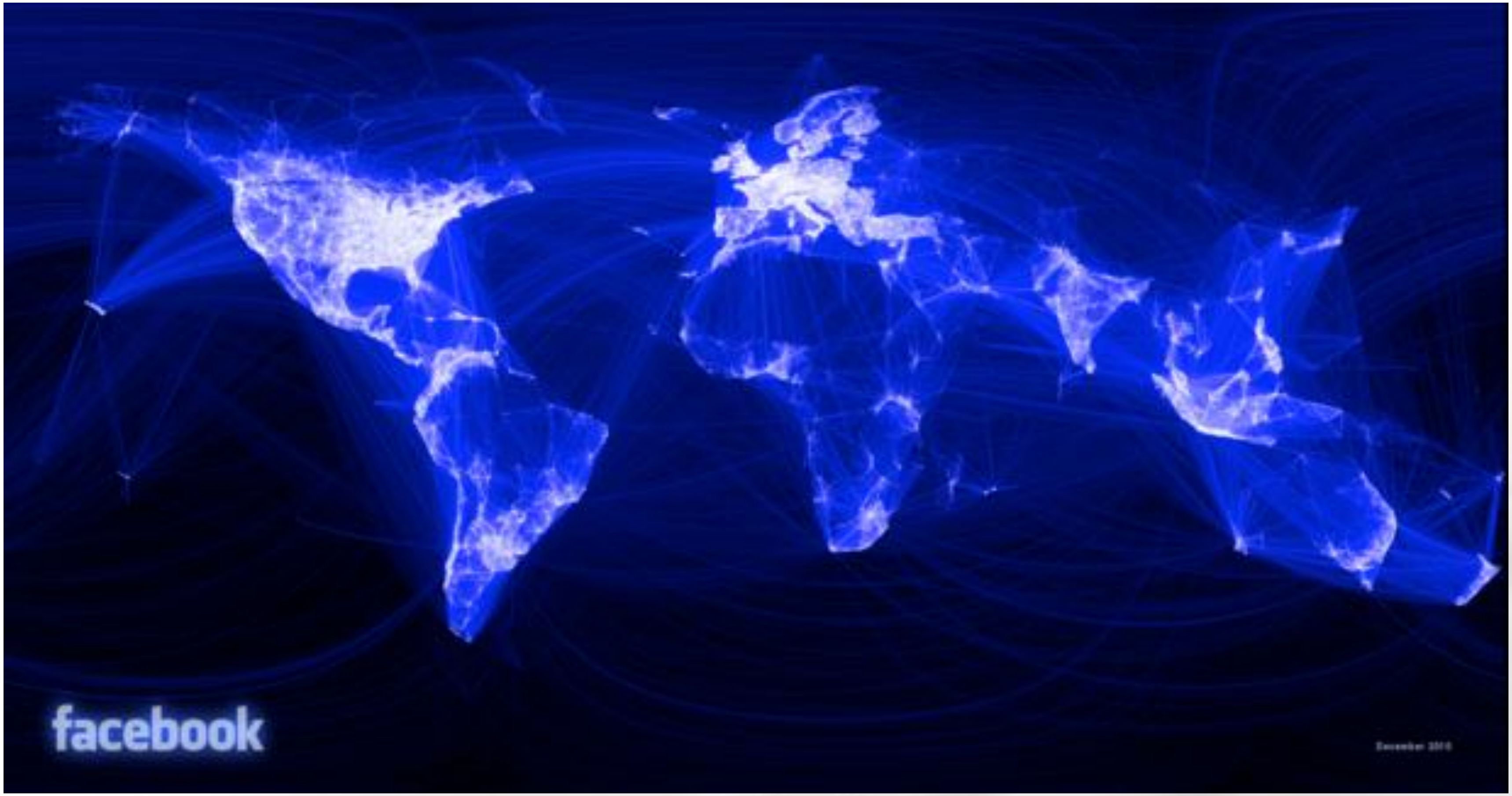
Searching through time in the New York Times, *Human Computer and Information Retrieval* 2010.

MemeTracker: News Analysis



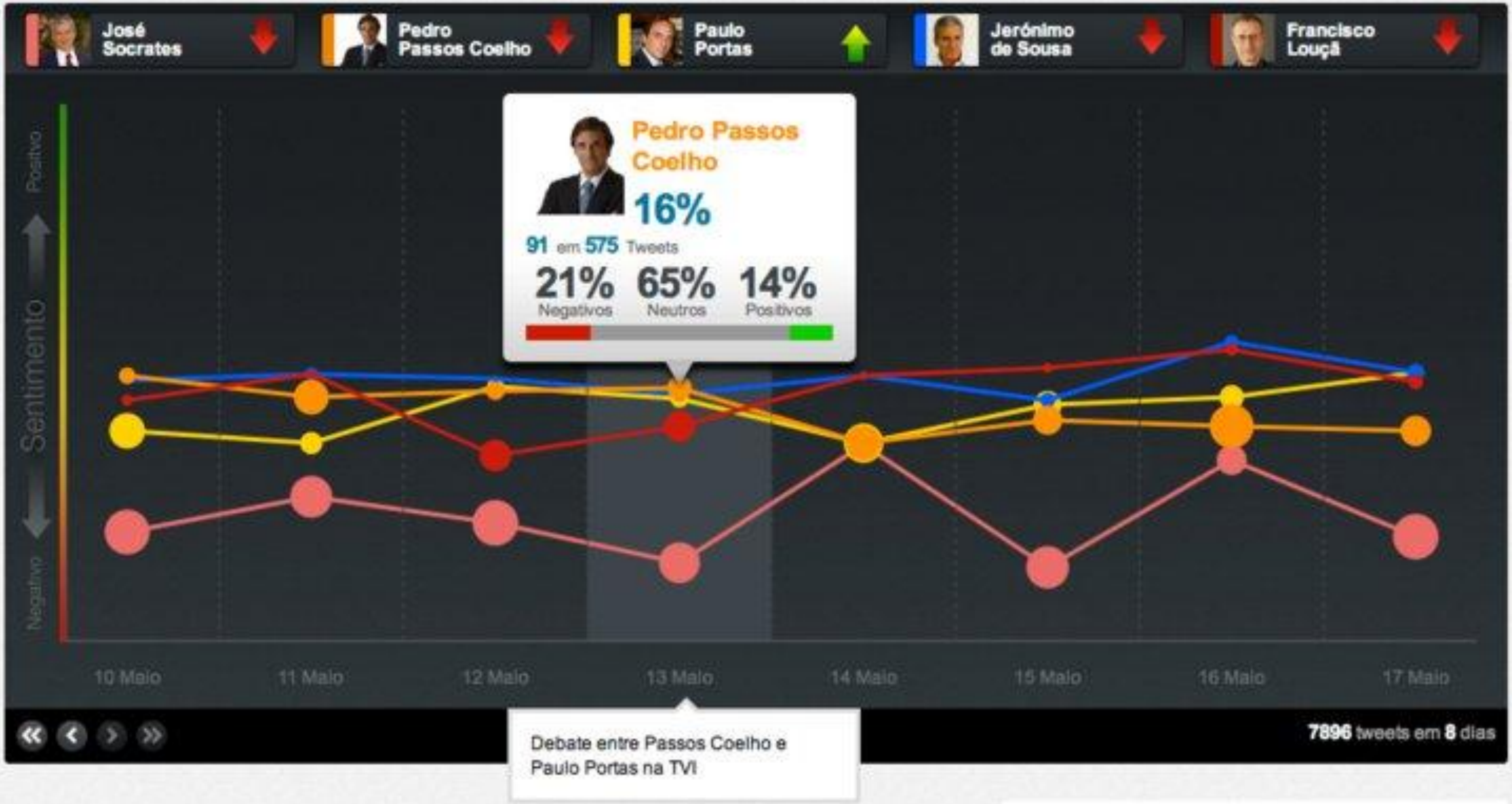
Meme-tracking and the Dynamics of the News Cycle, *Knowledge Discovery and Data Mining* 2009.

Facebook: Social Analysis



Paul Butler created this friend relationship visualization map using Facebook data.

Twitómetro: Sentiment Analysis



Pedro Passos Coelho

16%

91 em 575 Tweets

21% Negativos | 65% Neutros | 14% Positivos

Debate entre Passos Coelho e Paulo Portas na TVI

7896 tweets em 8 dias

http://dmir.inesc-id.pt/project/Reaction

NYT Archive: Forecast Events

In deep drought, at 104 degrees, dozens of Africans are dying.

New York Times 02/17/2006

Angola cholera cases rise sharply after floods.

New York Times 01/30/2007

Mining the Web to Predict Future Events, *Web Search and Data Mining* 2013.

What can we do with all this data?

all kinds of machine learning over time

model the past and predict the future

But, what do most web archive users need?

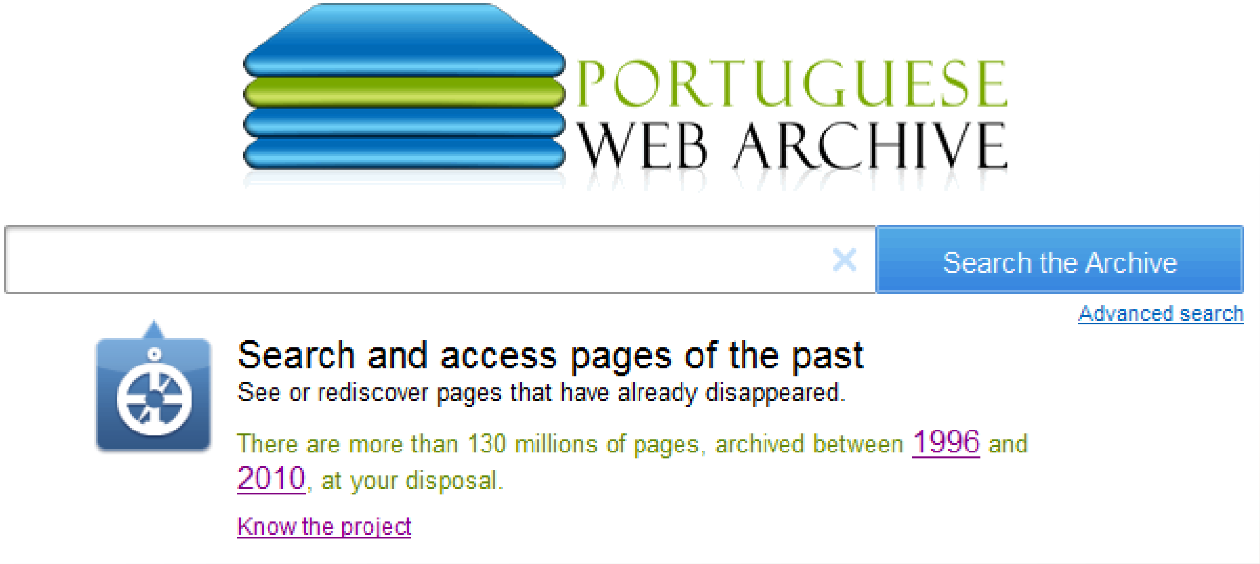
Use Cases

- **User** visits a missing bookmark
- **Journalist** investigates past information
- **Webmaster** recovers the lost site
- **Historian** searches for digital documents
- **Web designer** creates portfolio of sites
- **Professor** downloads missing slides
- **Lawyer** looks for evidences

Users don’t understand Web Archives

What do you want? I don’t know!

