LASIGE
Large-Scale Informatics Systems Laboratory

U LISBOA | UNIVERSIDADE DE LISBOA

Information Search in Web Archives

Miguel Costa

Advisor: Prof. Mário J. Silva
Co-Advisor: Prof. Francisco Couto

Department of Informatics, Faculty of Sciences, University of Lisbon

PhD thesis defense, Lisbon, Portugal
November 4, 2014

The Web is Ephemeral

- 50 days - 50% of documents are changed

  (Cho and Garcia-Molina. 2000)

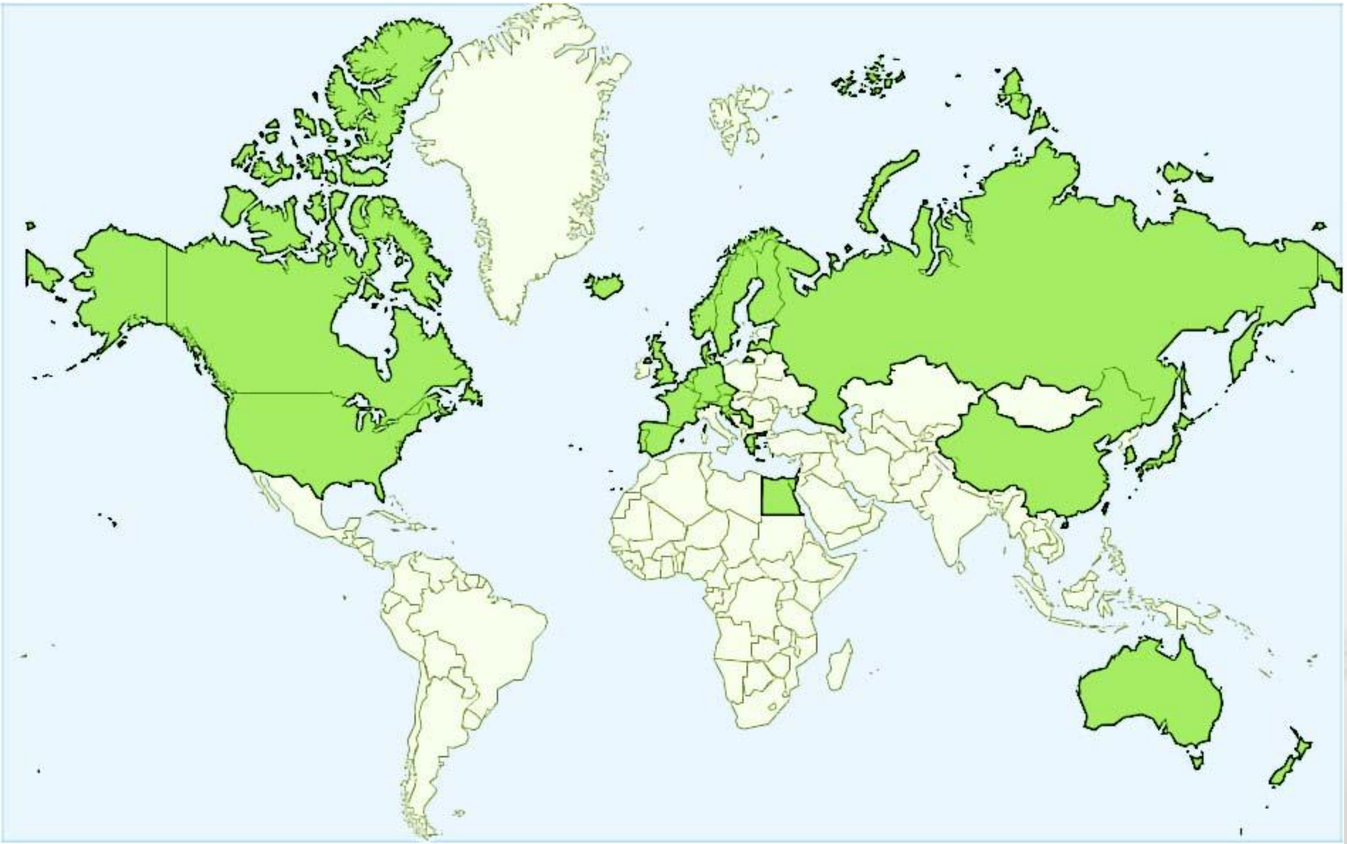
- 1 year - 80% of documents become inaccessible

  (Ntoulas, Cho and Olson. 2004)

