Time-Aware Chi-squared for Document Filtering over Time

Tom Kenter
ISLA, University of Amsterdam
tom.kenter@uva.nl

David Graus
ISLA, University of Amsterdam
d.p.graus@uva.nl

Edgar Meij
Yahoo! Research Barcelona
emeij@yahoo-inc.com

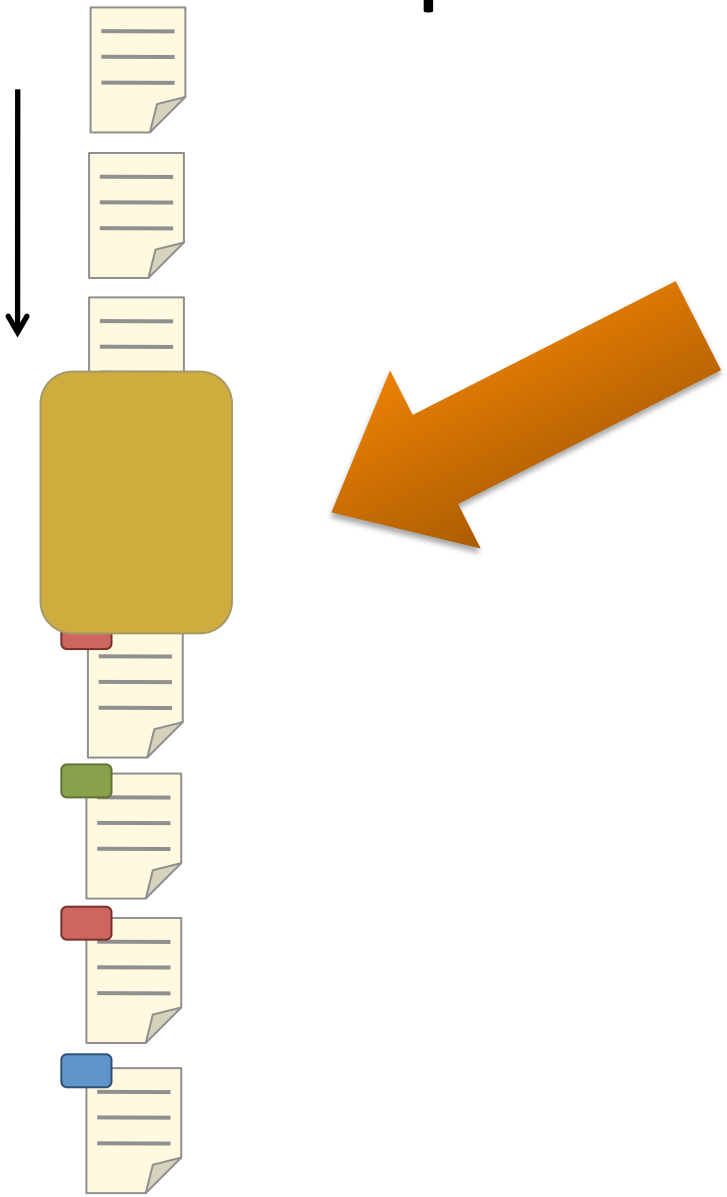
Maarten de Rijke
ISLA, University of Amsterdam
derijke@uva.nl

TAIA 2013
SIGIR 2013 Workshop on Time-aware Information Access
Thursday, August 1st 2013
Dublin, Ireland

