

# Learning Temporal-Dependent Ranking Models

tisboa

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

Google books goethe	Search Books
Faust By Johann Wolfgang von Go	the, Bayard Taylor
Continue Peace Reserves (27) Bay Bay Search in this book drive drive drive drive Get this Albenos Alb	RAUST. A TRAGEDY BY JOHANN WOLFGANG VON GOETHE.
View all related books.a	The First Part. TRANSLATED, IN THE ORIGINAL METRES,



Web photo galleries

#### Forums

p	Notes States States	36	0	Martin 1 Martin
	Community Brief Glassen Brieferer Blassener Blassener			
	um in Tan Jan 20, 2015 J.19 pm am Porson Enden			Ver and send a
	SOUTPEDIA	1466	10000	1007 7001
Cener				
0	Annisemeersenis Raad vee Final before publice anyorkerel Montonie (several matter) statustere fann	133	185	Han Jun 27, 2005 6.14 ( feat.flat) #0
-				
0	photo Stupport Ger lety with institution and running physical 2.6 is here. Please <b>de not</b> post buy reports, feature requests or MSD related questions here. Pedienters Davies, Malamine Jano, Davied Jano	100012	792842	The law 29, 2005 2119 (
0	Converting from their baset articles of Dard decision liked high face a soundar about a converter? With to other a converter packaget high baset, pack baseters produced their baset articles are based to be a sound to be a sound to be a sound to be a sound to be a water increased to be about the support from waters increased to be about the support from the sound baset baset to be a support from the sound baset baset to be a support from the sound baset	3577	(1341	Sector 28, 2000 2,271 (seconds =0)
0	physite Descension Concerse physite here, phone do not put suggert requests, buy repets, or feature requests' tor-physite restes suestans and descension guines Serveral Descenses Descenses	13185	87575	ha he 28, 2005 3.55 p Statu #G
-	Madifications			
0	MOD Release Annexemption & Support All see MOS interact in an <u>SUC cluster</u> will be annumed in term. All signed for released MODs reach to take pace in term. Release Reports 10,000 annum, NOD Same	427	11877	The Jun 28, 2005 2.14 p
0	MOD Requests Gal a NOD requests? hexist news, since physikil does not have afficial NOD Authors, as such, physikils not requested for developing your requested NOD - Meeterin Teaching, Jacob 2000 2000	29433	99942	Ter Jan 20, 2010 2-28 p
0	NODe in Development A plan for NOD Autors to and and reasons tradition without still in development. (No under within this forum should be used within a live environment). Reasons traditional plan, mod hum	2494	112791	Ther Serie 20, 2000 3:04 (
0	MOD Writers Discussion Concessor forum for MOD Writer requiring MDD Development Memory Hopping, Non, NOD Jon	3439	20813	Tax Ion 28, 2008 3 12 1

#### Blogs



#### **Online newspapers**

Browse Catalog Skin He	¢.			
Mani Herald  Milwaukee-Jour HOME Pu	AGE MY TINES TODAY'S PAPER VI	DED MOST POPULAR TIMES TOPICS	TimesSelect	Free 14-Day T
-576-5 PEPE	ers Wanted.		Jork Eimes	The
PR Online	DE 🛊		Search (B)	* Get Home Delive
Chapters in the second	The A Conservative Faces a Socialist in French Runoff Faces a Conservative A conservative and Socialist in French Runoff Faces a Conservative A conservative and Socialist in French Conservative and Socialist in French Conservative Transformer to the Top Socialist in French Conservative Top Socialist in French Conservative Advance Socialist in French Conservative		variant the question of the season? re raised the question of Access of the season? Access of the season? Acces	a, writes

