LASIGE
Large-Scale Informatics Systems Laboratory

inesc id lisboa

U LISBOA | UNIVERSIDADE DE LISBOA

Learning Temporal-Dependent Ranking Models

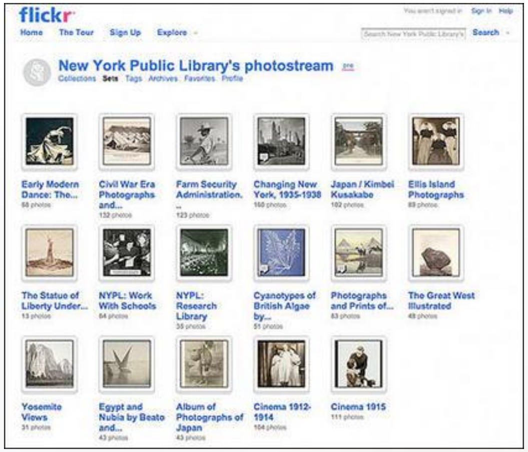
Miguel Costa, Francisco Couto, Mário Silva
LaSIGE @ Faculty of Sciences, University of Lisbon
IST/INESC-ID, University of Lisbon

37th Annual ACM SIGIR Conference, Gold Coast, Australia
July 10, 2014

Our Memory is in Digital Form

E-books

Web photo galleries



Forums

Blogs

Online newspapers

Social networks

The Web is Ephemeral

- 50 days - 50% of documents are changed

  (Cho and Garcia-Molina. 2000)

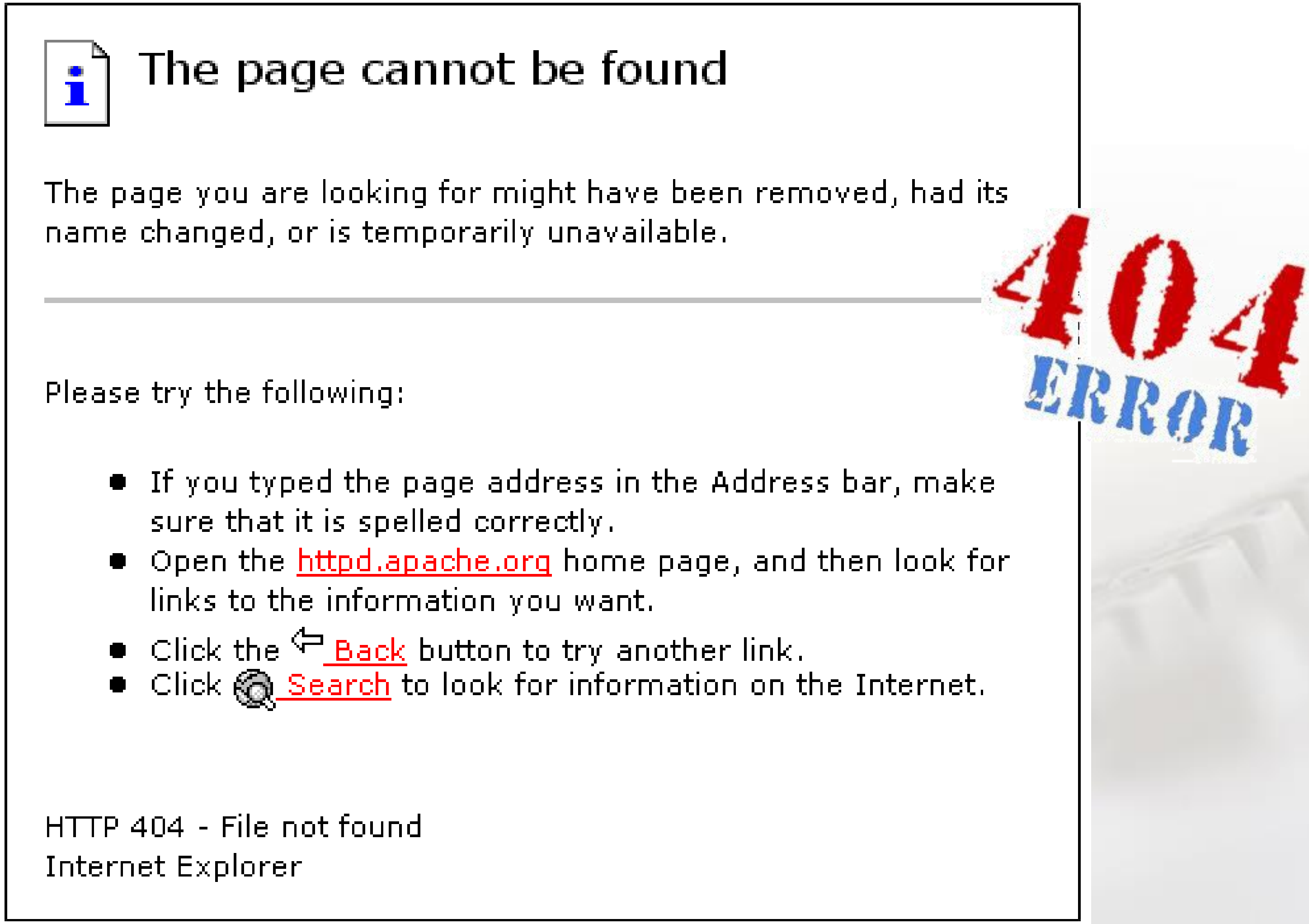
- 1 year - 80% of documents become inaccessible