PWA Search System



Search and access pages of the past

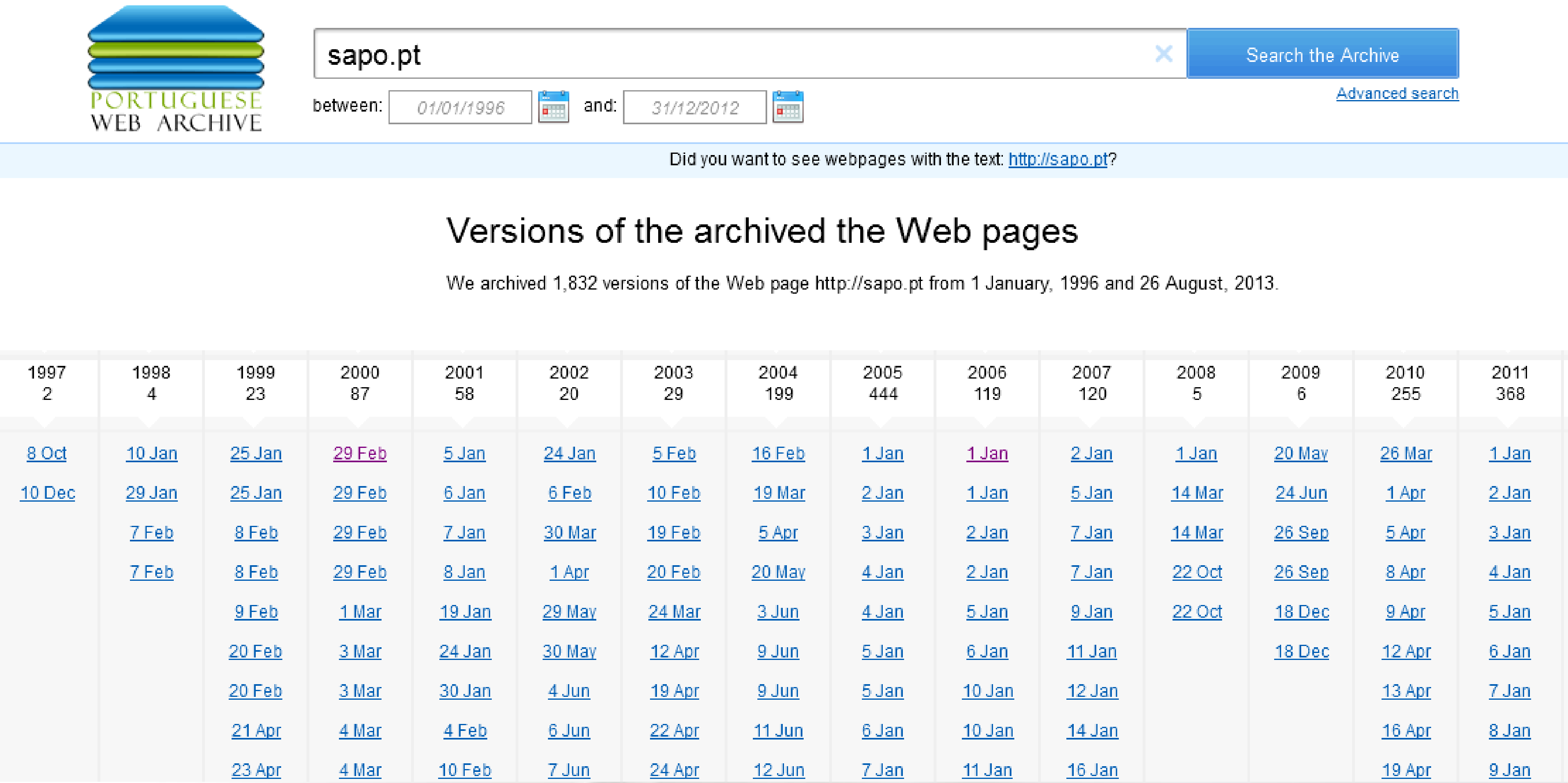
See or rediscover pages that have already disappeared.

There are more than 130 millions of pages, archived between 1996 and 2010, at your disposal.

Know the project

- Available since 2010: http://archive.pt
- 1.2 billion documents
  - searchable by full-text and URL
  - range between 1996 and 2012

URL Search



PORTUGUESE
WEB ARCHIVE

sapo.pt ✕ Search the Archive

between: 01/01/1996 and: 31/12/2012

Advanced search

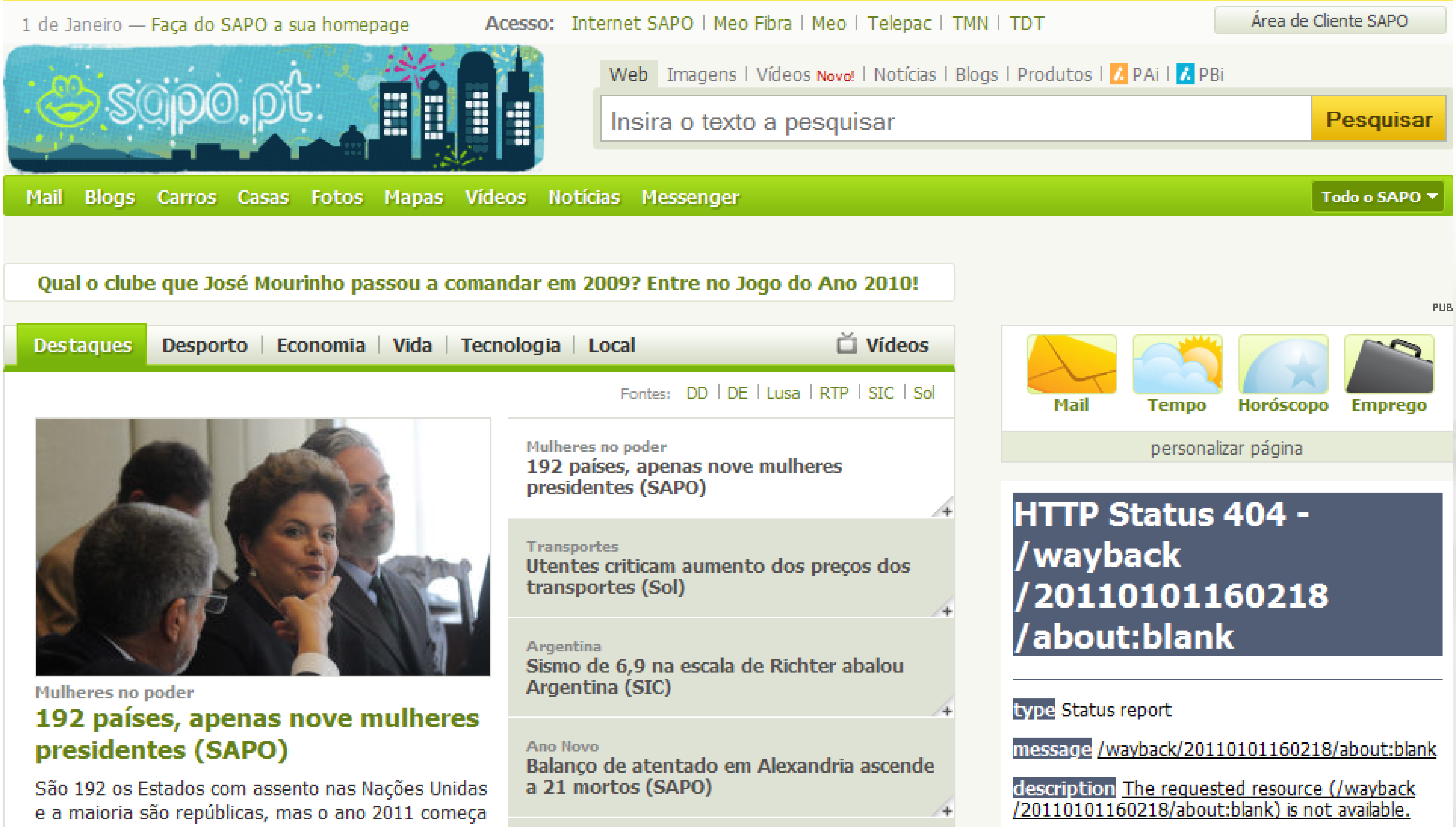
Did you want to see webpages with the text: http://sapo.pt?

Versions of the archived the Web pages

We archived 1,832 versions of the Web page http://sapo.pt from 1 January, 1996 and 26 August, 2013.

1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
2	4	23	87	58	20	29	199	444	119	120	5	6	255	368
8 Oct	10 Jan	25 Jan	29 Feb	5 Jan	24 Jan	5 Feb	16 Feb	1 Jan	1 Jan	2 Jan	1 Jan	20 May	26 Mar	1 Jan
10 Dec	29 Jan	25 Jan	29 Feb	6 Jan	6 Feb	10 Feb	19 Mar	2 Jan	1 Jan	5 Jan	14 Mar	24 Jun	1 Apr	2 Jan
	7 Feb	8 Feb	29 Feb	7 Jan	30 Mar	19 Feb	5 Apr	3 Jan	2 Jan	7 Jan	14 Mar	26 Sep	5 Apr	3 Jan
	7 Feb	8 Feb	29 Feb	8 Jan	1 Apr	20 Feb	20 May	4 Jan	2 Jan	7 Jan	22 Oct	26 Sep	8 Apr	4 Jan
		9 Feb	1 Mar	19 Jan	29 May	24 Mar	3 Jun	4 Jan	5 Jan	9 Jan	22 Oct	18 Dec	9 Apr	5 Jan
		20 Feb	3 Mar	24 Jan	30 May	12 Apr	9 Jun	5 Jan	6 Jan	11 Jan		18 Dec	12 Apr	6 Jan
		20 Feb	3 Mar	30 Jan	4 Jun	19 Apr	9 Jun	5 Jan	10 Jan	12 Jan			13 Apr	7 Jan
		21 Apr	4 Mar	4 Feb	6 Jun	22 Apr	11 Jun	6 Jan	10 Jan	14 Jan			16 Apr	8 Jan
		23 Apr	4 Mar	10 Feb	7 Jun	24 Apr	12 Jun	7 Jan	11 Jan	16 Jan			19 Apr	9 Jan

SAPO.PT 2011



1 de Janeiro — Faça do SAPO a sua homepage

Acesso: Internet SAPO | Meo Fibra | Meo | Telepac | TMN | TDT

Área de Cliente SAPO

Web | Imagens | Vídeos Novo! | Notícias | Blogs | Produtos | PAi | PBi

Insira o texto a pesquisar

Pesquisar

Mail Blogs Carros Casas Fotos Mapas Vídeos Notícias Messenger

Todo o SAPO

Qual o clube que José Mourinho passou a comandar em 2009? Entre no Jogo do Ano 2010!

PUB

Destaques | Desporto | Economia | Vida | Tecnologia | Local | Vídeos

Fontes: DD | DE | Lusa | RTP | SIC | Sol

Mulheres no poder

192 países, apenas nove mulheres presidentes (SAPO)

São 192 os Estados com assento nas Nações Unidas e a maioria são repúblicas, mas o ano 2011 começa

Mulheres no poder

192 países, apenas nove mulheres presidentes (SAPO)

Transportes

Utentes criticam aumento dos preços dos transportes (Sol)

Argentina

Sismo de 6,9 na escala de Richter abalou Argentina (SIC)

Ano Novo

Balanço de atentado em Alexandria ascende a 21 mortos (SAPO)