TREC 2013
Thursday, November 21 2013
Gaithersburg, United States

CLIN24
Friday, January 17 2014
Leiden

Topic Filtering over Time



Classifier

Multinomial Naive Bayes classifier with feature selection

- Evolving topics
- Online learning

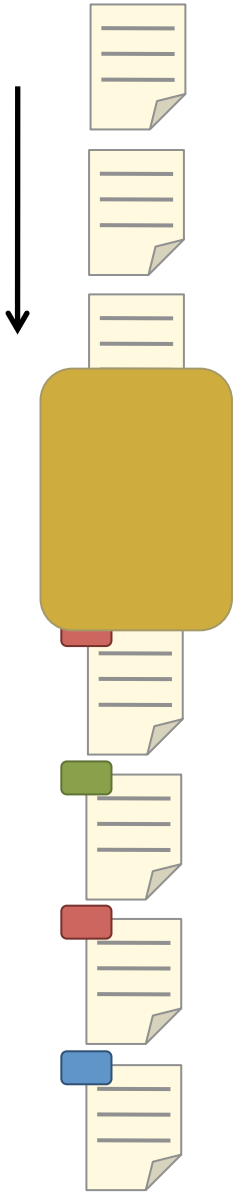
Feature selection with χ^2

Contingency table

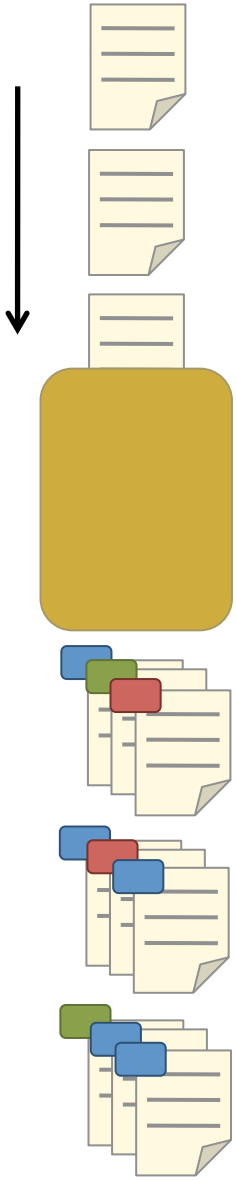
	w	¬w	
c	a	b	C
¬c	c	d	¬C
	W	¬W	N

$$\chi^2 = \frac{N(ad - bc)^2}{C \cdot \neg C \cdot \neg W \cdot W}$$