  (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

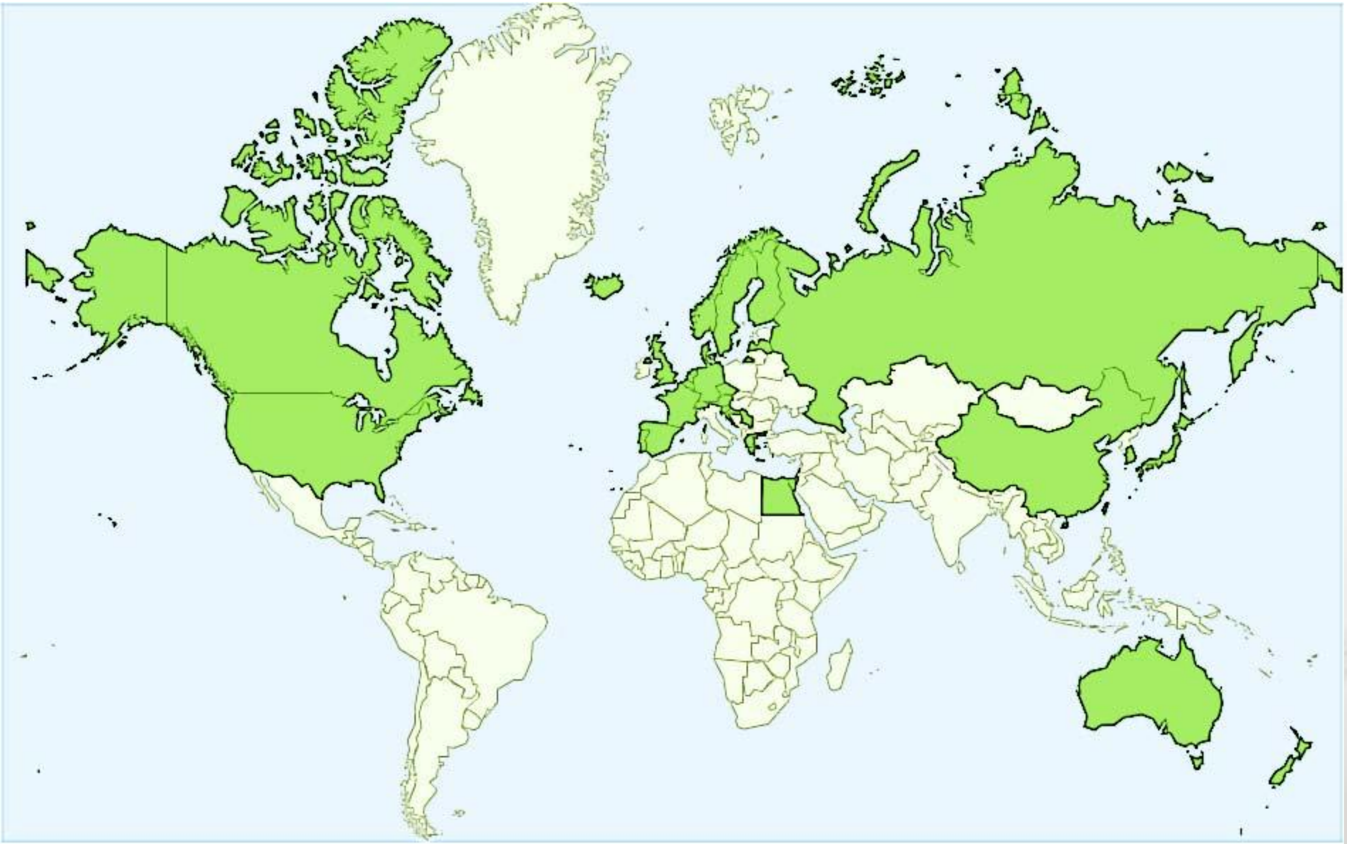
The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

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

2014: Web Archiving Initiatives



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

PWA Search System

- Available since 2010: http://archive.pt
- 1.2 billion documents
  - searchable by full-text and URL
  - range between 1996 and 2013