- 27 months - 13% of web references disappear

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

2014: Web Archiving Initiatives



- +68 initiatives in 33 countries
- +534 billions of web contents since 1996 (17 PB)

- Available since 2010: http://archive.pt
- 1.2 billion documents

Objective of PhD Thesis

Problem:

- it is hard to find past information with current Web Archive Information Retrieval (WAIR) systems

Objective:

- study the problems of WAIR and propose solutions

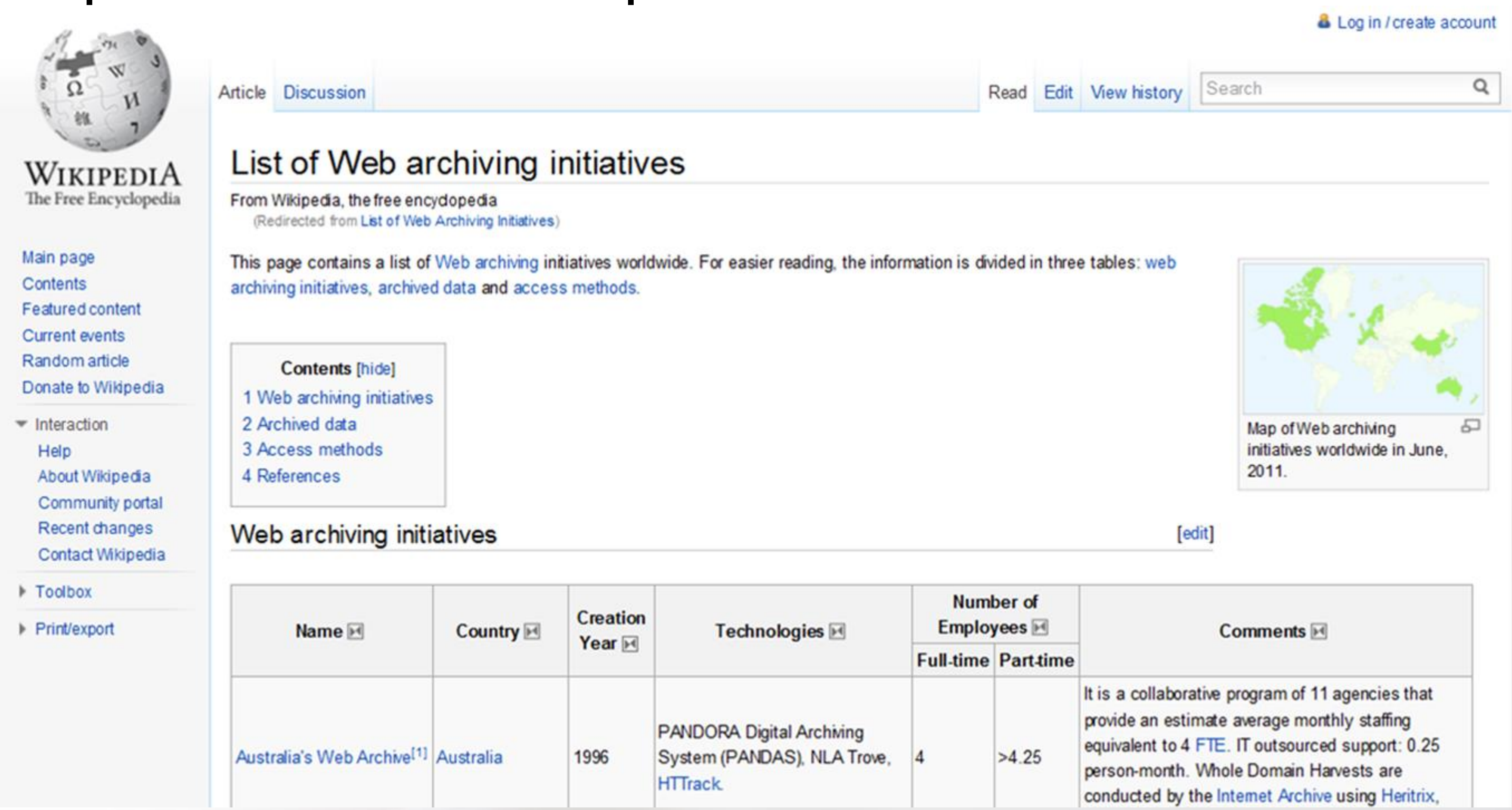
Contributions

1. Understanding WAIR systems
   - What is the state-of-the-art in WAIR?
   - What is the status of web archiving initiatives?
   - How are web archiving initiatives evolving?
2. Understanding web archive users
   - Does the state-of-the-art in WAIR meet the users' information needs?
   - Why, what and how do web archive users search?
   - What functionalities would like the users to see implemented?
   - What are the specificities of web archive users?
3. Improving WAIR systems
   - How to improve WAIR?
   - How to evaluate WAIR systems?
   - What is the search effectiveness of the state-of-the-art in WAIR?