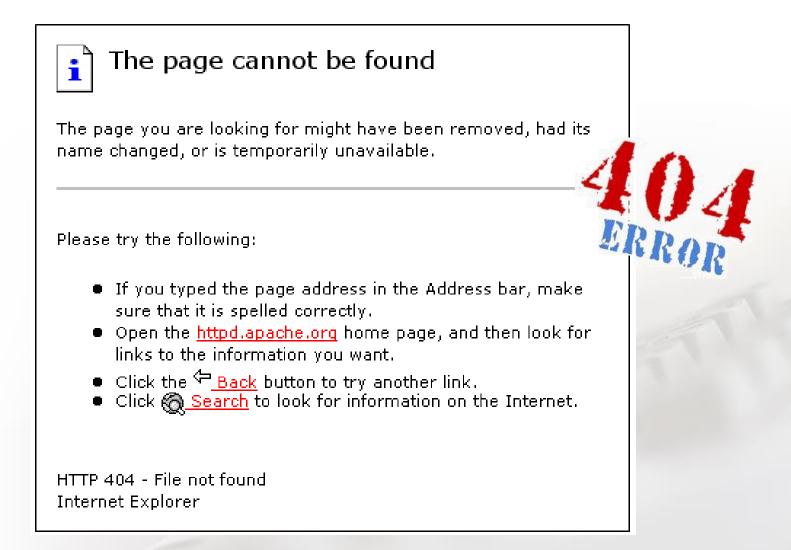
#### Social networks



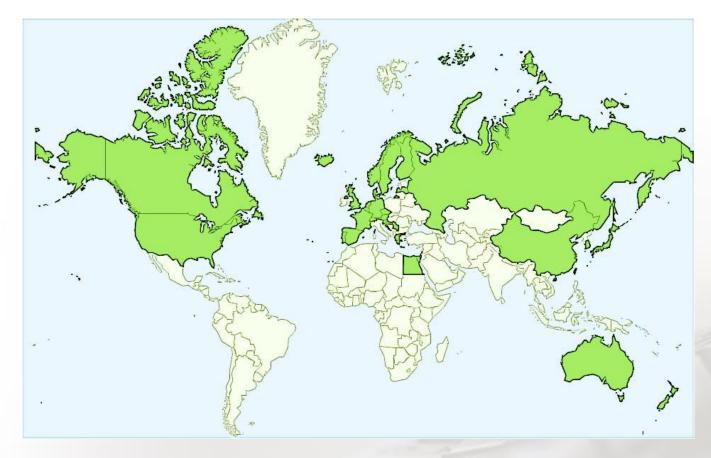
### The Web is Ephemeral

- 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 (<u>http://webcitation.org/</u>. 2007)

## Will we face a Digital Dark Age?



#### 2014: Web Archiving Initiatives



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

## **PWA Search System**



Search the Archive

Advanced search

×



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  $\underline{1996}$  and  $\underline{2010},$  at your disposal.

Know the project

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

#### **URL Search**

7



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

## SAPO.PT 1997



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

- <u>Novidades</u>
   <u>Novos Links, Congressos, ...</u>
- <u>Ensino e Investigação</u> <u>Universidades, Institutos, Escolas, ...</u>
- <u>Comunicação Social</u> Jomais, <u>Rádios</u>, <u>Televisão</u>, ....
- <u>Entretenimento</u>
   <u>Desportos, Fora de Casa, Música, ...</u>
- <u>Serviços de Informação</u> <u>Software, Mailing Lists, IRC, ...</u>

- <u>Comércio, Indústria e Serviços</u> <u>Serviços, Informática, Saúde, Lojas, ...</u>
- <u>Páginas Pessoais</u>
   <u>Páginas pessoais, Lista de E-Mails</u>
- <u>Sociedade e Cultura</u> <u>Museus, Hospitais, Religião, Governo, ...</u>
- <u>Regional</u> <u>Câmaras Municipais, Turismo, Timor, ...</u>
- <u>Computadores e Internet</u> <u>Docs, Web Designers, Software, ISPs, ...</u>