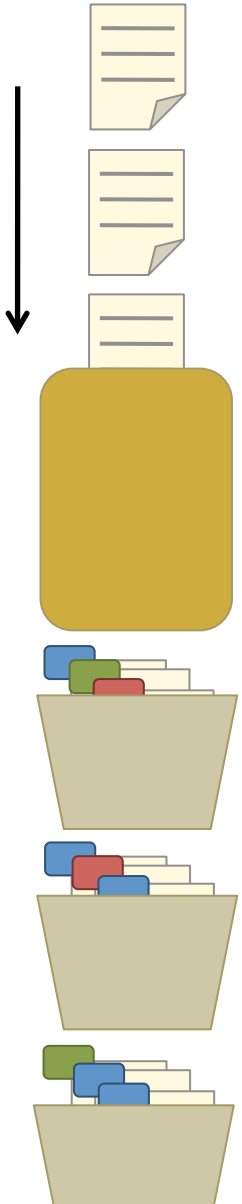
χ^2 over time



χ^2 over time

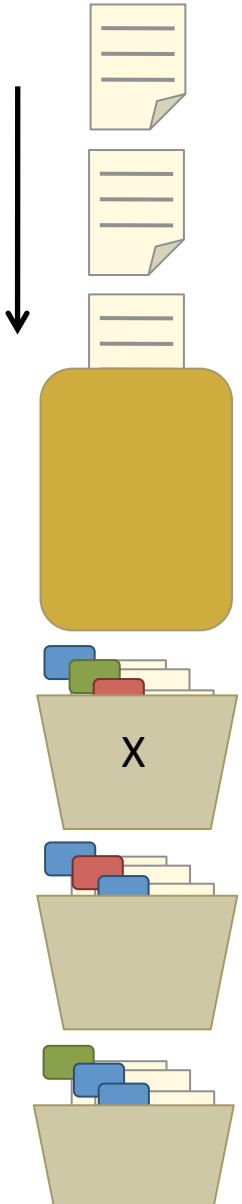


χ^2 over time



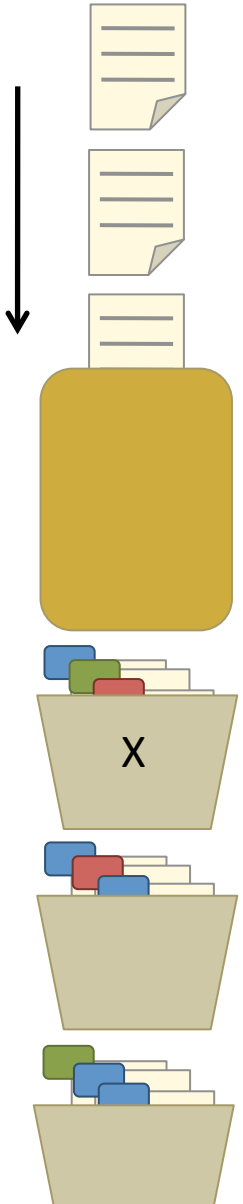
Time batches

χ^2 over time



	w	¬w	
c	a_x	b_x	C_x
¬c	c_x	d_x	$¬C_x$
	W_x	$¬W_x$	N_x

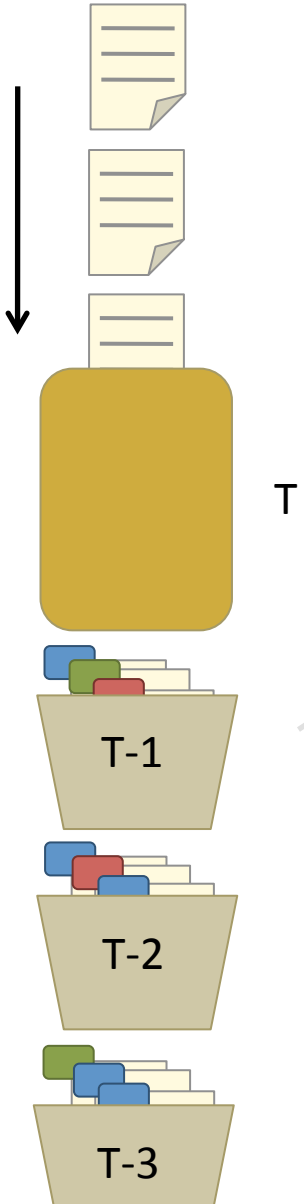
χ^2 over time



	w	¬w	
c	a_x	b_x	C_x
¬c	c_x	d_x	$\neg C_x$
	W_x	$\neg W_x$	N_x

$$\chi^2_x = \frac{N_x(a_x d_x - b_x c_x)^2}{C_x \cdot \neg C_x \cdot \neg W_x \cdot W_x}$$