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

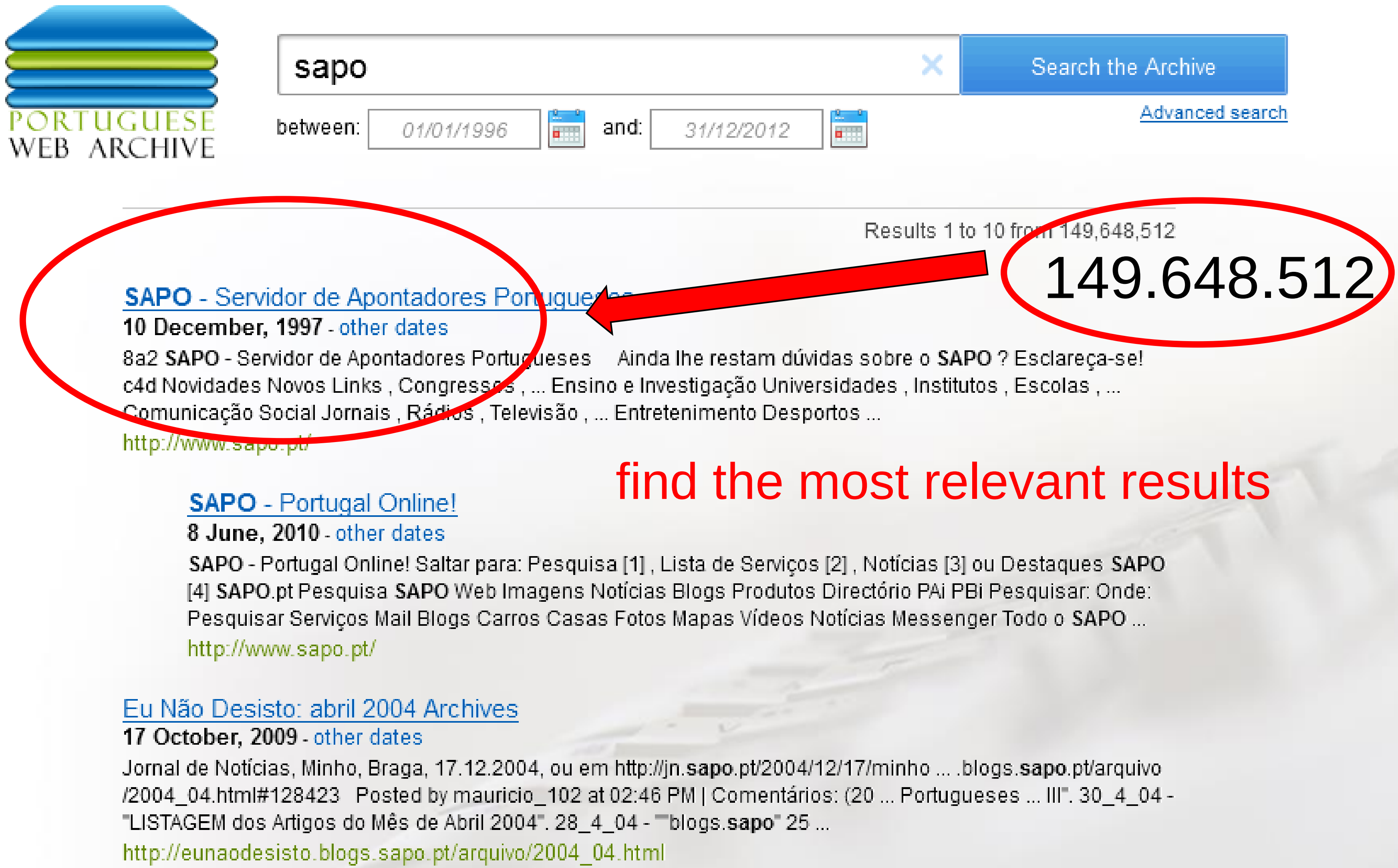
We archived 1,832 versions of the Web page http://sapo.pt from 1 January, 1996 and 26 August, 2013.

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 1997


SAPO
Informações
Correio Electrónico
TOP SAPO
Novidades
Imagens Satélite
Pesquisar
Opções
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



How to find the best search results for a given query in a **Web Archive**?