Mail | Tempo | Horóscopo | Emprego

personalizar página

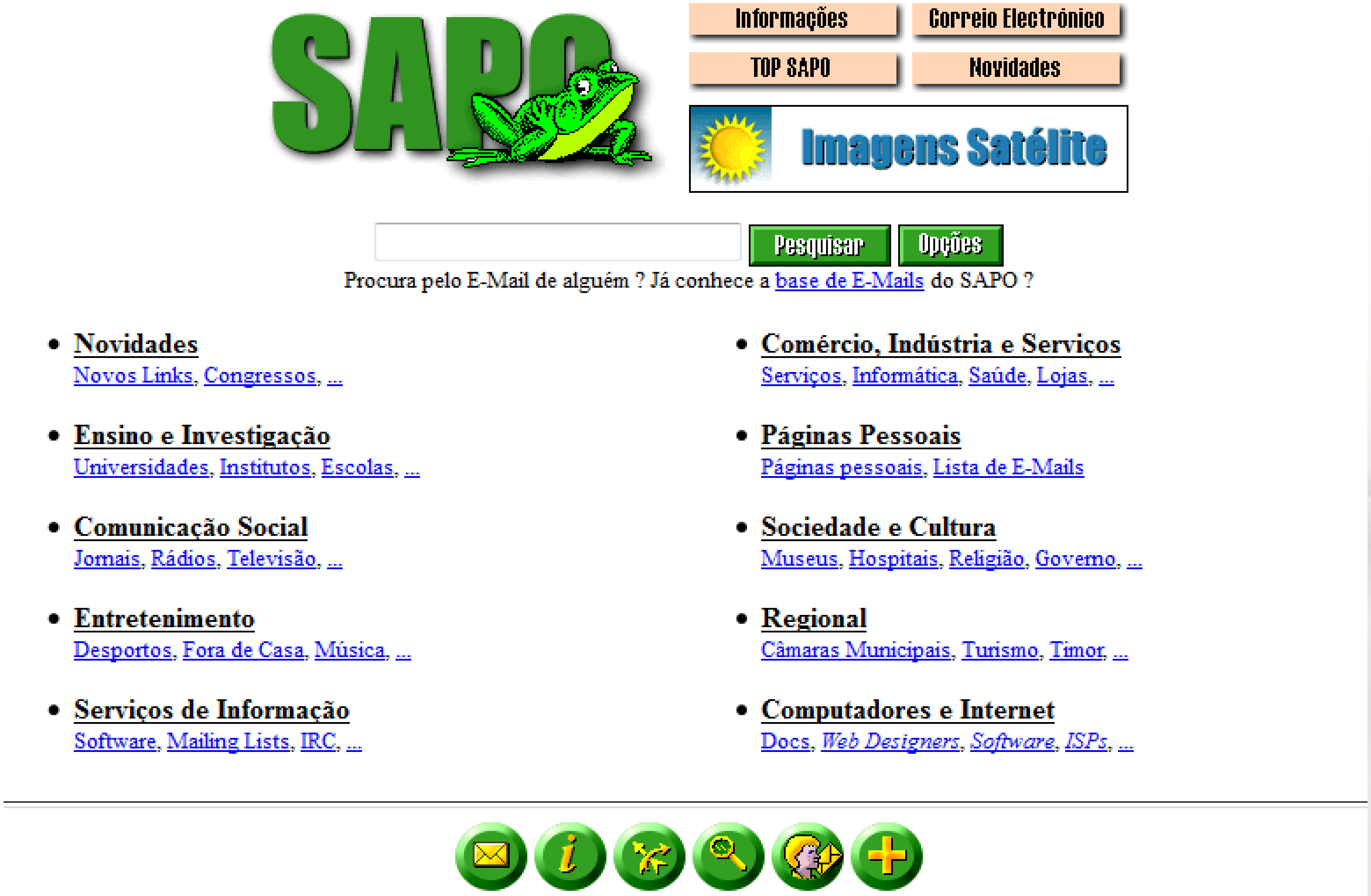
HTTP Status 404 - /wayback /20110101160218 /about:blank

type Status report

message /wayback/20110101160218/about:blank

description The requested resource (/wayback /20110101160218/about:blank) is not available.

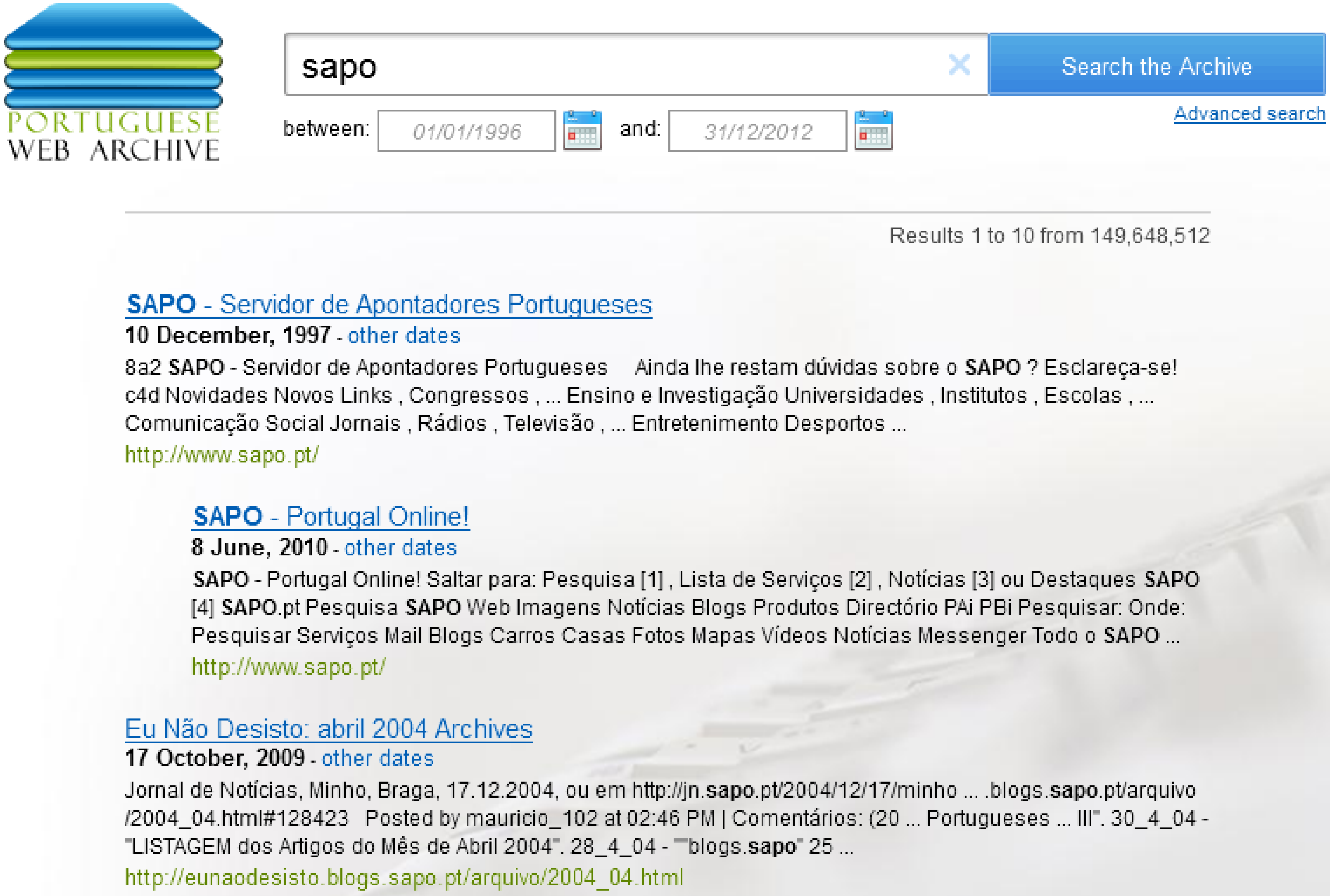
SAPO.PT 1997



Procura pelo E-Mail de alguém ? Já conhece a base de E-Mails do SAPO ?

- **Novidades**
  Novos Links, Congressos, ...
- **Ensino e Investigação**
  Universidades, Institutos, Escolas, ...
- **Comunicação Social**
  Jornais, Rádios, Televisão, ...
- **Entretenimento**
  Desportos, Fora de Casa, Música, ...
- **Serviços de Informação**
  Software, Mailing Lists, IRC, ...
- **Comércio, Indústria e Serviços**
  Serviços, Informática, Saúde, Lojas, ...
- **Páginas Pessoais**
  Páginas pessoais, Lista de E-Mails
- **Sociedade e Cultura**
  Museus, Hospitais, Religião, Governo, ...
- **Regional**
  Câmaras Municipais, Turismo, Timor, ...
- **Computadores e Internet**
  Docs, *Web Designers*, *Software*, *ISPs*, ...

Full-text Search



Results 1 to 10 from 149,648,512

SAPO - Servidor de Apontadores Portugueses

10 December, 1997 - other dates

8a2 **SAPO** - Servidor de Apontadores Portugueses Ainda lhe restam dúvidas sobre o **SAPO** ? Esclareça-se! c4d Novidades Novos Links , Congressos , ... Ensino e Investigação Universidades , Institutos , Escolas , ... Comunicação Social Jornais , Rádios , Televisão , ... Entretenimento Desportos ...

http://www.sapo.pt/

SAPO - Portugal Online!

8 June, 2010 - other dates

SAPO - Portugal Online! Saltar para: Pesquisa [1] , Lista de Serviços [2] , Notícias [3] ou Destaques **SAPO** [4] **SAPO**.pt Pesquisa **SAPO** Web Imagens Notícias Blogs Produtos Directório PAi PBi Pesquisar: Onde: Pesquisar Serviços Mail Blogs Carros Casas Fotos Mapas Vídeos Notícias Messenger Todo o **SAPO** ...

http://www.sapo.pt/

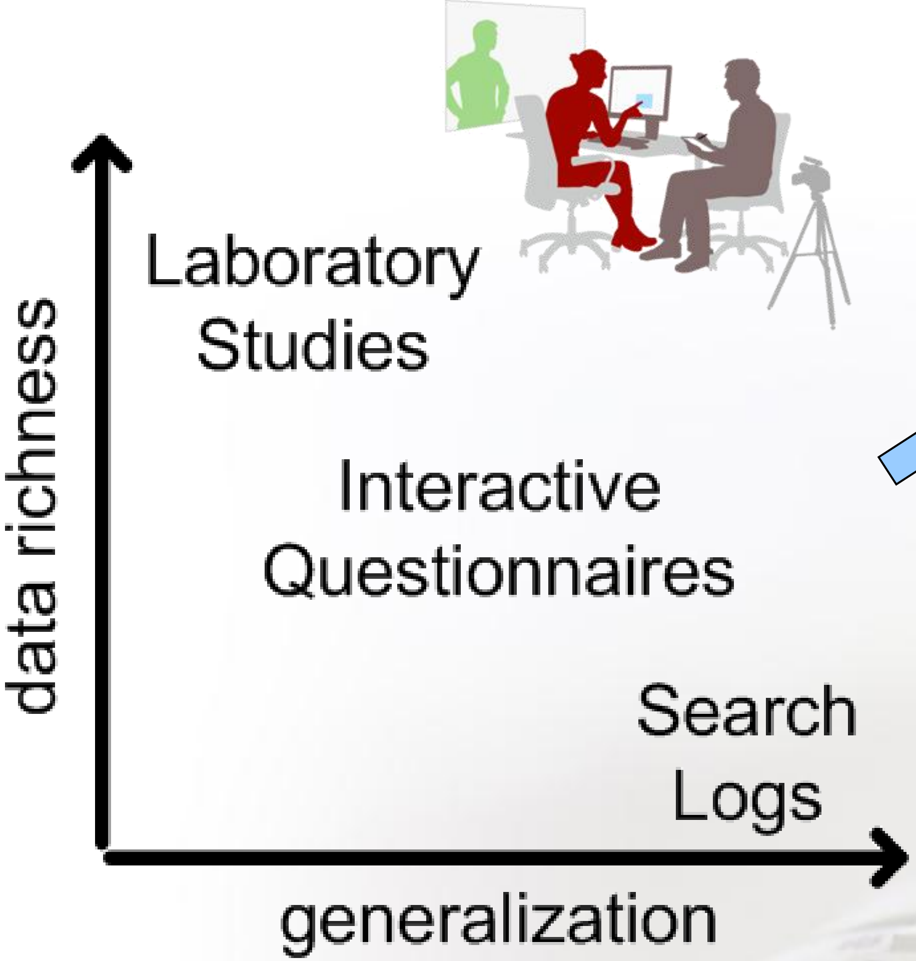
Eu Não Desisto: abril 2004 Archives

17 October, 2009 - other dates

Jornal de Notícias, Minho, Braga, 17.12.2004, ou em http://jn.**sapo**.pt/2004/12/17/minho ... .blogs.**sapo**.pt/arquivo /2004_04.html#128423 Posted by mauricio_102 at 02:46 PM | Comentários: (20 ... Portugueses ... III". 30_4_04 - "LISTAGEM dos Artigos do Mês de Abril 2004". 28_4_04 - ""blogs.**sapo**" 25 ...

http://eunaodesisto.blogs.sapo.pt/arquivo/2004_04.html

Understanding the Users' Information Needs



```
[03/02/2012  21:16:11] QUERY fcul
[03/02/2012  21:16:19] CLICK RANK=1
```

Understanding the Information Needs of Web Archive Users, *10th International Web Archiving Workshop* 2010.