Understanding WAIR Systems

Methodology: 2 Surveys

- conducted in 2010 and 2014.
- questionnaires and public information.



Log in / create account

Article | Discussion | Read | Edit | View history | Search

WIKIPEDIA The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

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.

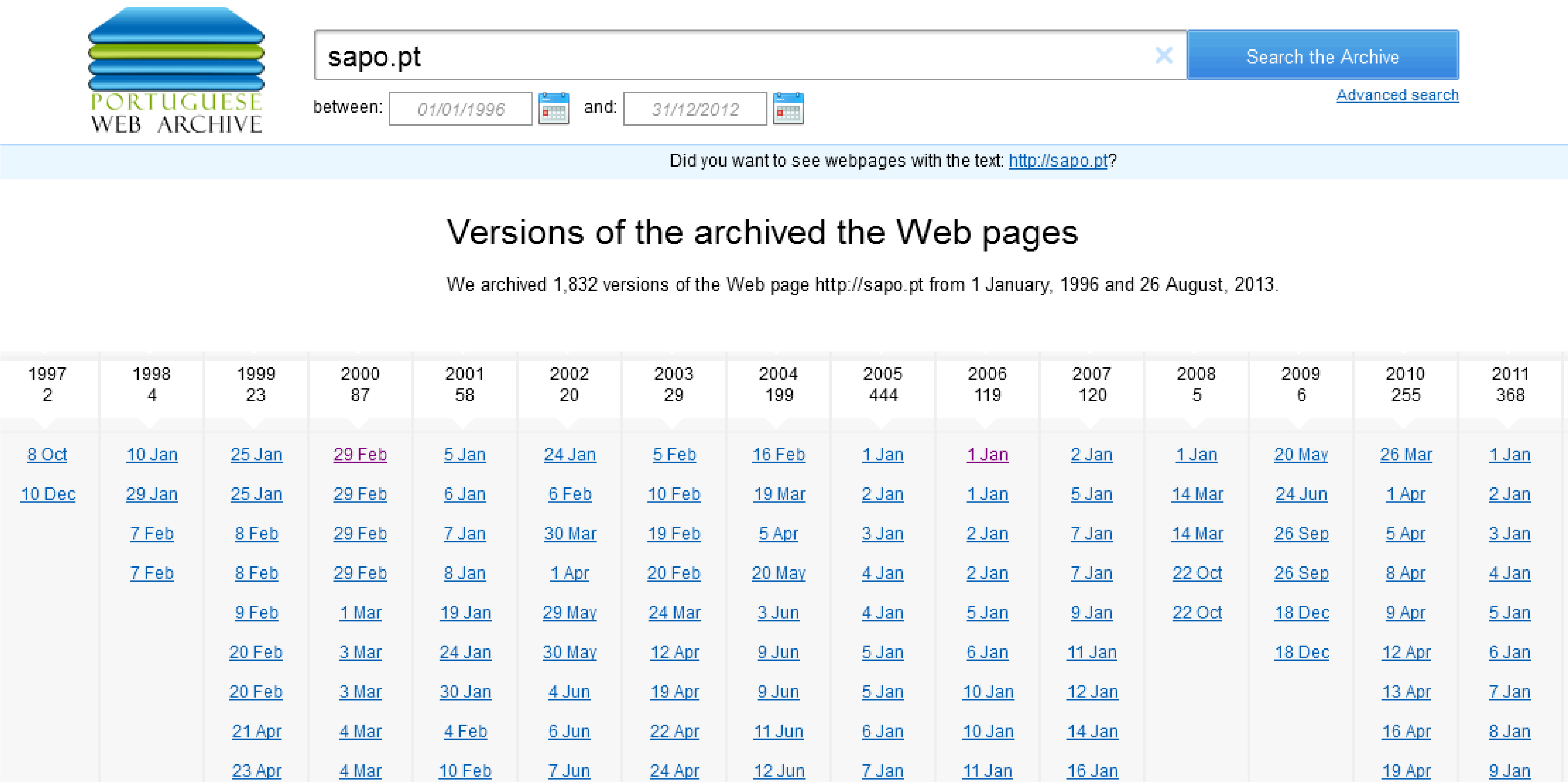
Contents [hide]

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,

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

What is the State-of-the-Art? 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?