#### **Full-text Search**

149.648.512



Results 1 to 10 from 149,648,512

#### SAPO - Servidor de Apontadores Portuguer

10 December, 1997 - other dates

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

http://www.sapo.pt/

#### find the most relevant results

#### 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

# 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

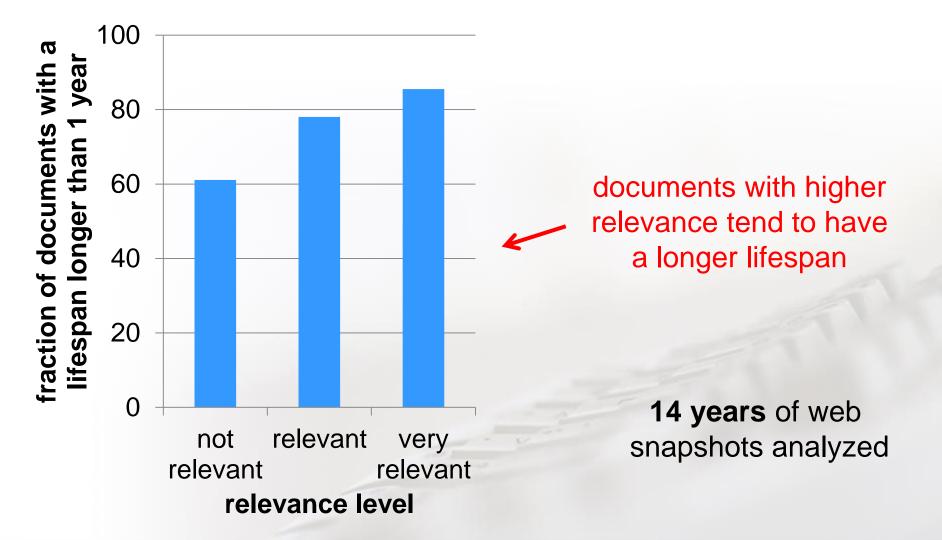
# **Temporal Features**

## Long-term Document Persistence

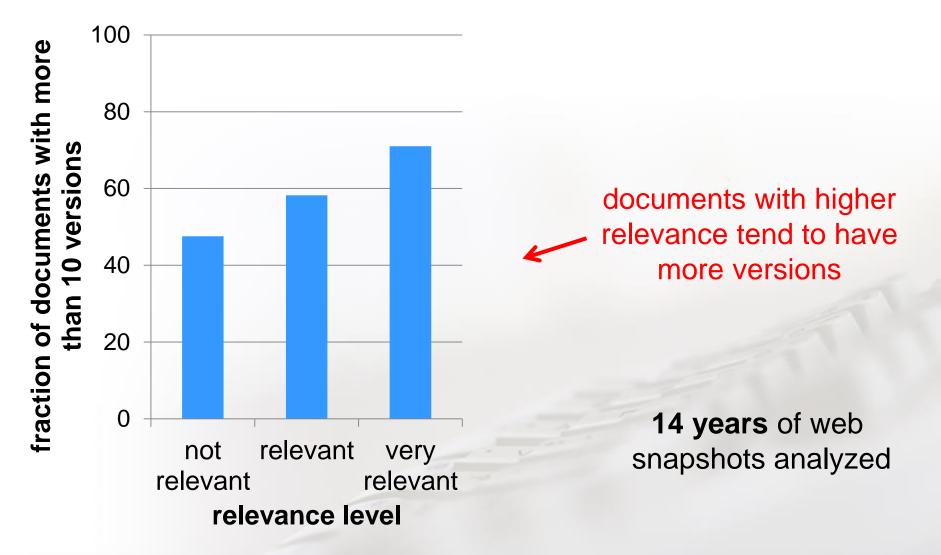
- Predominant user information need: navigational.
- Query-independent ranking features do not work well
  - Much smaller volume of clicks
  - Sparser web-graphs
- We need alternatives