What are the Users' Information Needs?

- **Navigational** – 53% to 81%
  - seeing a web page in the **past** or how it evolved

- **Informational** – 14% to 38%
  - collecting information about a topic written in the **past**

What is the best tool to support **navigational** (and **informational**) information needs?

Searching vs Analytical tools

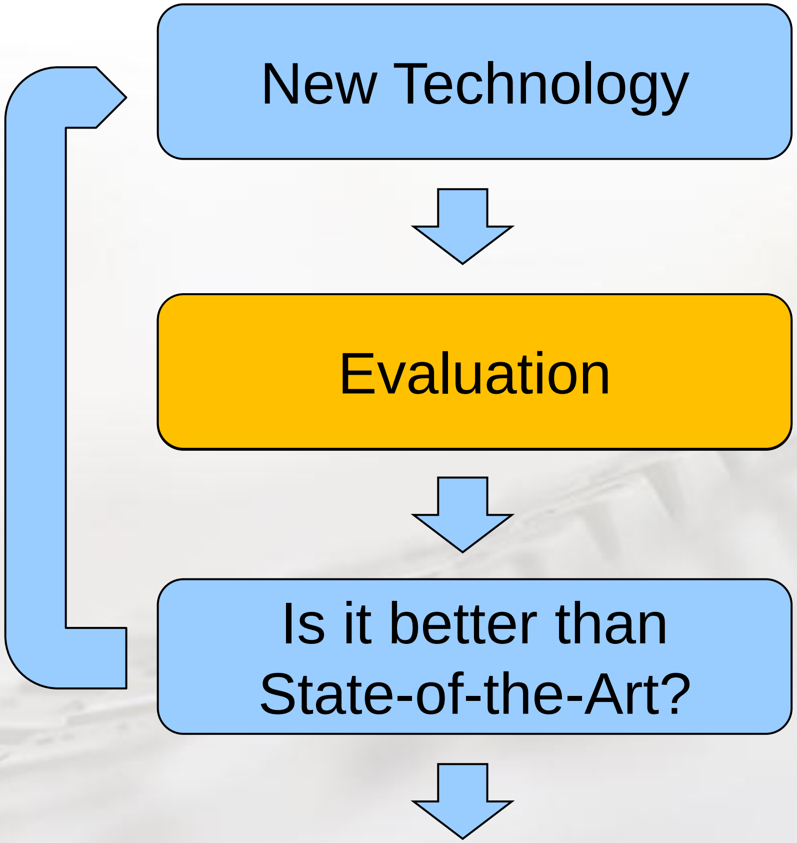
State-of-the-Art

- **URL Search** – Internet Archive’s Wayback Machine
    - difficult to remember or unknown

- **Full-text Search** – Lucene extensions (NutchWAX & Solr)
    - does not scale for large collections
    - slow searches
    - poor quality results

~~How to improve?~~

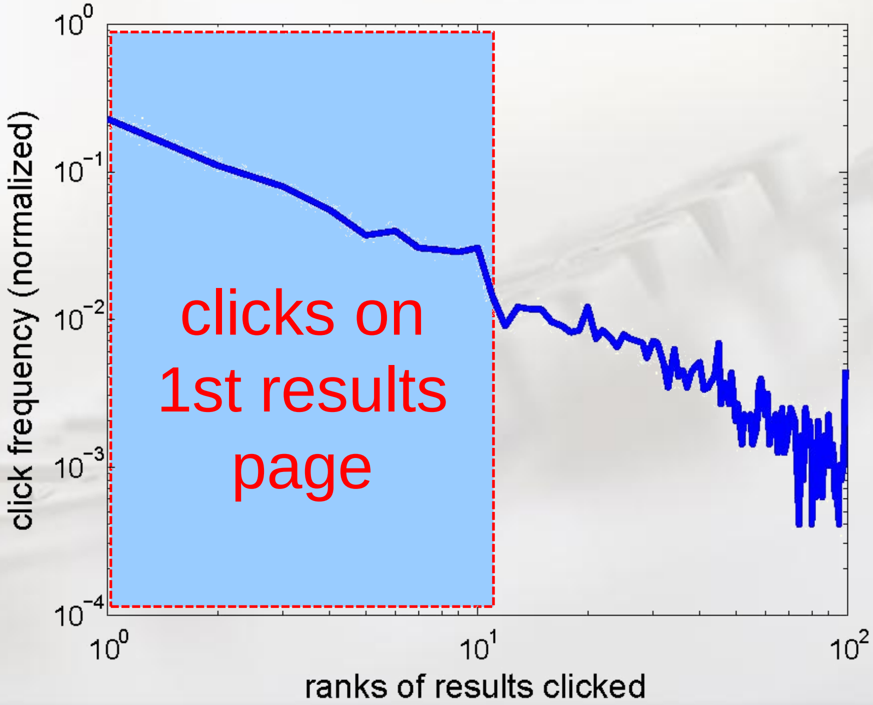
How to evaluate?



Evaluation Methodology

- Test Collection (Cranfield Paradigm):
  - Corpus
    - **What** are the typical web collections?
  - Topics
    - **Why**, **what** and **how** do users search?
  - Relevance Judgments
    - **What** is relevant for users?
  - Measures
    - **What** and **how** many documents do users see?
- Wrong assumptions lead to wrong conclusions

Searching and Clicking



clicks on 1st results page

Evaluation Methodology

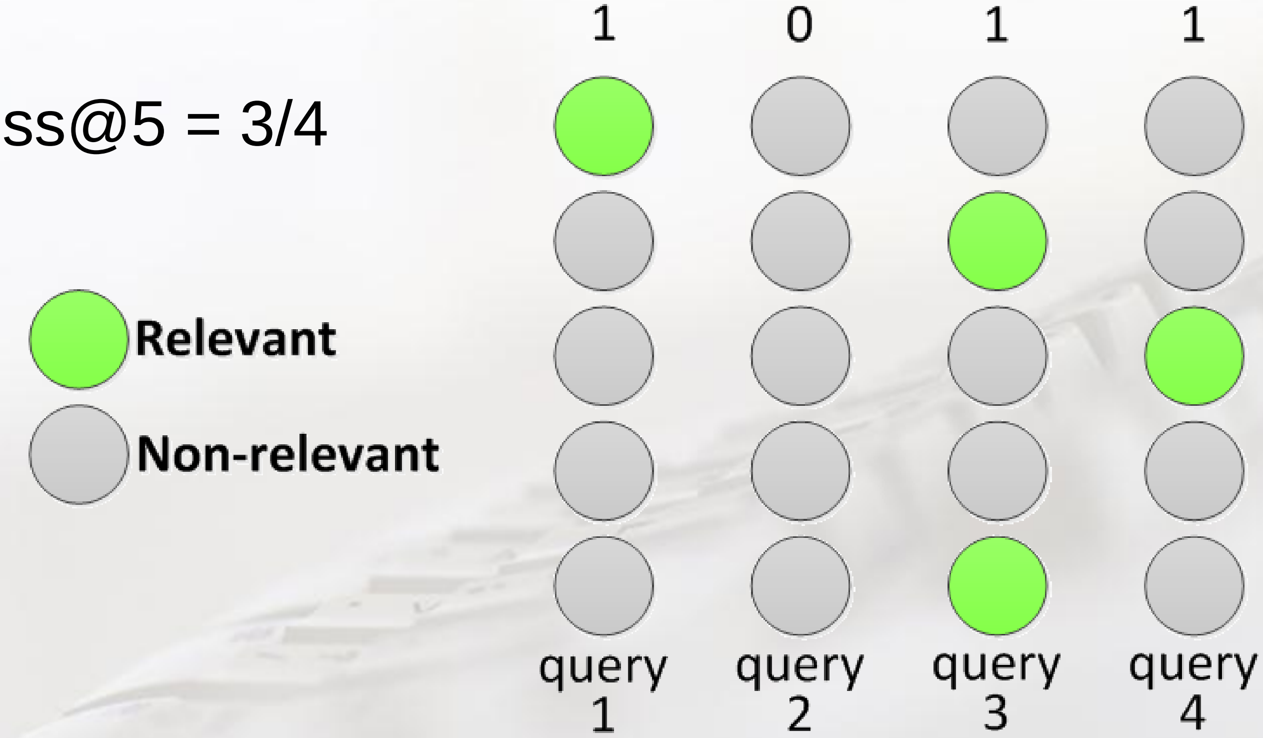
- Test Collection (Cranfield Paradigm):
  - **Corpus**: 6 web collections, 255M contents, 8.9TB
    - broad crawls, selective crawls, integrated collections
  - **Topics**: 50 navigational
    - I need the page of Público newspaper between 1996 and 2000.
  - **Relevance Judgments**: 3 judges, 3-level scale of relevance, 267 822 versions assessed
  - **Measures**: (S@k, NDCG@k, P@k | k=1,5,10)
    - only 14% see the second page (> top 10)

Evaluating Web Archive Search Systems, *13th International Conference on Web Information System Engineering* 2012.