Typical solution: combine a set of proven ranking features using learning-to-rank (L2R) algorithms

Contributions

We describe how to leverage the **temporal dimension** of web data by:

1. designing novel ranking features that exploit correlations between archived data and relevance

2. designing a novel ranking framework that learns models considering variations of data over time

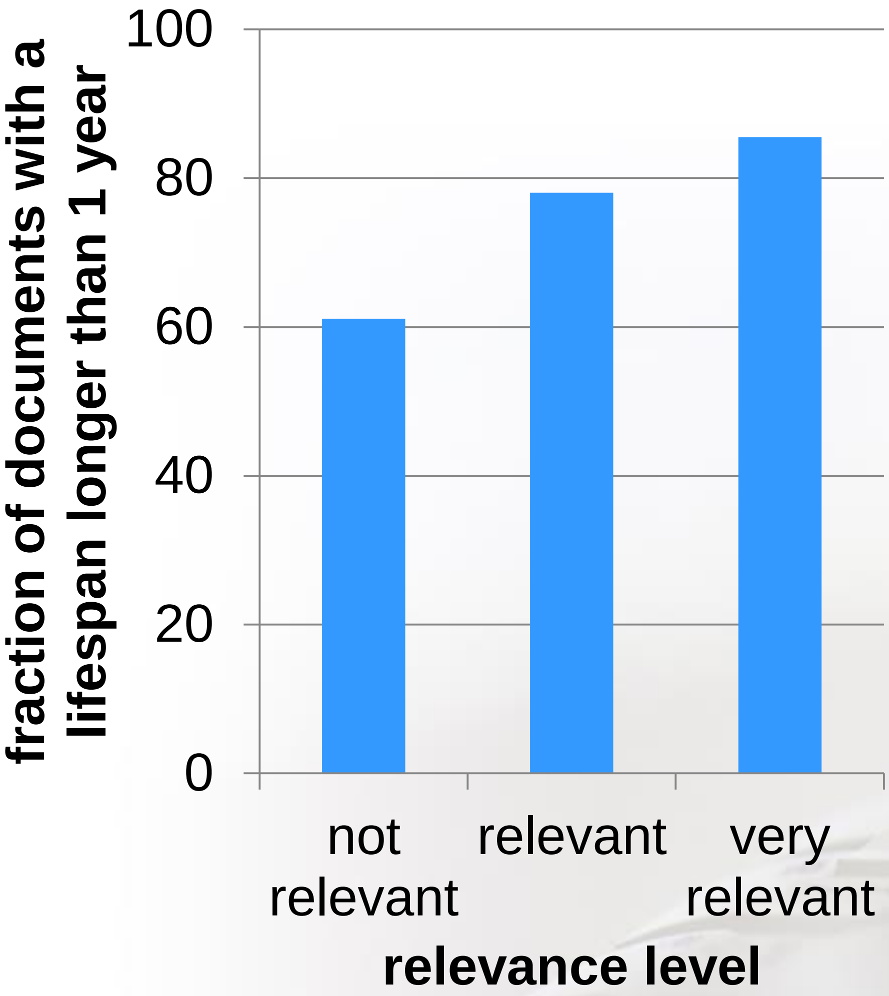
Temporal Features

Long-term Document Persistence

- Predominant user information need: **navigational**.
- Query-independent ranking features do not work well
  - Much smaller volume of clicks
  - Sparser web-graphs
- We need alternatives

- Are long-term persistent documents more relevant?
- How to measure persistence?
  - lifespan
  - number of versions