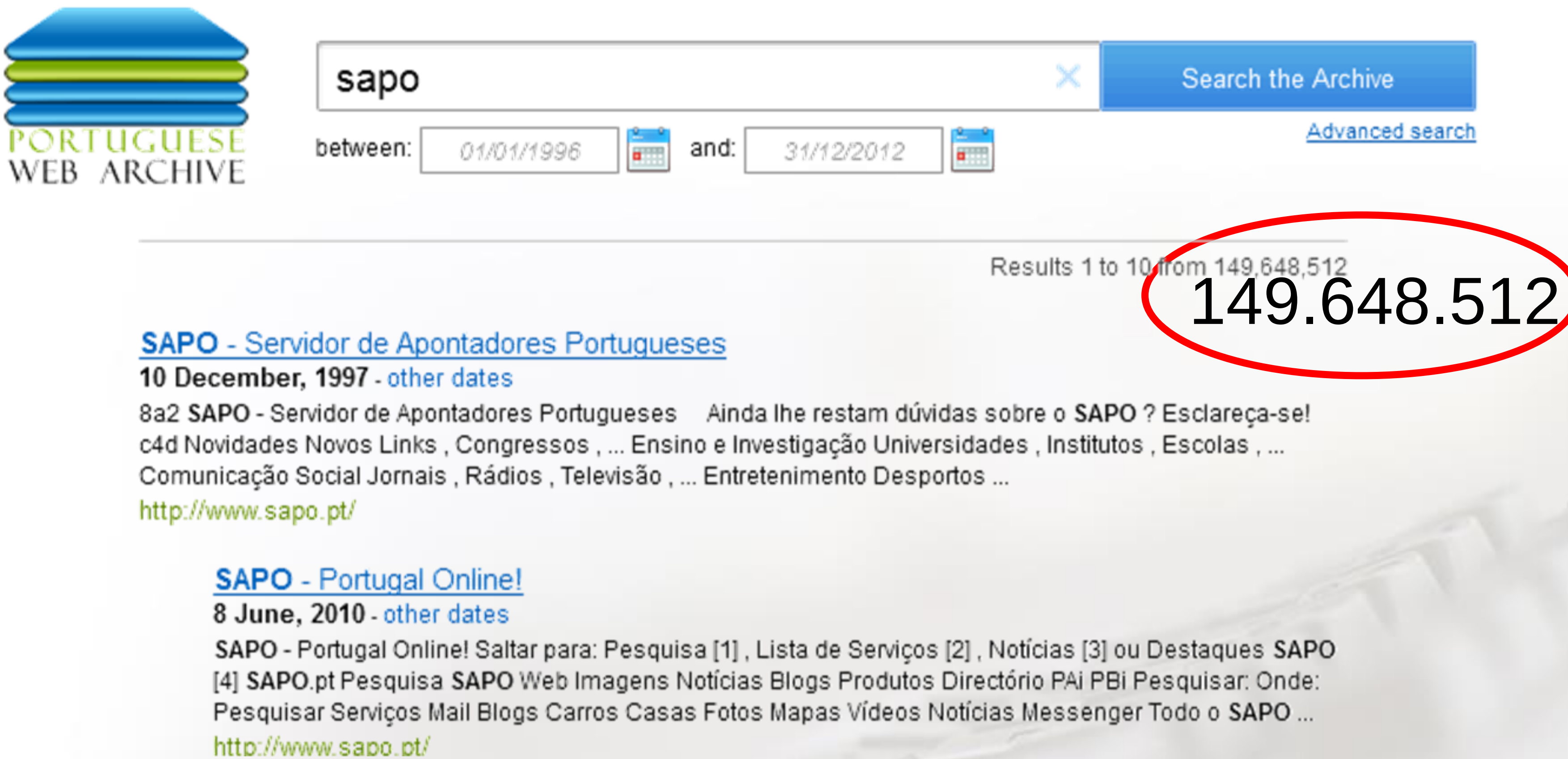
Versions of the archived the Web pages

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

- Technology based on the **Wayback Machine**.
- **Problem:** URLs are hard to remember or unknown.

What is the State-of-the-Art? Full-text Search



- Technology based on **Lucene** extensions (NutchWAX & Solr).
- **Problem:** poor relevance rankings.

Understanding Web Archive Users

Methodology: 3 Data Collecting Methods

Laboratory Studies

Online Questionnaires

Search Log Mining

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

data richeness

generalization

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**
- **Transactional** – 5% to 16%
  - downloading an old file or recovering a site from the **past**

Problems:

- Search engine technology optimized for different needs.
- Some needs are not supported by current technology.