Evaluation Metric: Success@k

- 1 if a relevant version has been found on the top-k
- 0 otherwise

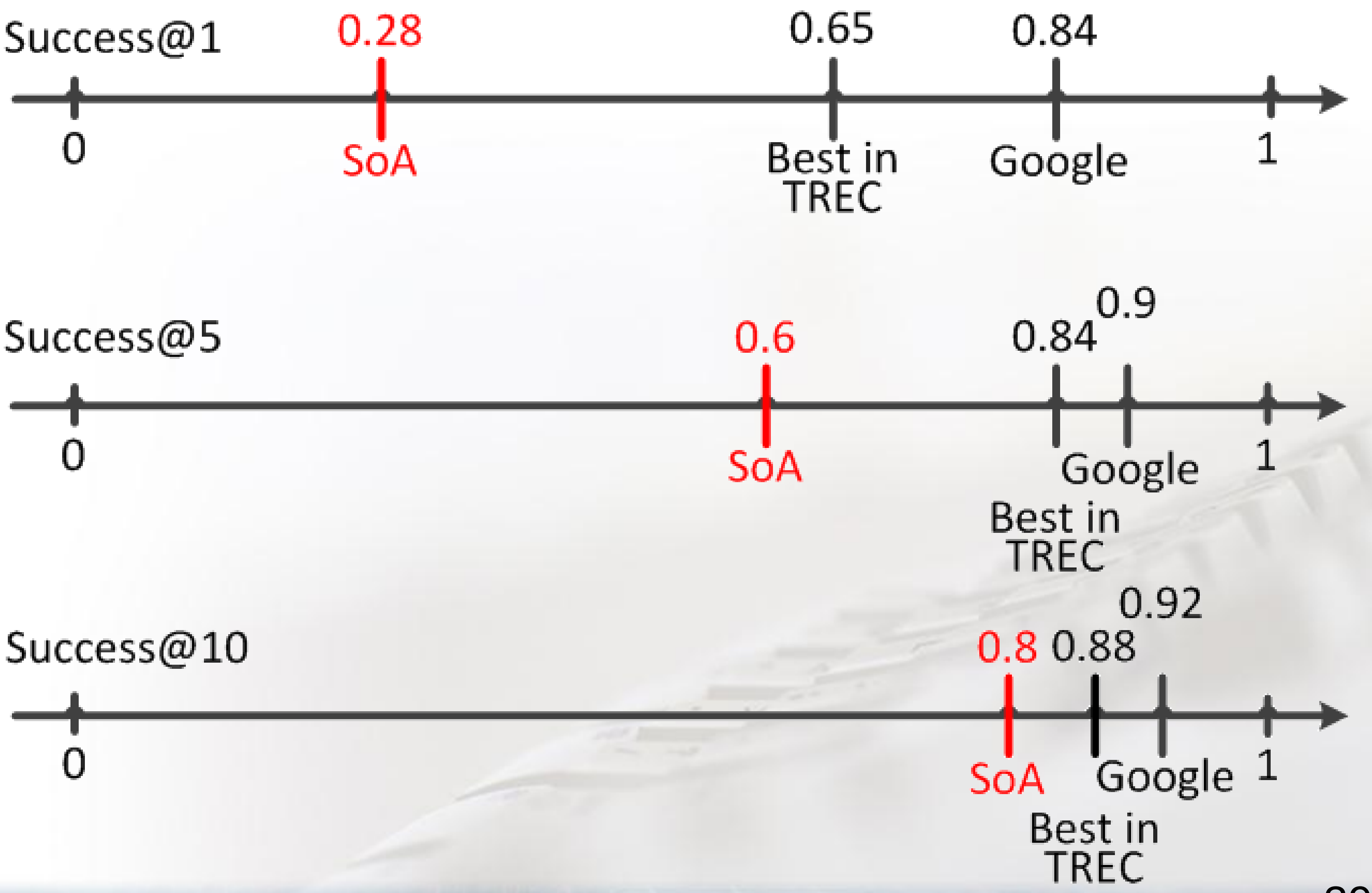
- Example:
  - avg. Success@5 = 3/4



Results

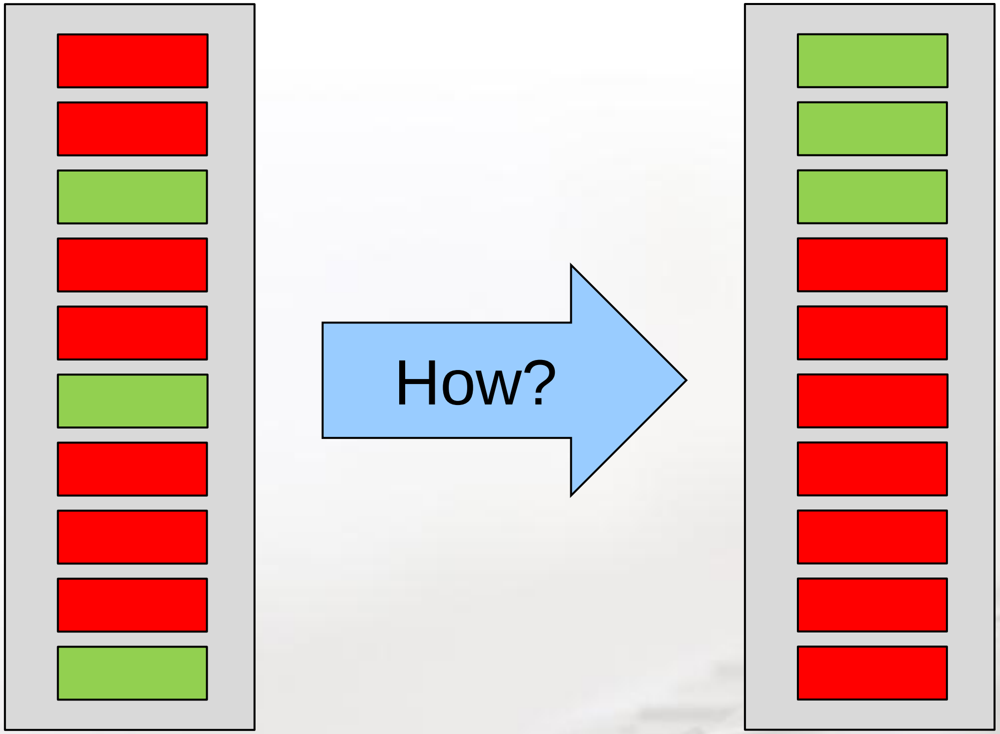
Search Effectiveness of the State-of-the-Art

State-of-the-Art (SoA) Effectiveness



How to improve?

Goal: Maximize Relevance



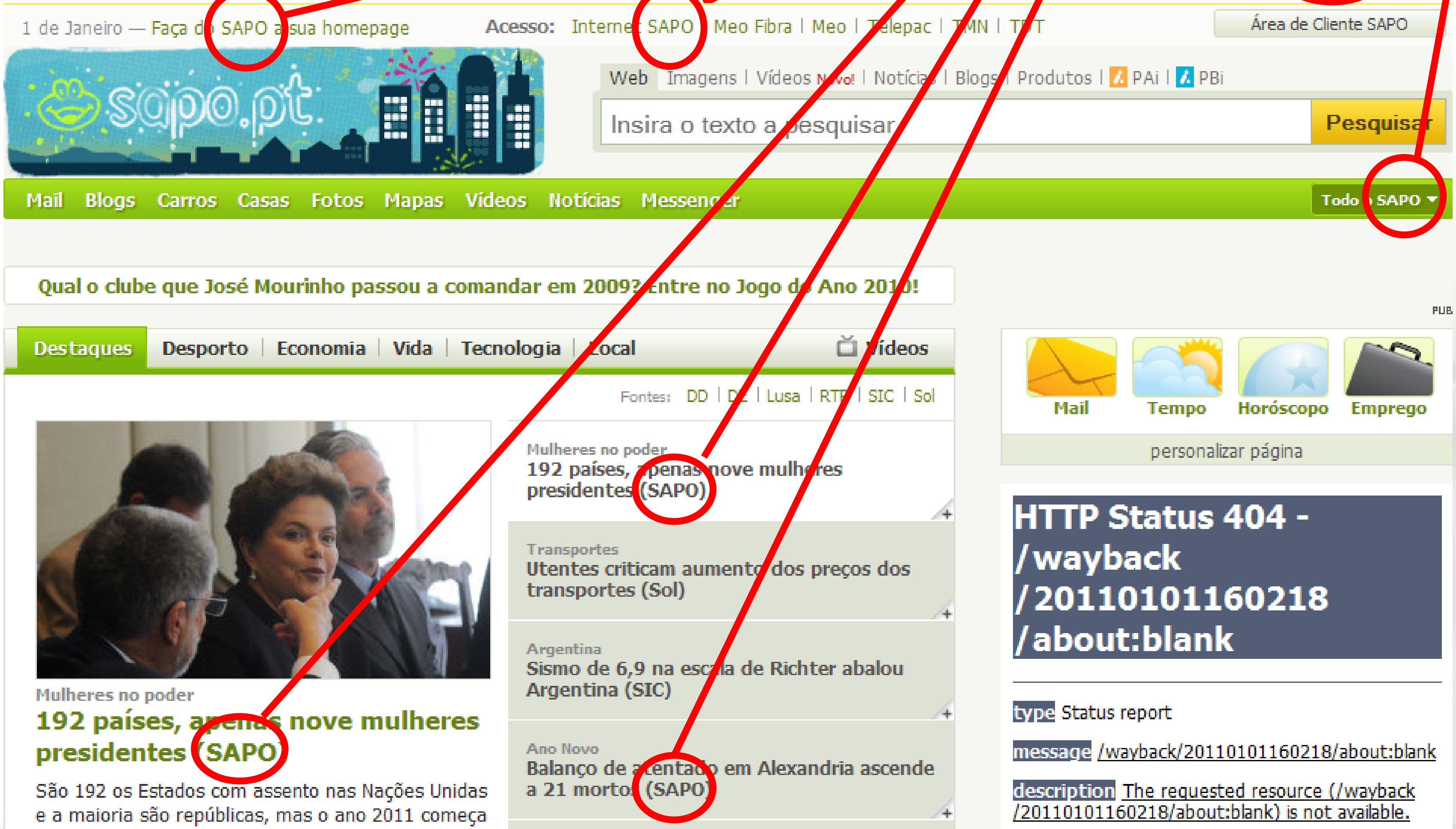
🟩 Relevant document

🟥 Non-relevant document

State-of-the-Art: Ranking Features

Query: sapo (toad in English)

http://www.sapo.pt



1 de Janeiro — Faça do SAPO a sua homepage

Acesso: Internet SAPO | Meo Fibra | Meo | Telepac | TMN | TNT

Área de Cliente SAPO

Web | Imagens | Vídeos Novo! | Notícias | Blogs | Produtos | PAi | PBi

Insira o texto a pesquisar

Pesquisar

Mail Blogs Carros Casas Fotos Mapas Vídeos Notícias Messenger

Todo o SAPO

Qual o clube que José Mourinho passou a comandar em 2009? Entre no Jogo do Ano 2010!

PUB

Destaques | Desporto | Economia | Vida | Tecnologia | Local | Vídeos

Fontes: DD | DE | Lusa | RTP | SIC | Sol

Mulheres no poder

192 países, apenas nove mulheres presidentes (SAPO)

São 192 os Estados com assento nas Nações Unidas e a maioria são repúblicas, mas o ano 2011 começa

Mulheres no poder

192 países, apenas nove mulheres presidentes (SAPO)

Transportes

Utentes criticam aumento dos preços dos transportes (Sol)

Argentina

Sismo de 6,9 na escala de Richter abalou Argentina (SIC)

Ano Novo

Balanço de atentado em Alexandria ascende a 21 mortos (SAPO)

Mail | Tempo | Horóscopo | Emprego

personalizar página

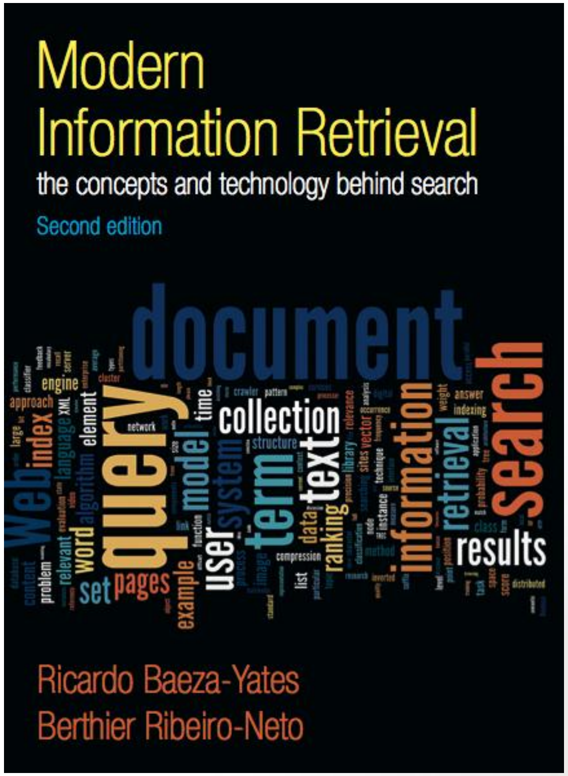
HTTP Status 404 - /wayback /20110101160218 /about:blank

type Status report

message /wayback/20110101160218/about:blank

description The requested resource (/wayback /20110101160218/about:blank) is not available.

What other Features can we use?



