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
Our Memory is in Digital Form

E-books

Web photo galleries

Forums

Blogs

Online newspapers

Social networks
The Web is Ephemereral

• 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 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)
List of Web archiving initiatives

From Wikipedia, the free encyclopedia

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.

Contents
1 Web archiving initiatives
2 Archived data
3 Access methods
4 References

Web archiving initiatives

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Creation Year</th>
<th>Technologies</th>
<th>Number of Employees</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia's Web Archive[1]</td>
<td>Australia</td>
<td>1996</td>
<td>PANDORA Digital Archiving System (PANDAS), NLA Trove, HTTrack.</td>
<td>4</td>
<td>Full-time: 4, Part-time: &gt;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.</td>
</tr>
<tr>
<td>Our digital island, a Tasmanian Web Archive[2]</td>
<td>Australia</td>
<td>1996</td>
<td>HTTrack, Experimentally: Web Curator, Hentix and Wayback Machine</td>
<td>1</td>
<td>Archival access tools and...</td>
</tr>
</tbody>
</table>
2013: Worldwide Web Archiving Initiatives

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

2008

Welcome to the Tomba project.

Publishing tools, such as Blogger, enable web publishers. Never before in history have so many web archives been available for the public to browse.

If we do not archive the current web content, it will be lost.

The Internet Archive collects and stores a comprehensive archive of the web for posterity, while satisfying the needs of today's web users.

As a result, several countries are creating national web archives to preserve their national history.

Portugal is now beginning its national web archive for the Web (Portuguese Web Archive).

See or rediscover pages of the past.

There are more than 2.7 million pages from 2008.

Search examples:

- Expo 98
- Euro 2004
- Elections in 2004
- University of Lisbon

US elections

between 01/01/1996 and 01/12/2009

US Elections
6 January, 2007 - other dates

US Elections The Embassy Consular Services Portugal and What direction for America if the results become clear? many people are still speculating on http://www.american-embassy.pt/DepOLU.S.Embassy Lisbon, Portugal - Government & Other Links
18 April, 2004 - other dates


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 x POLITIK x 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/europaeisch_us_wahl
1.8 Billion Archived Files (52 TB)
What can we do with all this data?
### Portuguese WA: Web Characterization

<table>
<thead>
<tr>
<th>Media type</th>
<th>% contents 2005</th>
<th>% contents 2008</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text/html</td>
<td>61.2%</td>
<td>57.8%</td>
<td>-5.5%</td>
</tr>
<tr>
<td>Image/jpeg</td>
<td>22.6%</td>
<td>22.8%</td>
<td>+1.2%</td>
</tr>
<tr>
<td>Image/gif</td>
<td>11.4%</td>
<td>9.4%</td>
<td>-17.4%</td>
</tr>
<tr>
<td>Text/pdf</td>
<td>1.6%</td>
<td>1.9%</td>
<td>+18.5%</td>
</tr>
<tr>
<td>Other</td>
<td>3.2%</td>
<td>8.1%</td>
<td>-</td>
</tr>
</tbody>
</table>

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

Meme-tracking and the Dynamics of the News Cycle, *Knowledge Discovery and Data Mining* 2009.
Paul Butler created this friend relationship visualization map using Facebook data.
Twitómetro: Sentiment Analysis

http://dmir.inesc-id.pt/project/Reaction
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!
• Available since 2010: http://archive.pt
• 1.2 billion documents
  – searchable by full-text and URL
  – range between 1996 and 2012
### 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.
Qual o clube que José Mourinho passou a comandar em 2009? Entre no Jogo do Ano 2010!

1 de Janeiro — Faça do SAPO a sua homepage

Web | Imagens | Vídeos Novo! | Notícias | Blogs | Produtos | APAi | APBi

Pesquisar

Mail | Blogs | Carros | Casas | Fotos | Mapas | Vídeos | Notícias | Messenger

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

Destaques | Desporto | Economia | Vida | Tecnologia | Local | Videos

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

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

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

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.
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, ...
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
http://www.sapo.pt/

Eu Não Desisto: abril 2004 Archives
17 October, 2009 - other dates
Understanding the Users’ Information Needs

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?

New Technology
Evaluation
Is it better than State-of-the-Art?
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

click frequency (normalized)

clicks on 1st results page

ranks of results clicked
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)

Evaluation Metric: Success@k

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

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

<table>
<thead>
<tr>
<th>Query</th>
<th>Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>
Search Effectiveness of the State-of-the-Art
State-of-the-Art (SoA) Effectiveness

Success@1

0

SoA

0.28

0.65

Best in TREC

0.84

Google

Success@5

0

SoA

0.6

0.84

Google

0.9

Best in TREC

Success@10

0

SoA

0.8

0.88

Google

0.92

Best in TREC
How to improve?
Goal: Maximize Relevance

How?
State-of-the-Art: Ranking Features

Query: sapo (toad in English)
What other Features can we use?
Vocabulary Change Over Time

The most persistent terms are descriptive of the main topic.

documents with higher relevance tend to have a longer lifespan
documents with higher relevance tend to have more versions
Modeling Temporal Information

\[ f_{\text{versions}}(d) = \log_y(x) \]

Parameters:
- \( x = \text{number of versions of document } d \)
- \( y = \text{maximum number of versions of a document in the collection} \)

Assumption: persistent documents are more relevant
Search Effectiveness of the State-of-the-Art + #Versions
New Ranking Model: $f_{\text{SoA}} + f_{\text{Versions}}$
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
- ...

**URL based:**
- UrlLength
- UrlDepth
- UrlSlashes
- ...

**Temporal:**
- NumberVersions
- BoostOlder
- Age
- ...

[Image of a mixer with various features listed next to it]
Supervised Learning-to-Rank Framework

Loss function for minimization

\[
\bar{f}(x) = \sum_{n=1}^{p} \lambda_n \ast f_n(x)
\]

\[
= \sum_{i=1}^{m} L(y^{(i)}, f(x^{(i)}))
\]
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:

<table>
<thead>
<tr>
<th>Rel.</th>
<th>Query</th>
<th>Features</th>
<th>Doc. Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>qid:21</td>
<td>1:0.70 2:0.34 3:0.27 ... 68:0.86</td>
<td># id114746079</td>
</tr>
<tr>
<td>0</td>
<td>qid:22</td>
<td>1:0.05 2:0.18 3:0.14 ... 68:0.43</td>
<td># id172346033</td>
</tr>
<tr>
<td>1</td>
<td>qid:22</td>
<td>1:0.75 2:0.33 3:0.84 ... 68:0.54</td>
<td># id456334535</td>
</tr>
</tbody>
</table>

used in training
Evaluation Metrics: NDCG@k

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

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

$NDCG@5 = 1$  $NDCG@5 \approx 0.4$
Search Effectiveness of the 68 features
## Results of L2R Algorithms

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

• 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