Good news:

- Some needs may be supported by a high quality full-text search.

Improving WAIR

How to improve WAIR?

Previous studies show that temporal information:

- has been exploited to improve IR systems.
- can be extracted from web archives.

Hypothesis: state-of-the-art WAIR systems can be improved by exploiting temporal information intrinsic to web archives.

Exploiting Temporal Information

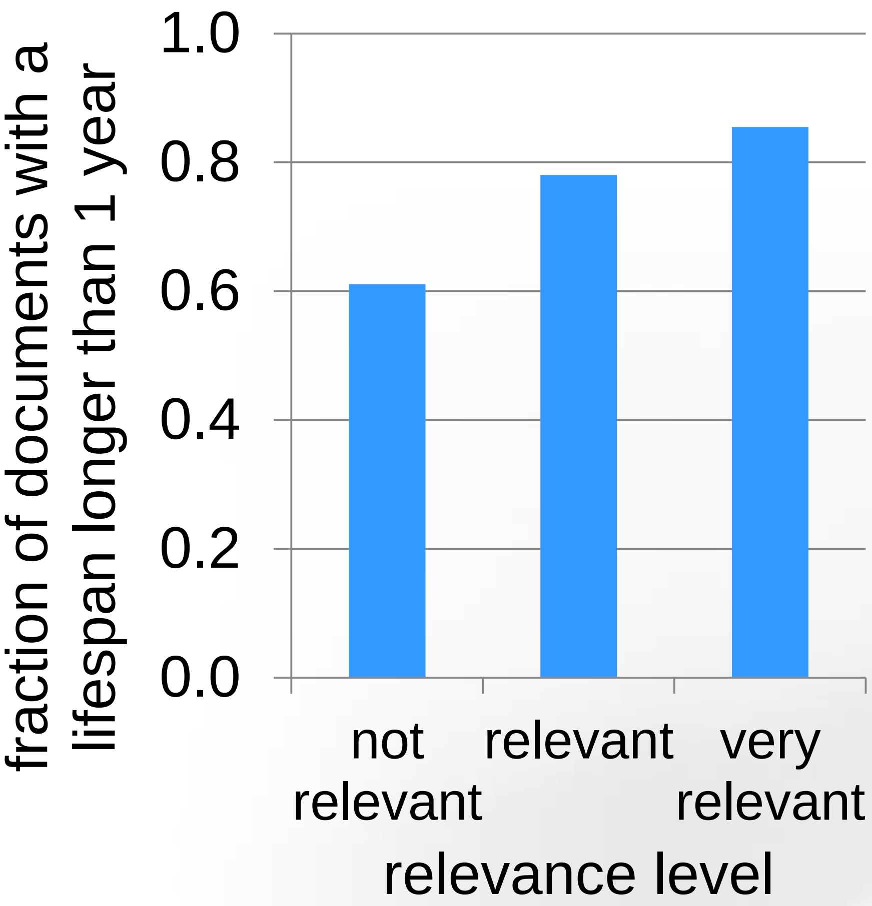
1. novel ranking features

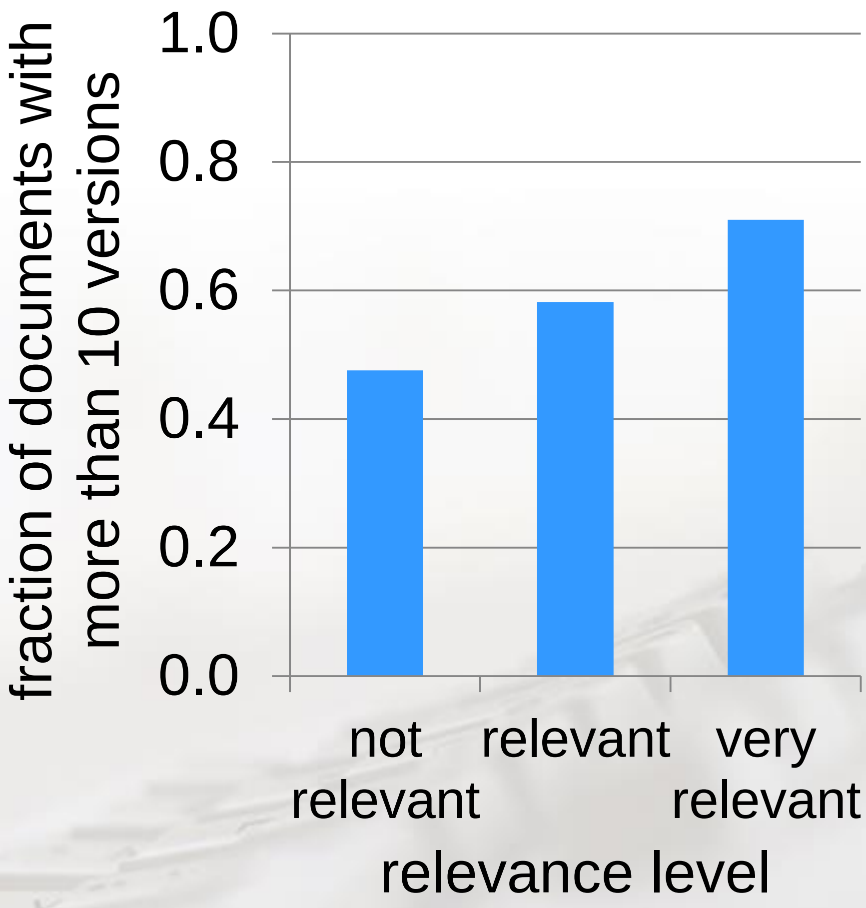
   Intuition: persistent documents are more relevant for navigational queries.