Modern Information Retrieval

the concepts and technology behind search

Second edition

Ricardo Baeza-Yates

Berthier Ribeiro-Neto

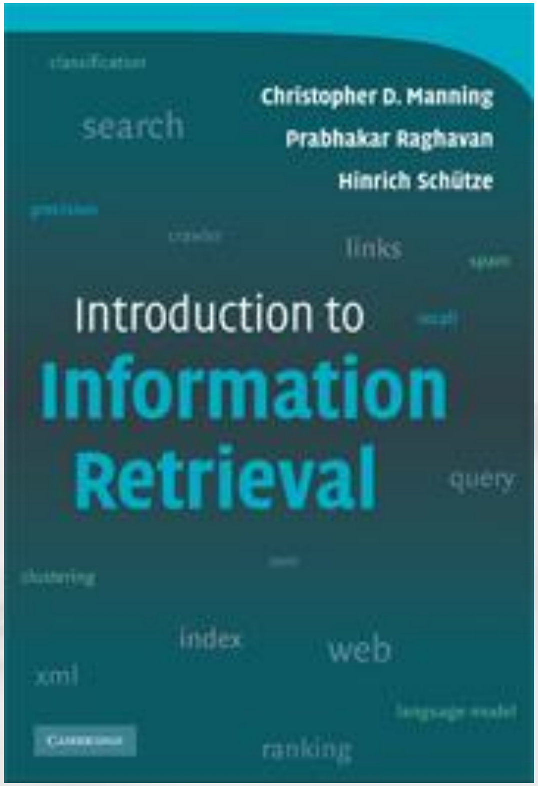
International Edition

Search Engines

INFORMATION RETRIEVAL IN PRACTICE

W. Bruce Croft • Donald Metzler • Trevor Strohman

PEARSON



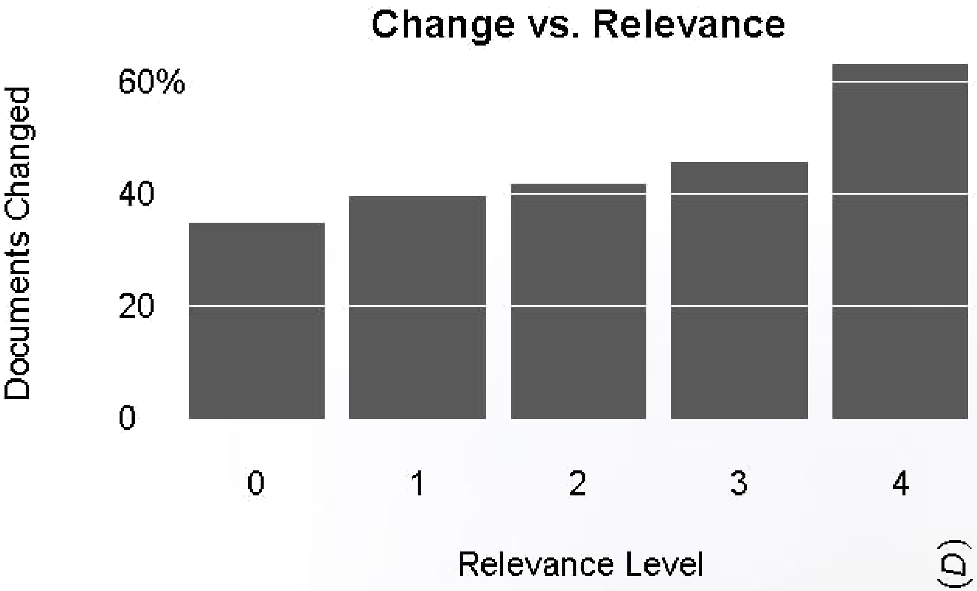
Christopher D. Manning

Prabhakar Raghavan

Hinrich Schütze

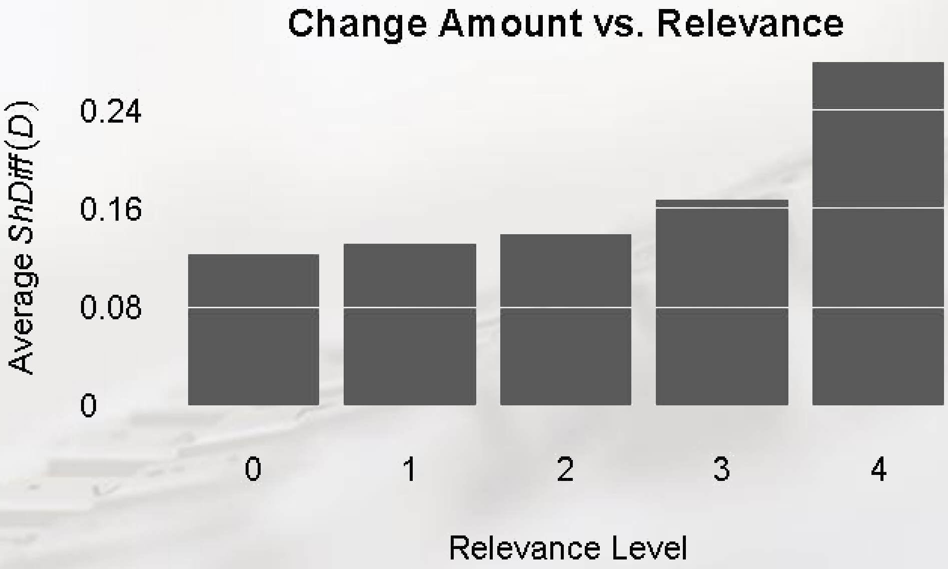
Introduction to Information Retrieval

Document Change Over Time



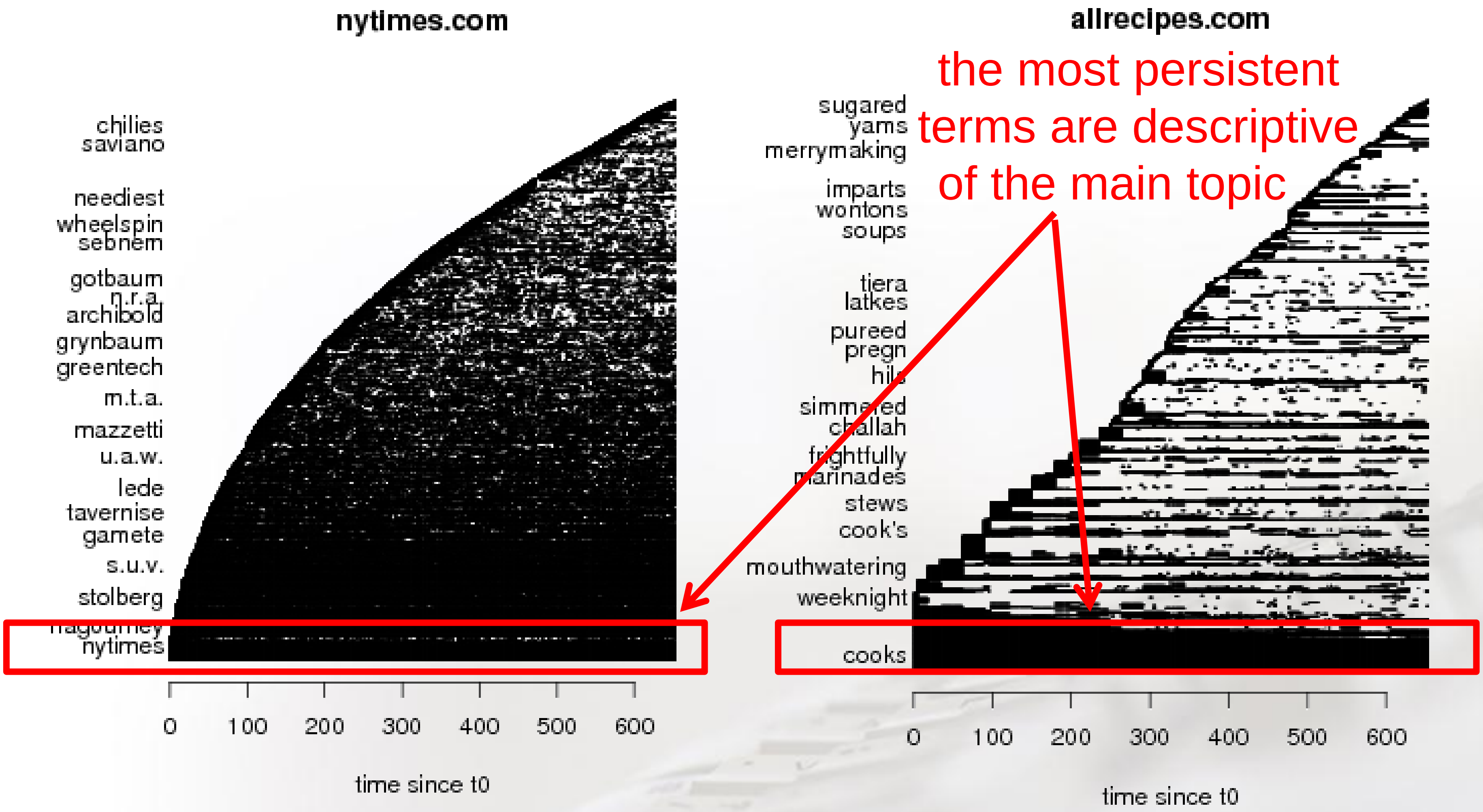
documents with higher relevance are more likely to change

documents with higher relevance tend to change to a greater degree



Leveraging Temporal Dynamics of Document Content in Relevance Ranking, *Web Search and Data Mining* 2010.