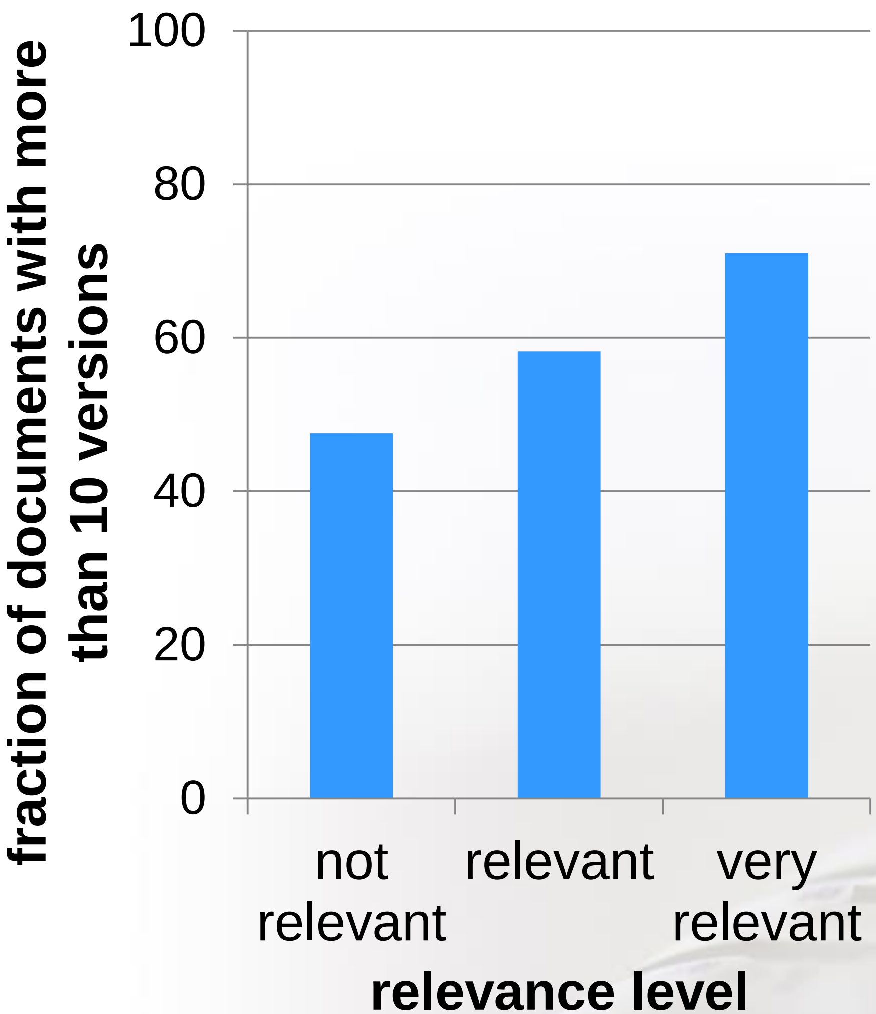
Lifespan & Relevance



documents with higher relevance tend to have a longer lifespan

14 years of web snapshots analyzed

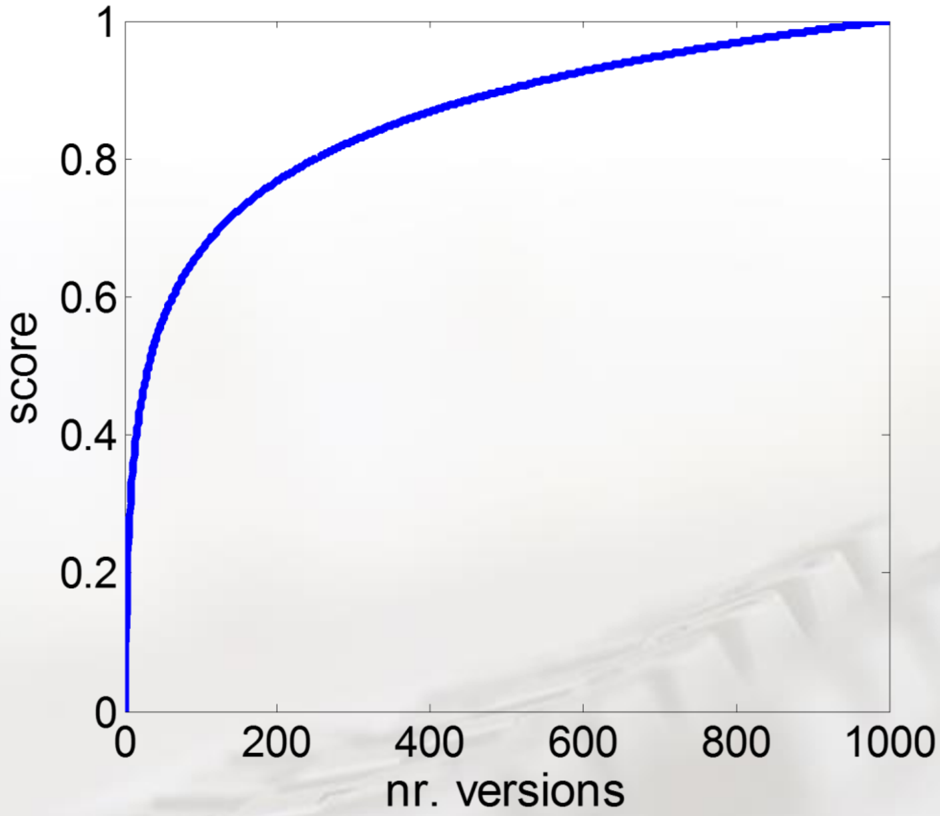
Versions & Relevance



documents with higher relevance tend to have more versions

14 years of web snapshots analyzed

Modeling Document Persistence



$$f(d) = log_y(x)$$

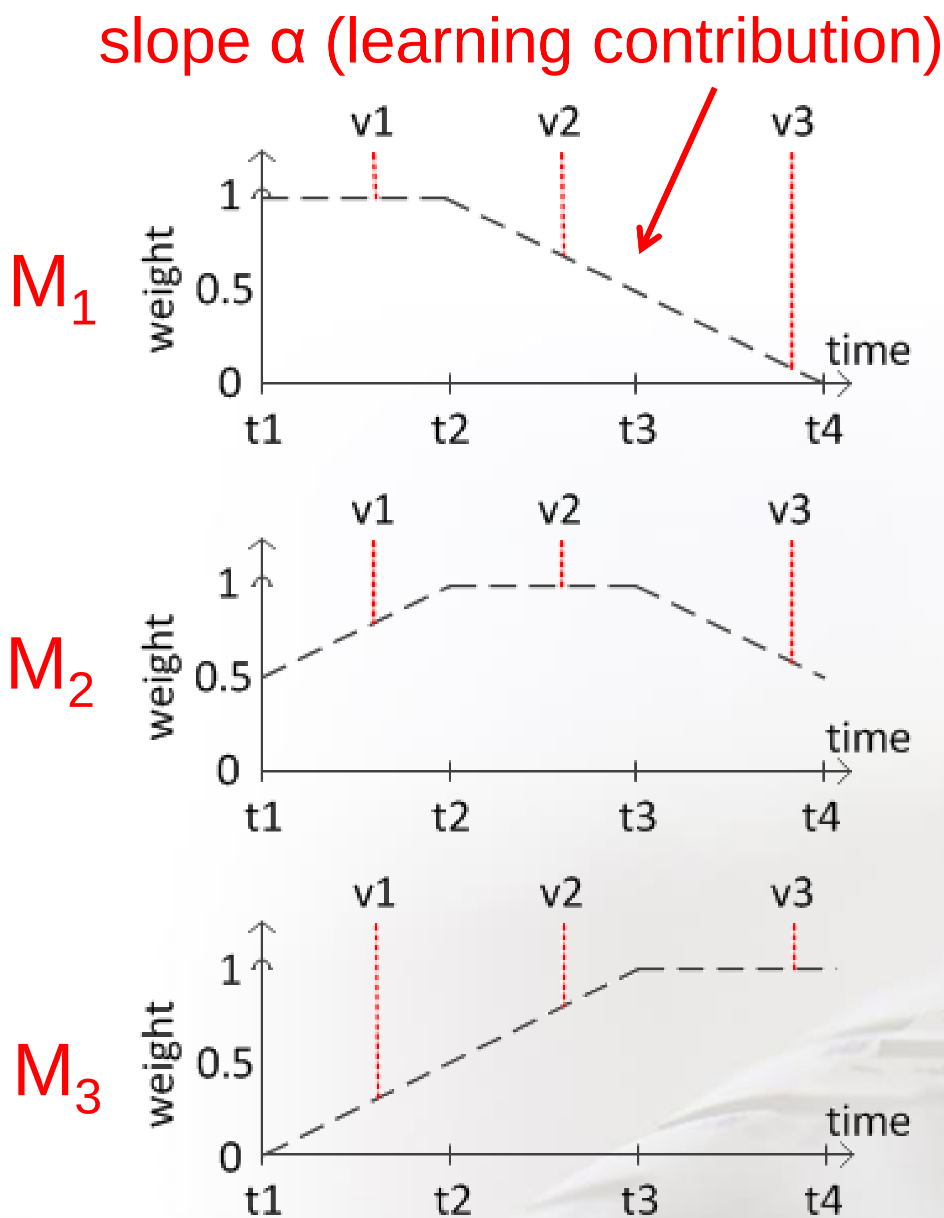
Parameters:
x = #versions/lifespan of document d
y = maximum #versions/lifespan of a document in the collection

Temporal-Dependent Ranking Models

Temporal-Dependent Ranking

- The web has different characteristics over time:
  - more sites and pages
  - longer contents
  - different technologies
  - slightly different language
  - denser web-graphs
- Should we use a single-model that fits all data?
  - No: [Kang & Kim 2003; Geng et al. 2008; Bian et al. 2010]

Temporal Intervals



- use all data (do not split data by time)
- closer periods are more likely to hold similar web characteristics

- Example:
  - 3 intervals
  - T= { [t1,t2] ,]t2,t3] ,]t3,t4] }

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

Global Loss Function

- Results of temporal models are sub-optimal and hard to combine.
- Minimize a global loss function (correlation and overlap between models are considered).

n = # temporal intervals

$$model_1, \ldots, model_n = argmin_{f_1, \ldots, f_n} \sum_{i=1}^{m} L\left(\sum_{j=1}^{n} \Upsilon(x_i, Tj)\ f_j(x_i, \omega), y_i\right)$$

- *Scoring follows the global loss function.*

$$score(x_i) = \sum_{j=1}^{n} \Upsilon(x_i, Tj)\ f_j(x_i, \omega)$$

Experimental Setup

Research Questions

- Do temporal features extracted from web archives improve Web Archive IR?
  - Created a L2R dataset
  - L2R algorithms used: AdaRank, RankSVM, Random Forests.
  - L2R algorithms compared using the dataset with and without temporal features.

- Does the temporal-dependent ranking framework outperforms L2R single-models?
  - L2R algorithms used: RankSVM and TD RankSVM.
  - Temporal-dependent models compared with single-models.