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

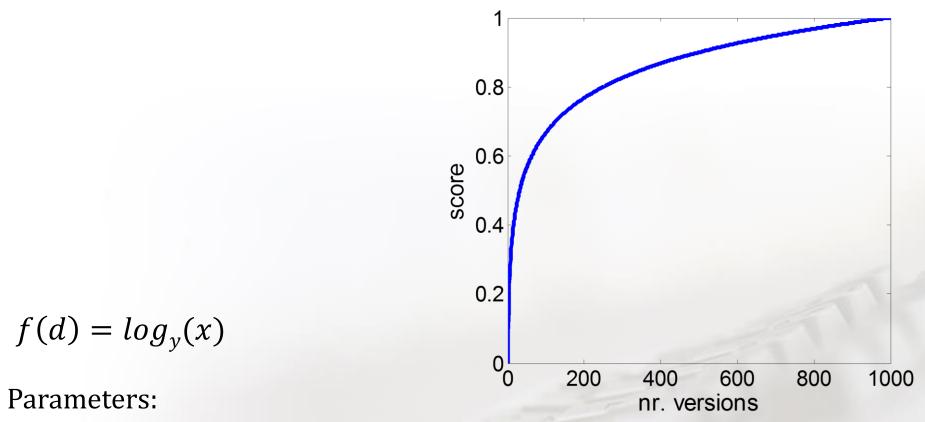
#### Lifespan & Relevance



#### # Versions & Relevance



#### **Modeling Document Persistence**

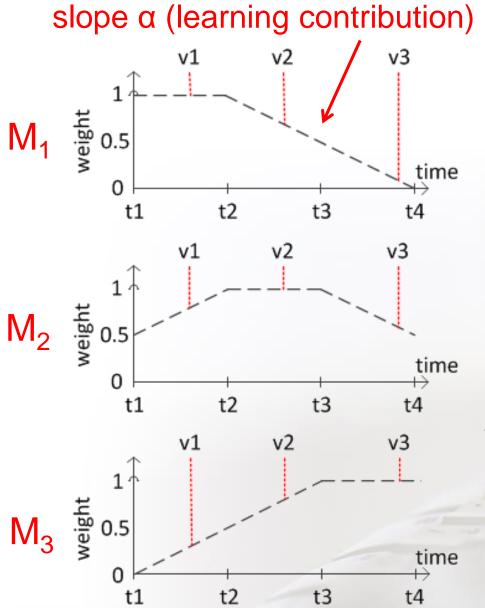


- x = #versions/lifespan of document d
- y = maximum #versions/lifespan of a document in the collection

# 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]

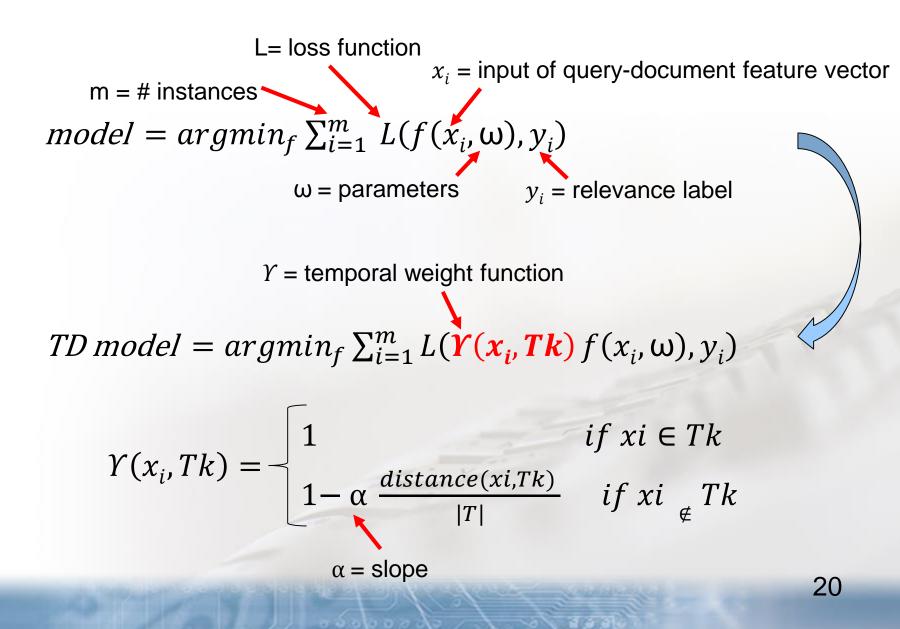


## **Temporal Intervals**

- 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**



## **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).

 $model_{1}, ..., model_{n} = argmin_{f_{1}, ..., f_{n}} \sum_{i=1}^{m} L\left(\sum_{j=1}^{n} \Upsilon(x_{i}, Tj) f_{j}(x_{i}, \omega), y_{i}\right)$ 

Scoring follows the global loss function.