2. novel ranking framework

   Intuition: ensemble of models learned for specific periods are more effective than a single ranking model.

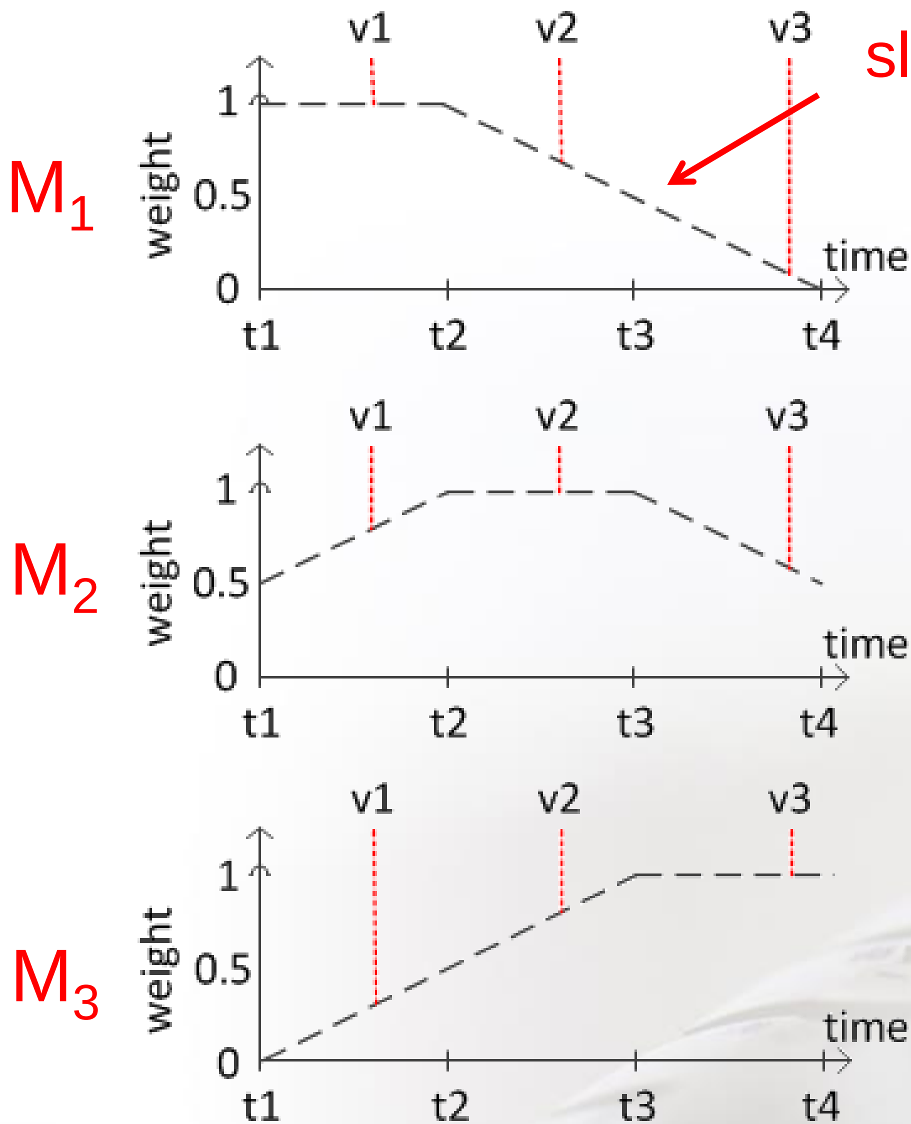
Temporal Ranking Features





documents with higher relevance tend to be more persistent (longer lifespan & more versions)

Temporal-Dependent Ranking Framework



- Learn a ranking model for each period.
- Use all data weighted by their temporal distance to the period.
- Combine models by minimizing a global loss function.