Dataset for L2R in Web Archives

- 39 608 quadruples <query, version, grade, features>
  - 50 queries randomly sampled from logs
  - 843 versions assessed on average per query
  - 3-level scale of relevance
  - 68 ranking features extracted (including temporal)

- LETOR file format:

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 Methodology

- Test Collection (based on Cranfield Paradigm):
  - **Corpus**: 6 web collections, 255M contents, 8.9TB
    - broad crawls, selective crawls, integrated collections
  - **Topics**: 50 navigational (with date range)
    - e.g. the page of Publico newspaper before 2000.
  - **Relevance Judgments**: 3 judges, 3-level scale of relevance, 267 822 versions assessed
  - **Metrics**: (NDCG@k, P@k | k=1,5,10)

- 5-fold cross-validation
  - 3 folders for training, 1 for validation, 1 for testing

Results

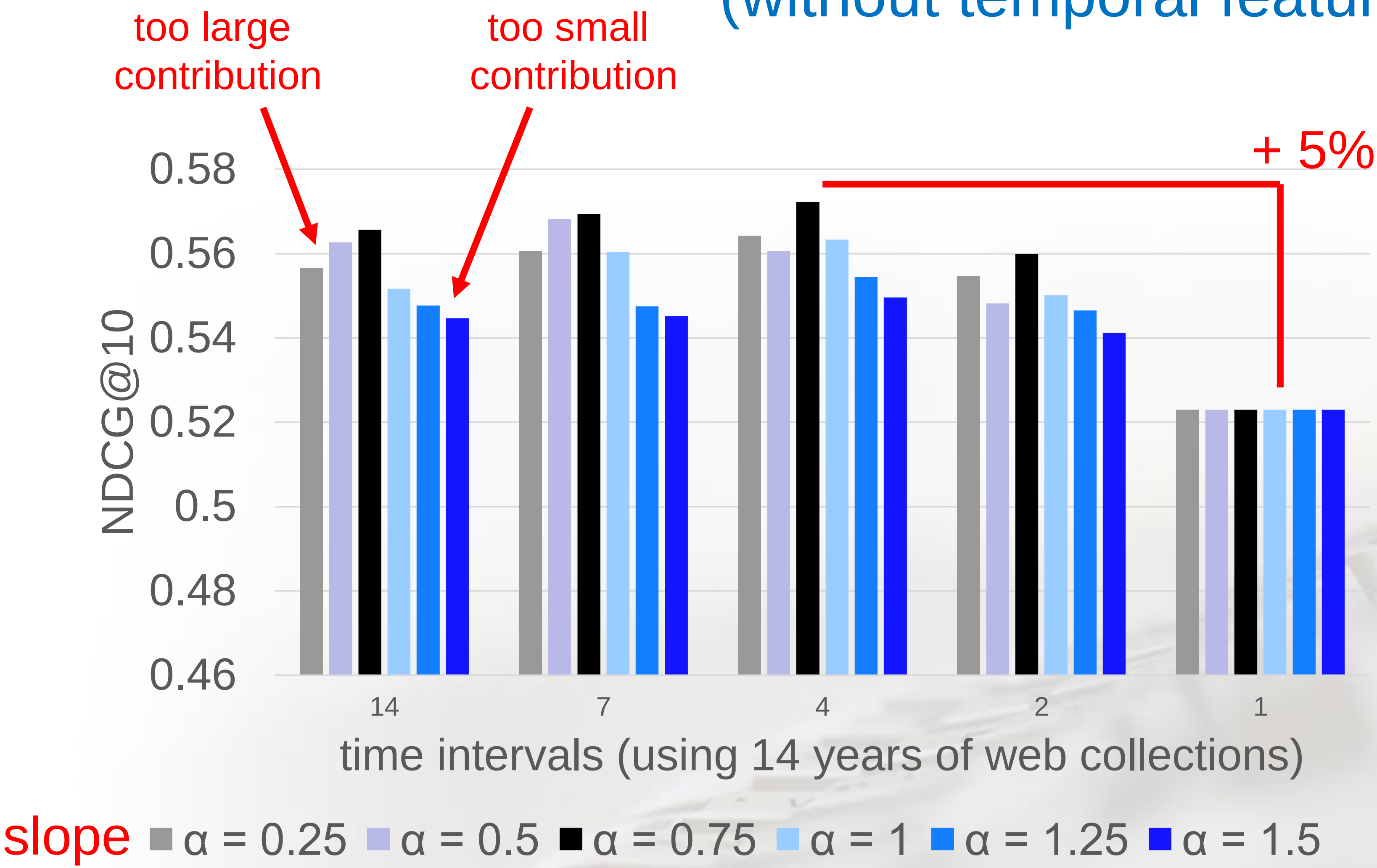
Temporal Features vs. Without Temporal Features