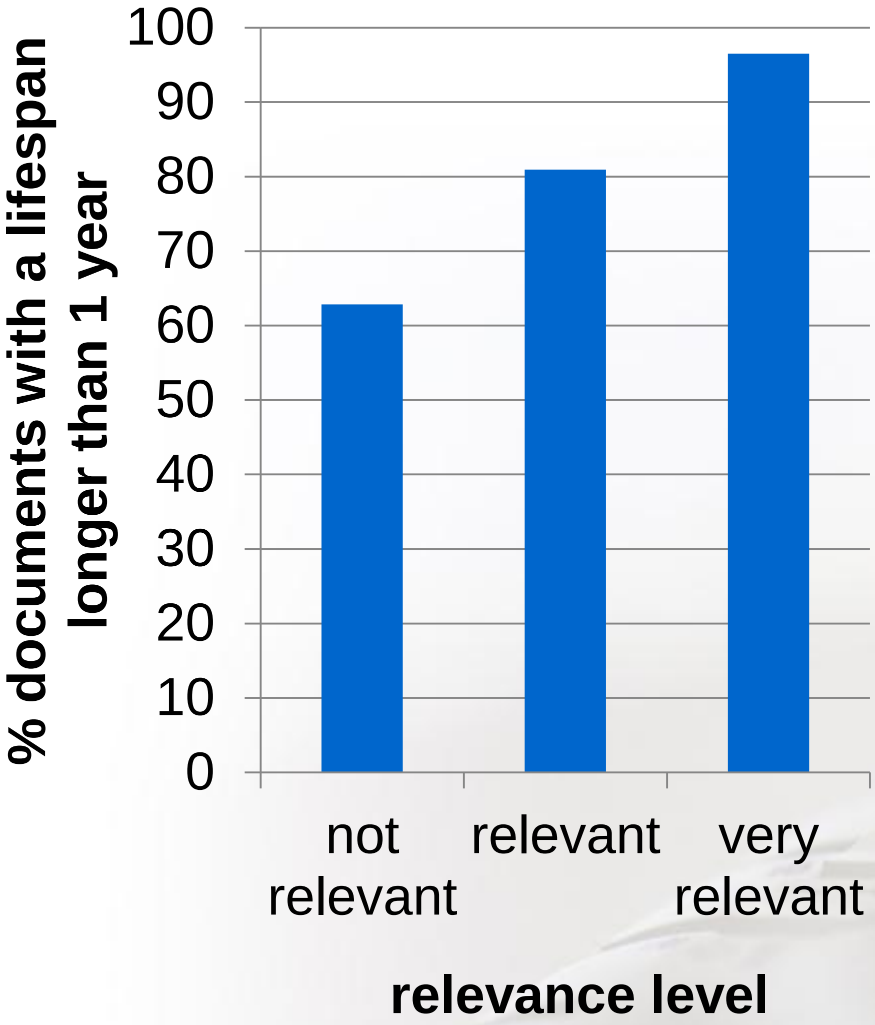
Vocabulary Change Over Time



the most persistent terms are descriptive of the main topic

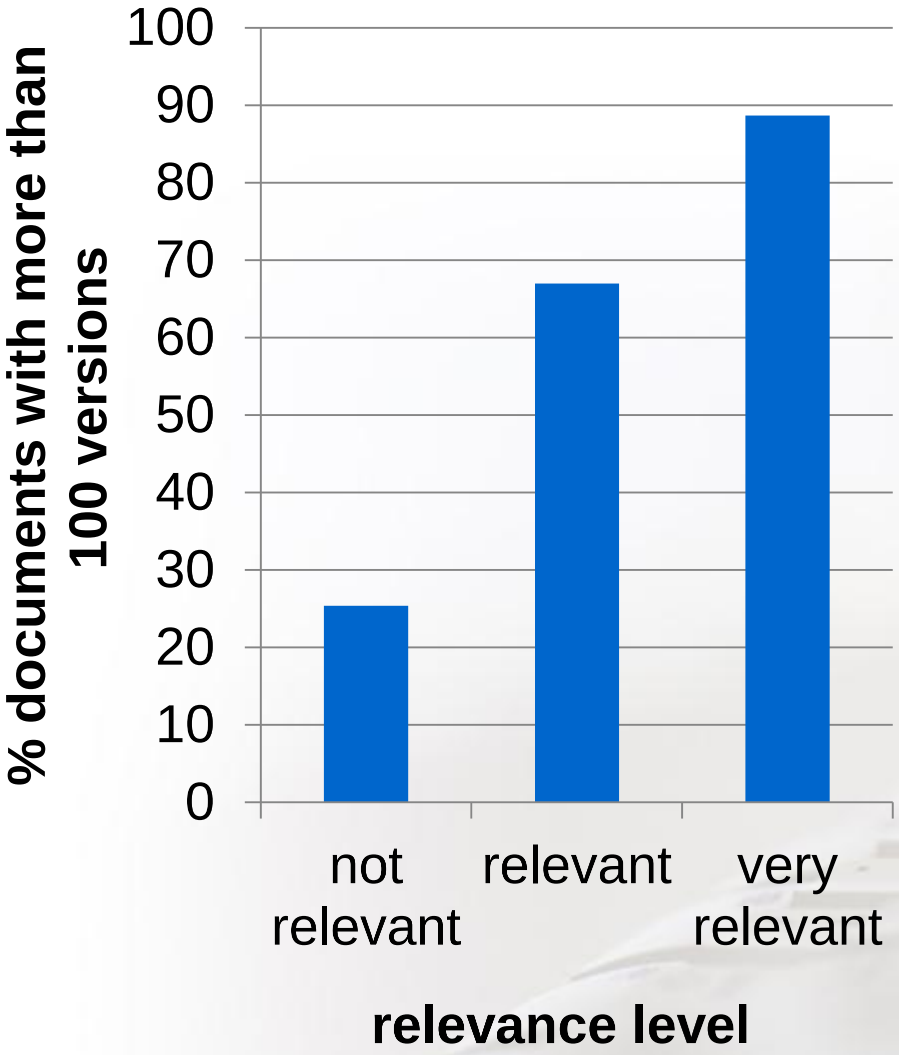
The Web Changes Everything: Understanding the Dynamics of Web Content, *Web Search and Data Mining* 2009.

Lifespan vs Relevance



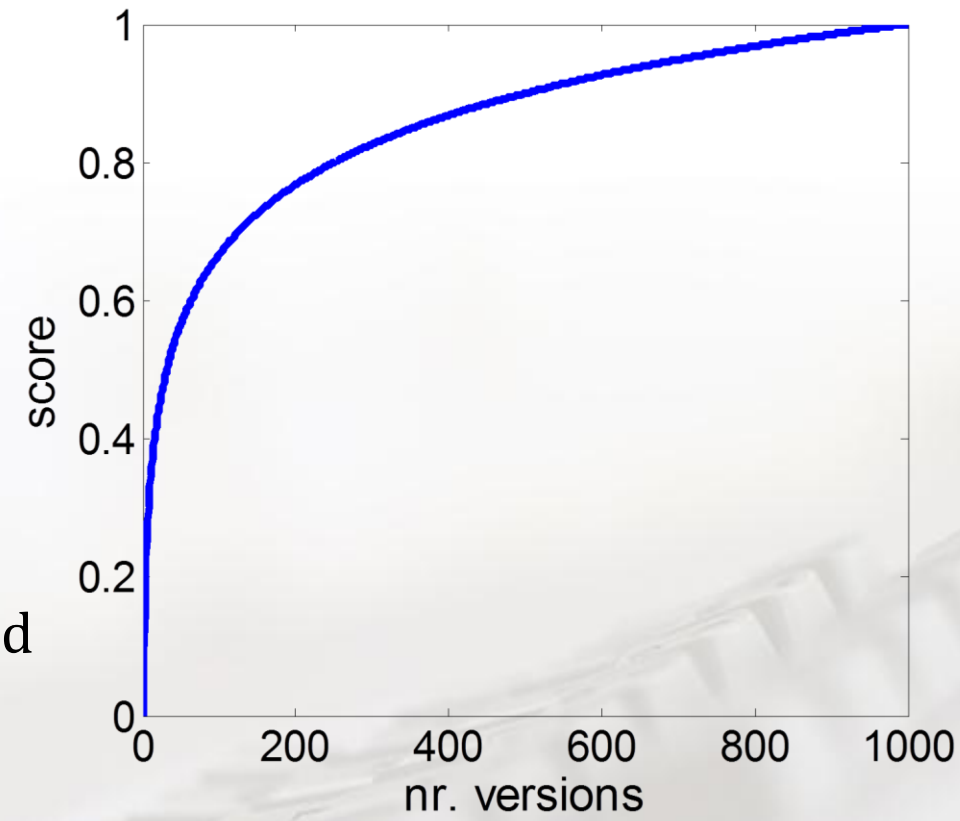
documents with higher relevance tend to have a longer lifespan

# Versions vs Relevance



documents with higher relevance tend to have more versions

Modeling Temporal Information



$$f_{Versions}(d) = log_{y}(x)$$

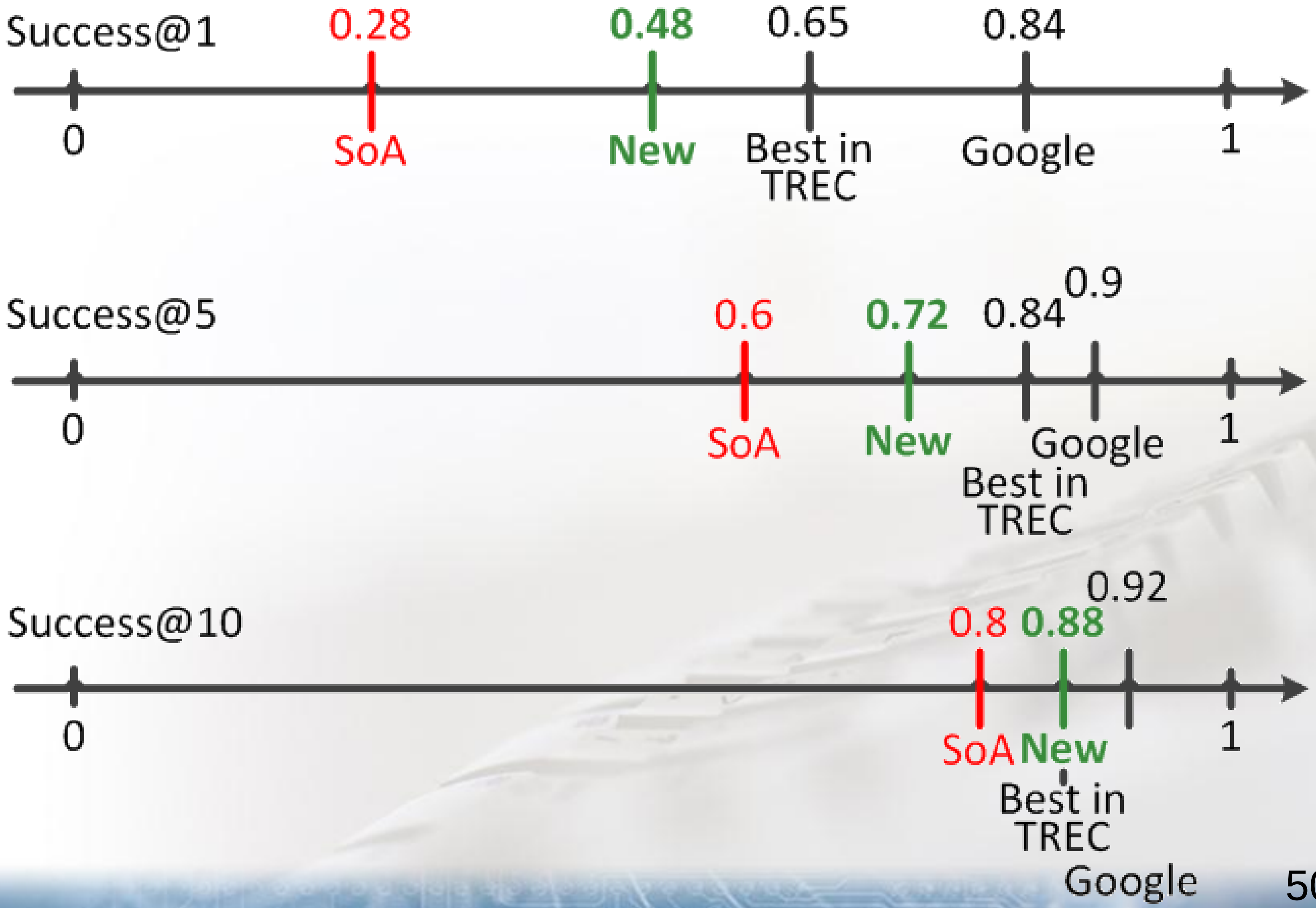
Parameters:
x = number of versions of document d
y = maximum number of versions of
a document in the collection

Assumption: persistent documents are more relevant

Results

Search Effectiveness of the State-of-the-Art + #Versions

New Ranking Model: $f_{SoA} + f_{Versions}$



If one is good, more is better

68 Ranking Features

Term-weighting:
BM25
Lucene
NutchWAX
...

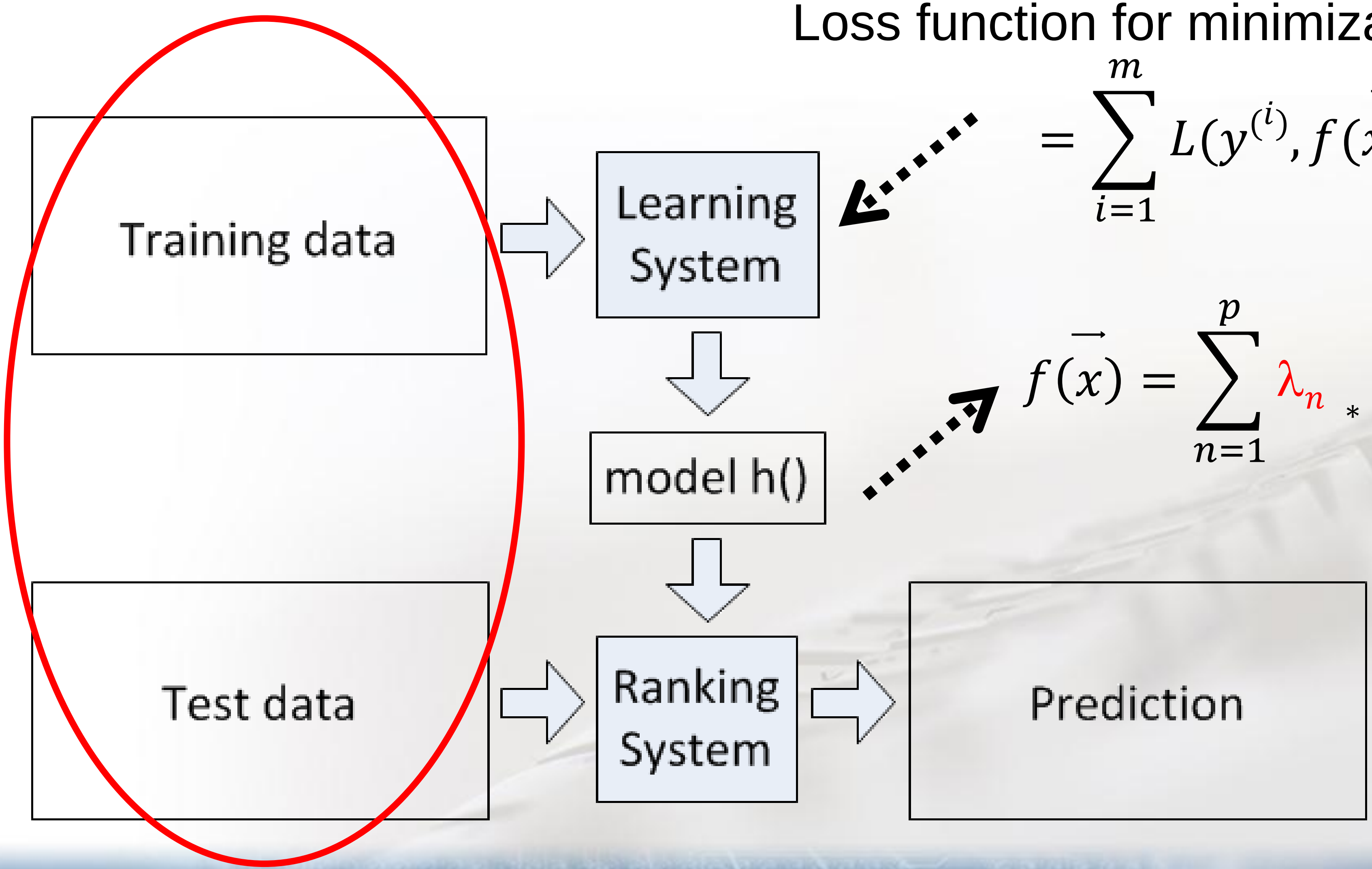
Term-distance:
MinPair
MinSpanOrdered
MinSpanUnordered
...