Temporal-Dependent Models

L= loss function

x_i = input of query-document feature vector

m = # instances

$$model = argmin_f \sum_{i=1}^{m} L(f(x_i, \omega), y_i)$$

ω = parameters

y_i = relevance label

Υ = temporal weight function

$$TD\,model = argmin_f \sum_{i=1}^{m} L(\boldsymbol{\Upsilon(x_i, Tk)}\, f(x_i, \omega), y_i)$$

$$\Upsilon(x_i, Tk) = \begin{cases} 1 & if\ xi \in Tk \\ 1 - \alpha \dfrac{distance(xi, Tk)}{|T|} & if\ xi \notin Tk \end{cases}$$

α = slope

Evaluation Methodology

Evaluation Methodology

- Test Collection (based on Cranfield Paradigm):
  - **Corpus**: 6 web collections, 255M contents, 8.9TB
  - **Topics**: 50 navigational (1/3 with date range)
  - **Relevance Judgments**: 3 judges, 3-level scale of relevance, 267 822 versions assessed
  - **Metrics**: (NDCG@k, P@k | k=1,5,10)

- Dataset for learning to rank (L2R):
  - 39 608 quadruples <query, version, grade, features>
  - 68 ranking features extracted (including temporal)
  - 5-fold cross-validation

Results &
Validation of Thesis

State-of-the-Art vs. Learning-to-Rank (L2R)

weak baseline

strong baseline

	State-of-the-Art		L2R algorithms (without temporal features)		
Metric	Lucene	NutchWAX	AdaRank	Rank SVM	Random Forests
NDCG@1	0.220	0.250	0.380	0.500	0.550
NDCG@5	0.157	0.215	0.427	0.485	0.610
NDCG@10	0.133	0.174	0.470	0.523	0.650

+ 30%

All results show a statistical significance of p<0.01 with a two-sided paired t-test.

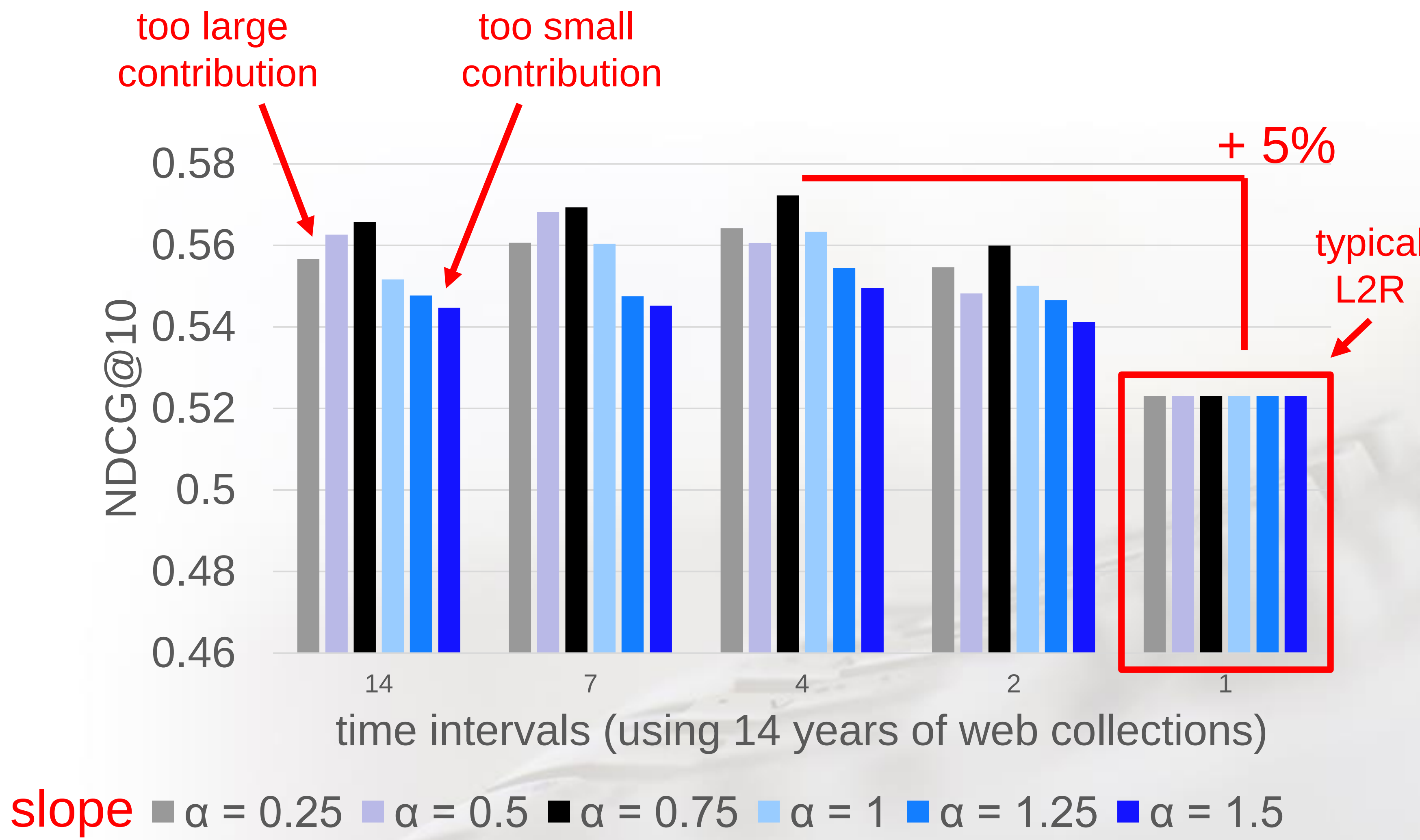
Temporal Features vs. Without Temporal Features