	L2R algorithms (without temporal features)			L2R algorithms (68 features)		
Metric	AdaRank	Rank SVM	Random Forests	AdaRank	Rank SVM	Random Forests
NDCG@1	0.380	0.500	0.550	0.400	0.530	**0.650**
NDCG@5	0.427	0.485	0.610	0.426	0.546	**0.665**
NDCG@10	0.470	0.523	0.650	0.476	0.571	**0.688**

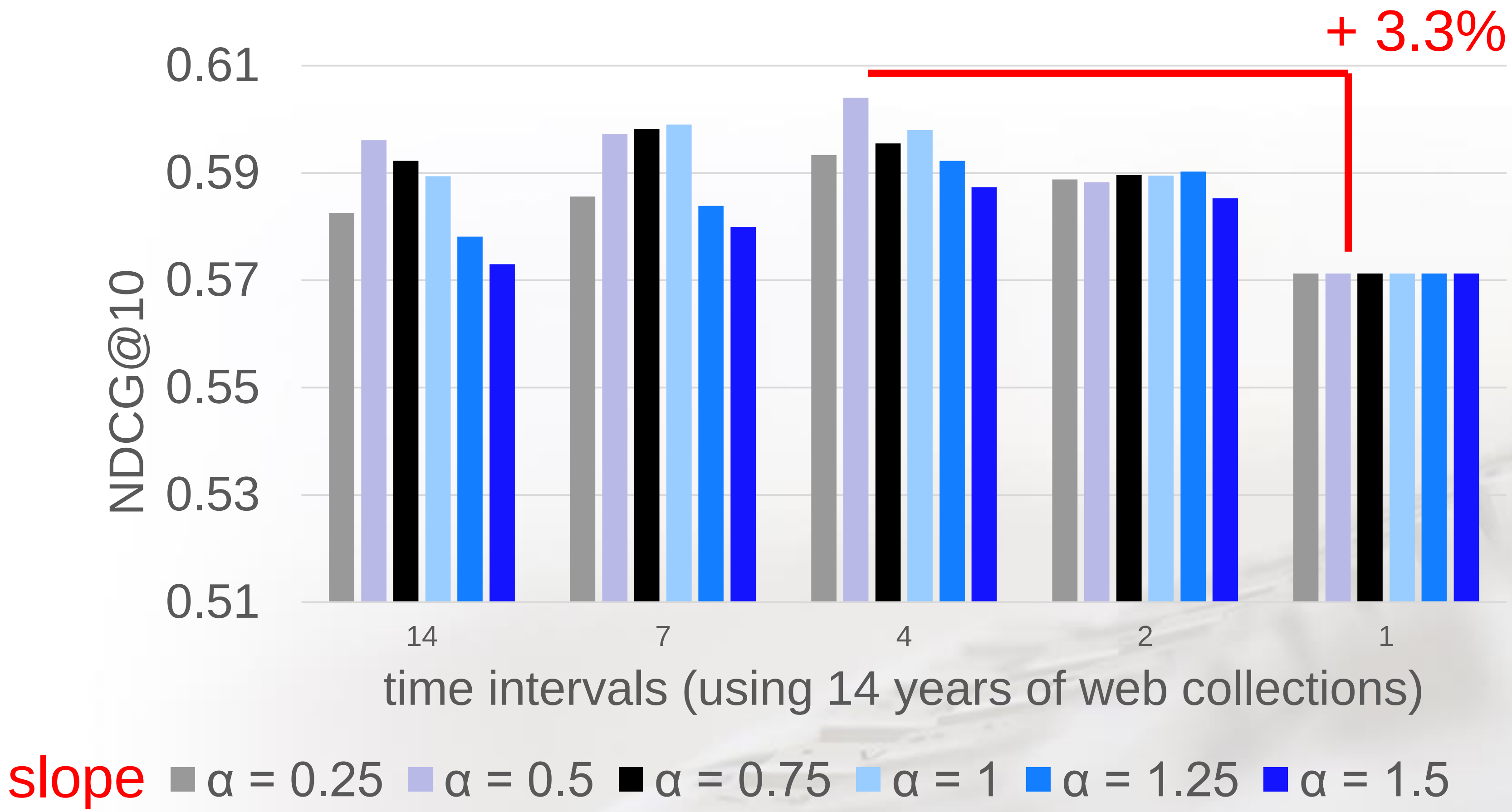
+ 10%

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)



Temporal-dependent models vs. Single-models (with temporal features)



Conclusions

Conclusions

- The evolution of web data over time can be exploited to improve the ranking of search results:
  - by designing novel temporal features
    - Relevant documents tend to have a longer lifespan and more versions.
  - by considering time when learning models
    - A model per period outperforms a single-model.

  (Combined techniques produce the best results)

- Web archives are an excellent source to provide temporal information to web search systems.

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 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. Questions?

migcosta@gmail.com