Web-graph based:
Inlinks
Outlinks
PageRank
...

Temporal:
NumberVersions
BoostOlder
Age
...

URL based:
UrlLength
UrlDepth
UrlSlashes
...

Supervised *Learning-to-Rank* Framework



Loss function for minimization

$$= \sum_{i=1}^{m} L(y^{(i)}, f(\vec{x}^{(i)}))$$

$$f(\vec{x}) = \sum_{n=1}^{p} \lambda_n * f_n(\vec{x})$$

Dataset for L2R in Web Archives

- 39 608 quadruples <query, version, relevance grade, features>
  - 50 **queries** (navigational topics)
  - 843 **versions** assessed on average per query
  - **3-level scale** of relevance
  - 68 ranking **features** extracted

- File Format:

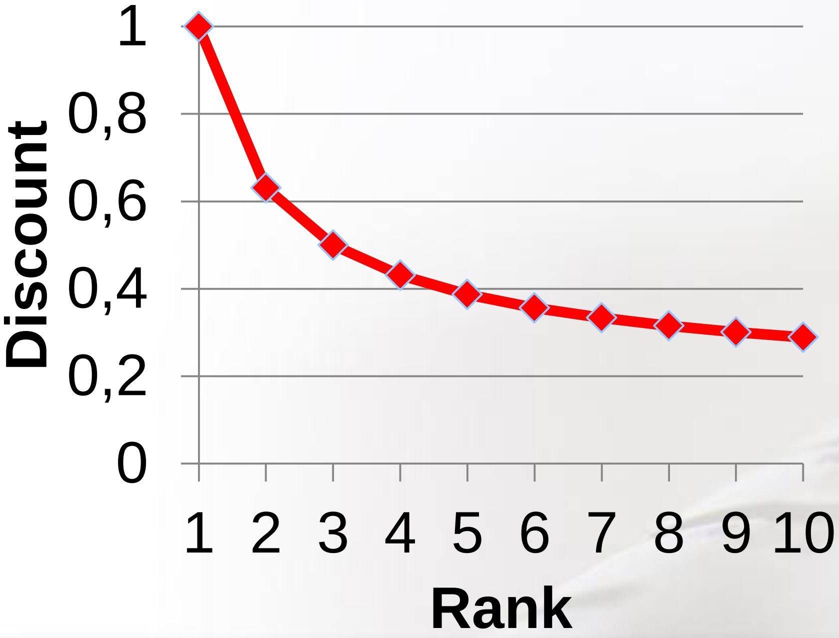
used in training

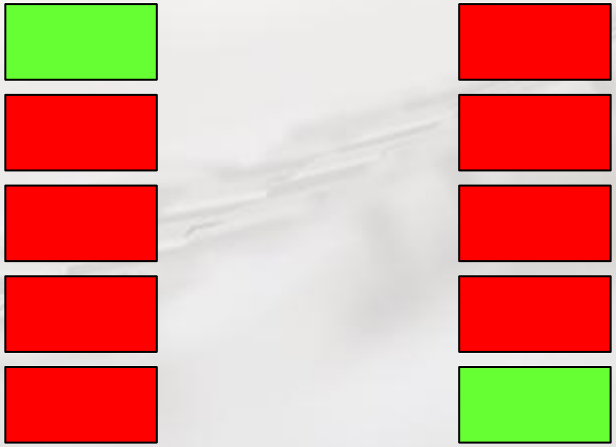
Rel.	Query	Features	Doc. Version
2	qid:21	1:0.70 2:0.34 3:0.27 ... 68:0.86	# id114746079
0	qid:22	1:0.05 2:0.18 3:0.14 ... 68:0.43	# id172346033
1	qid:22	1:0.75 2:0.33 3:0.84 ... 68:0.54	# id456334535

Evaluation Metrics: NDCG@k

- Normalized Discounted Cumulative Gain at cut-off k
  - total gain accumulated at a particular rank *p*

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i}-1}{log(1+i)}$$





NDCG@5 = 1

NDCG@5 ≈ 0.4

Results

Search Effectiveness of the 68 features

Results of L2R Algorithms

	State-of-the-Art		L2R algorithms (68 features)			
Metric	Lucene	NutchWAX	RankBoost	AdaRank	ListNet	Random Forests
NDCG@1	0.220	0.250	0.530	0.400	0.450	**0.650**
NDCG@5	0.157	0.215	0.535	0.426	0.432	**0.665**
NDCG@10	0.133	0.174	0.570	0.476	0.464	**0.688**

4x higher

All results show a statistical significance of p<0.01 against NutchWAX

Results

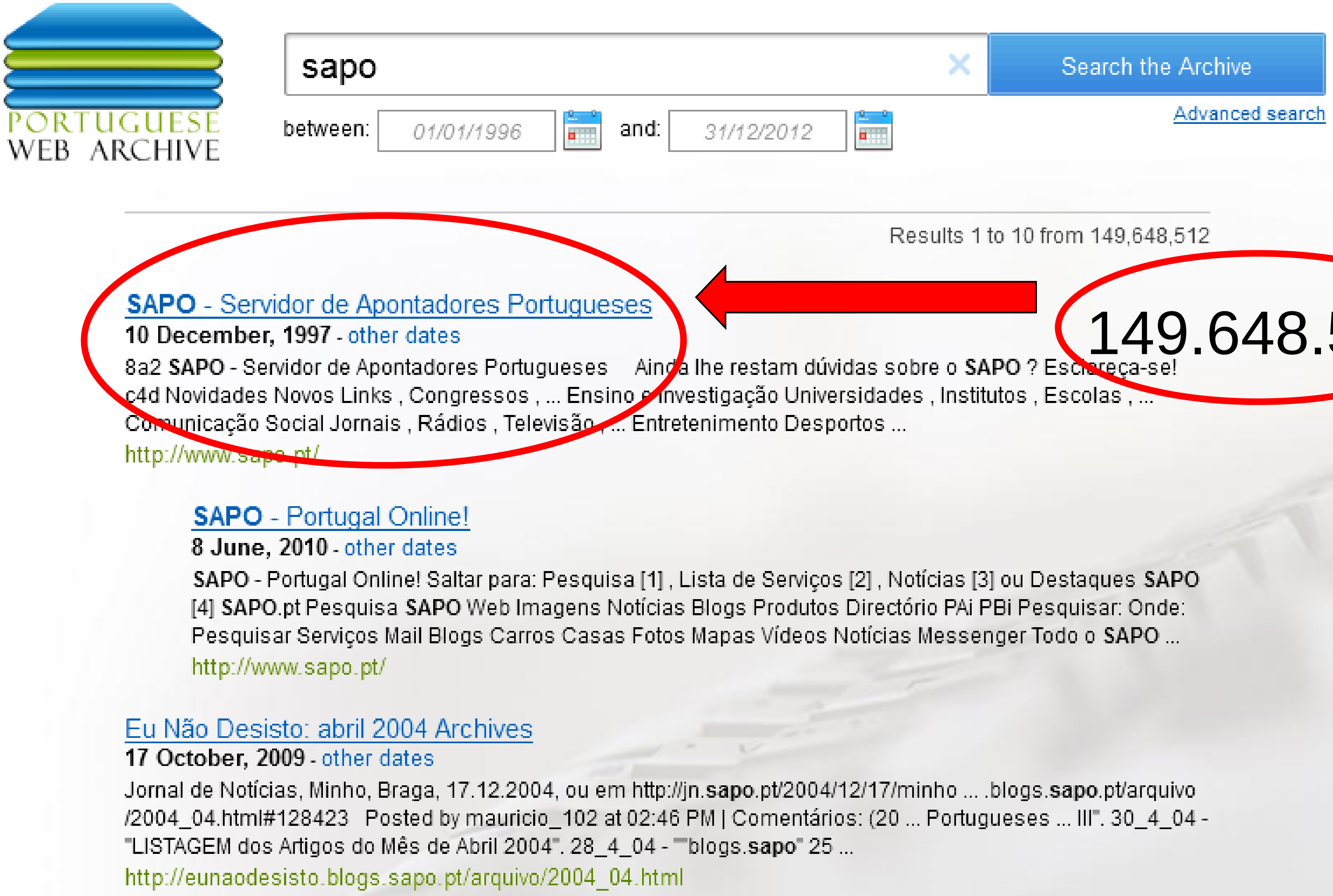
How much did the search effectiveness improve with the temporal features?

With Temporal Features is Better

- NDCG@1
  - RankBoost 0.53 > NT RankBoost 0.44 **+9%**
  - AdaRank 0.40 > NT AdaRank 0.38 **+2%**
  - ListNet 0.45 > NT ListNet 0.37 **+8%**
  - R. Forests 0.65 > NT R. Forests 0.55 **+10%**

- NDCG@10
  - RankBoost 0.57 > NT RankBoost 0.51 **+6%**
  - AdaRank 0.48 > NT AdaRank 0.47 **+1%**
  - ListNet 0.46 > NT ListNet 0.43 **+3%**
  - R. Forests 0.69 > NT R. Forests 0.65 **+4%**

Better Results = Happier Users



Results 1 to 10 from 149,648,512

SAPO - Servidor de Apontadores Portugueses
10 December, 1997 - other dates
8a2 **SAPO** - Servidor de Apontadores Portugueses Ainda lhe restam dúvidas sobre o **SAPO** ? Esclareça-se! c4d Novidades Novos Links , Congressos , ... Ensino e Investigação Universidades , Institutos , Escolas , ... Comunicação Social Jornais , Rádios , Televisão , ... Entretenimento Desportos ...
http://www.sapo.pt/

SAPO - Portugal Online!
8 June, 2010 - other dates
SAPO - Portugal Online! Saltar para: Pesquisa [1] , Lista de Serviços [2] , Notícias [3] ou Destaques **SAPO** [4] **SAPO**.pt Pesquisa **SAPO** Web Imagens Notícias Blogs Produtos Directório PAi PBi Pesquisar: Onde: Pesquisar Serviços Mail Blogs Carros Casas Fotos Mapas Vídeos Notícias Messenger Todo o **SAPO** ...
http://www.sapo.pt/

Eu Não Desisto: abril 2004 Archives
17 October, 2009 - other dates
Jornal de Notícias, Minho, Braga, 17.12.2004, ou em http://jn.**sapo**.pt/2004/12/17/minho ... .blogs.**sapo**.pt/arquivo /2004_04.html#128423 Posted by mauricio_102 at 02:46 PM | Comentários: (20 ... Portugueses ... III". 30_4_04 - "LISTAGEM dos Artigos do Mês de Abril 2004". 28_4_04 - ""blogs.**sapo**" 25 ...
http://eunaodesisto.blogs.sapo.pt/arquivo/2004_04.html

149.648.512

Conclusions

Conclusions

- Users need analytical tools for specific users + search tools for generic users.

- State-of-the-Art searching technology provides poor results.

- Temporal information intrinsic to web archives improves their search results.

- Learning-to-Rank technology greatly improves search results.

Resources

- Public service since 2010.
  - http://archive.pt
- Test collection to support evaluation.
  - https://code.google.com/p/pwa-technologies/wiki/TestCollection
- L2R dataset for web archive IR research.
  - http://code.google.com/p/pwa-technologies/wiki/L2R4WAIR
- All code available under the LGPL license.
  - https://code.google.com/p/pwa-technologies/

Thank you.

http://archive.pt

miguel.costa@fccn.pt