	L2R algorithms (without temporal features)			L2R algorithms (with temporal features)		
Metric	AdaRank	Rank SVM	Random Forests	AdaRank	Rank SVM	Random Forests
NDCG@1 NDCG@5 NDCG@10	0.380 0.427 0.470	0.500 0.485 0.523	0.550 0.610 0.650	0.400 0.426 0.476	0.530 0.546 0.571	**0.650** **0.665** **0.688**

+ 10%

All results show a statistical significance of $p<0.05$ with a two-sided paired t-test.

Temporal-Dependent Models vs. Single-models (without temporal features)



Conclusions

Conclusions

Answers to all research questions:

1. Understanding WAIR systems
   - Large increase of initiatives and volume of data, but smaller teams.
   - Only a small part of the web has been preserved.
   - State-of-the-art WAIR technology is optimized for different needs.
   - Some needs are not supported by state-of-the-art WAIR technology.
2. Understanding web archive users
   - Users have mostly navigational needs and then informational needs.
   - Users search as in web search engines.
   - Users prefer full-text search and older documents.
3. Improving WAIR systems
   - State-of-the-art WAIR systems have low search effectiveness.
   - An extension of the Cranfield paradigm can be used to evaluate WAIR.
   - State-of-the-art WAIR systems can be improved by exploiting temporal information intrinsic to web archives.

Resources

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

Publications

- Daniel Gomes, João Miranda and Miguel Costa, A Survey on Web Archiving Initiatives. In the 1st International Conference on Theory and Practice of Digital Libraries. September 2011.
- Miguel Costa and Mário J. Silva, Understanding the Information Needs of Web Archive Users. In the IPRES2010 10th International Web Archiving Workshop. September 2010.
- Miguel Costa and Mário J. Silva, Characterizing Search Behavior in Web Archives. In the WWW2011 1st Temporal Web Analytics Workshop. March 2011.
- Miguel Costa and Mário J. Silva, A Search Log Analysis of a Portuguese Web Search Engine. In the INForum - Simpósio de Informática. September, 2010.
- Miguel Costa and Mário J. Silva, Evaluating Web Archive Search Systems. In the 13th International Conference on Web Information System Engineering. November 2012.
- Miguel Costa and Mário J. Silva, Towards Information Retrieval Evaluation over Web Archives (poster). In the SIGIR 2009 Workshop on the Future of IR Evaluation. July 2009.
- Miguel Costa and Francisco M. Couto and Mário J. Silva, Learning Temporal-Dependent Ranking Models. In the 37th Annual ACM SIGIR Conference. July 2014.
- Daniel Gomes, Miguel Costa, David Cruz, João Miranda and Simão Fontes, Creating a Billion-Scale Searchable Web Archive. In the WWW2013 3rd Temporal Web Analytics Workshop. May 2013.

Thank you.