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

# **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

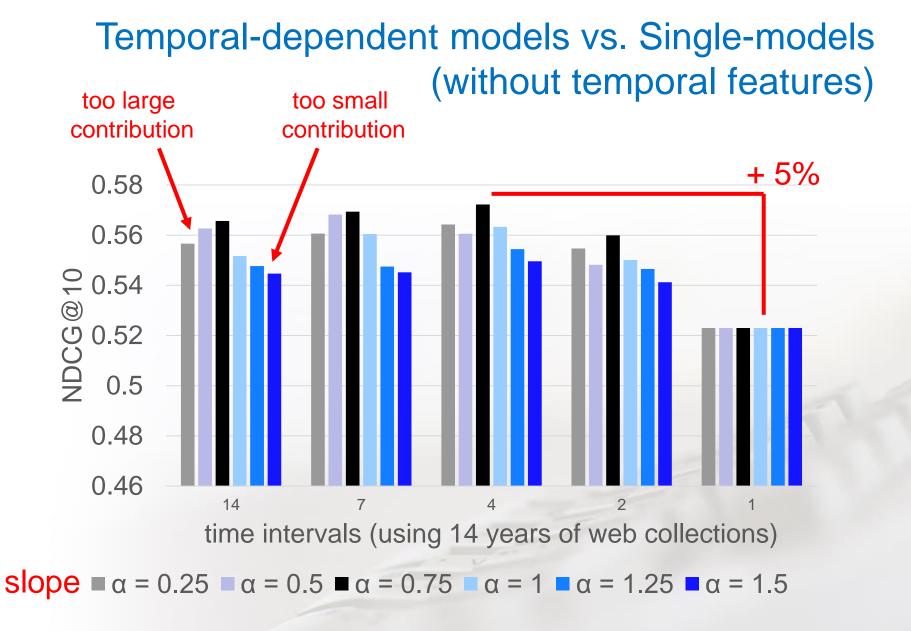


#### Temporal Features vs. Without Temporal Features

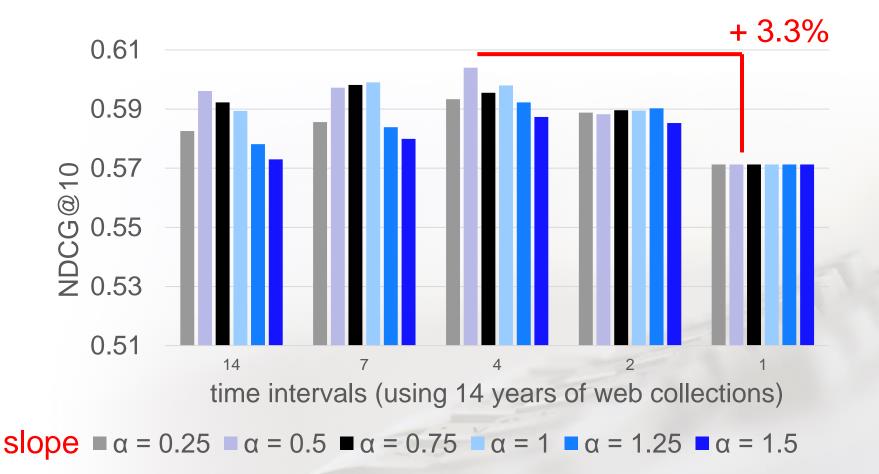
		R algorit empora	thms Il features)	L2R algorithms (68 features)		
Metric	AdaRank	Rank SVM	Random Forests	AdaRank	Rank SVM	Random Forests
NDCG@1 NDCG@5 NDCG@10	0.380 0.427 0.470	0.500 0.485 0.523	0.550 0.610 0.650	0.400 0.426 0.476	0.530 0.546 0.571	0.650 0.665 0.688

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

+ 10%



# 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:
  - <u>http://archive.pt</u>
- OpenSearch API:
  - <u>http://code.google.com/p/pwa-technologies/wiki/OpenSearch</u>
- Test collection to support evaluation:
  - <u>https://code.google.com/p/pwa-technologies/wiki/TestCollection</u>
- L2R dataset for web archive IR research:
  - <u>http://code.google.com/p/pwa-technologies/wiki/L2R4WAIR</u>
- All code available under the LGPL license:
  - <u>https://code.google.com/p/pwa-technologies/</u>

## Thank you. Questions?

#### migcosta@gmail.com