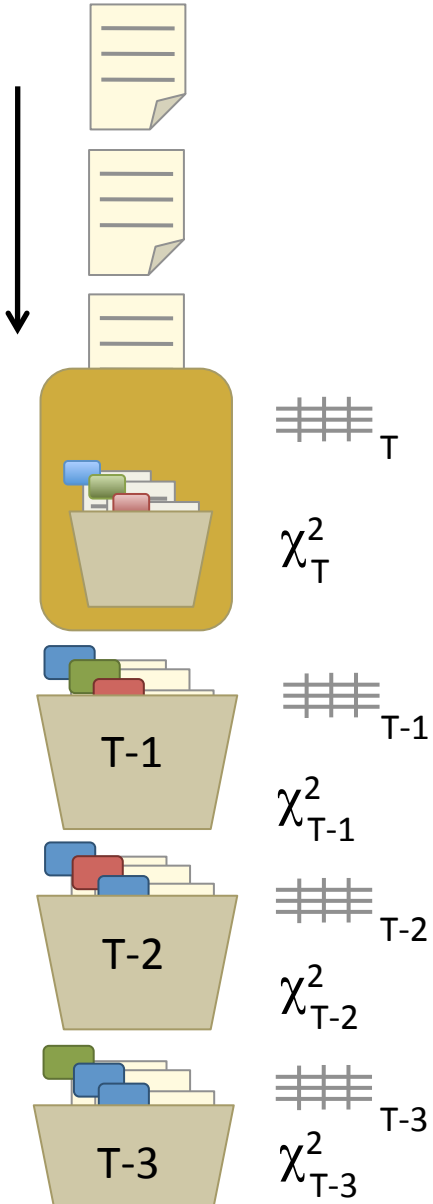
χ^2 over time



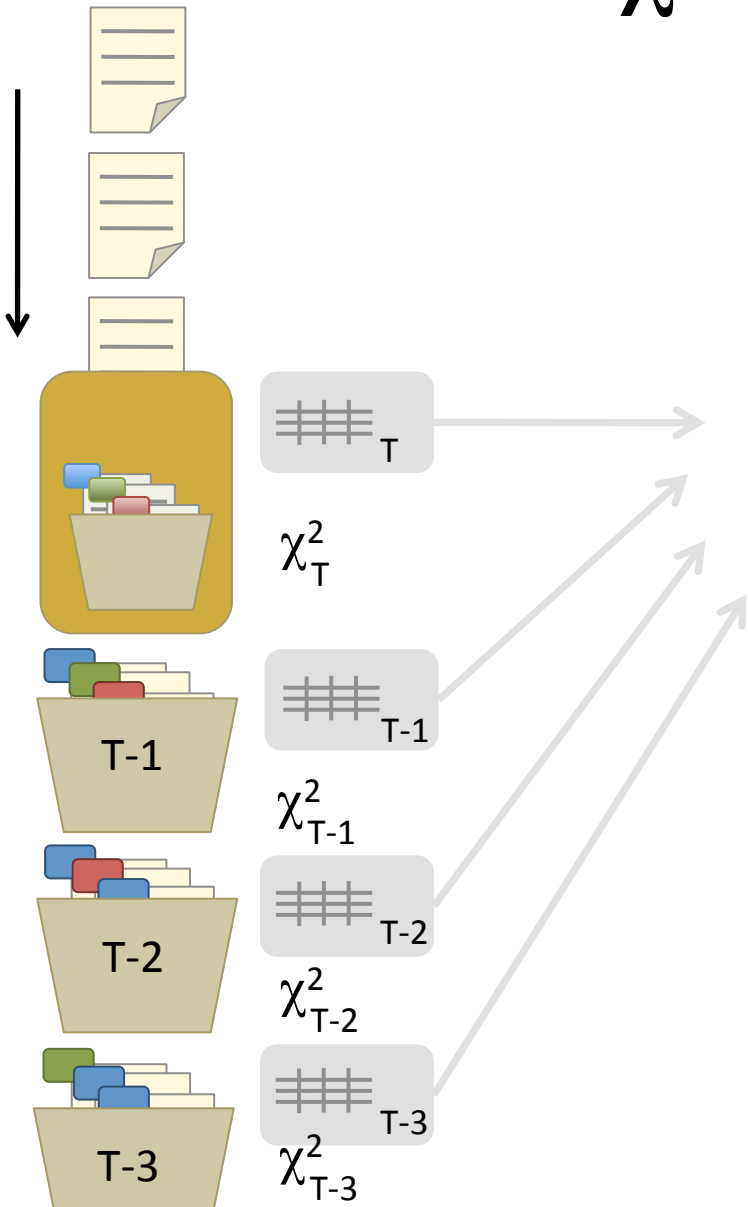
	w	¬w	
c	a_{T-1}	b_{T-1}	C_{T-1}
¬c	c_{T-1}	d_{T-1}	$\neg C_{T-1}$
	W_{T-1}	$\neg W_{T-1}$	N_{T-1}

$$\chi^2_{T-1} = \frac{N_{T-1}(a_{T-1}d_{T-1} - b_{T-1}c_{T-1})^2}{C_{T-1} \cdot \neg C_{T-1} \cdot \neg W_{T-1} \cdot W_{T-1}}$$

χ^2 over time



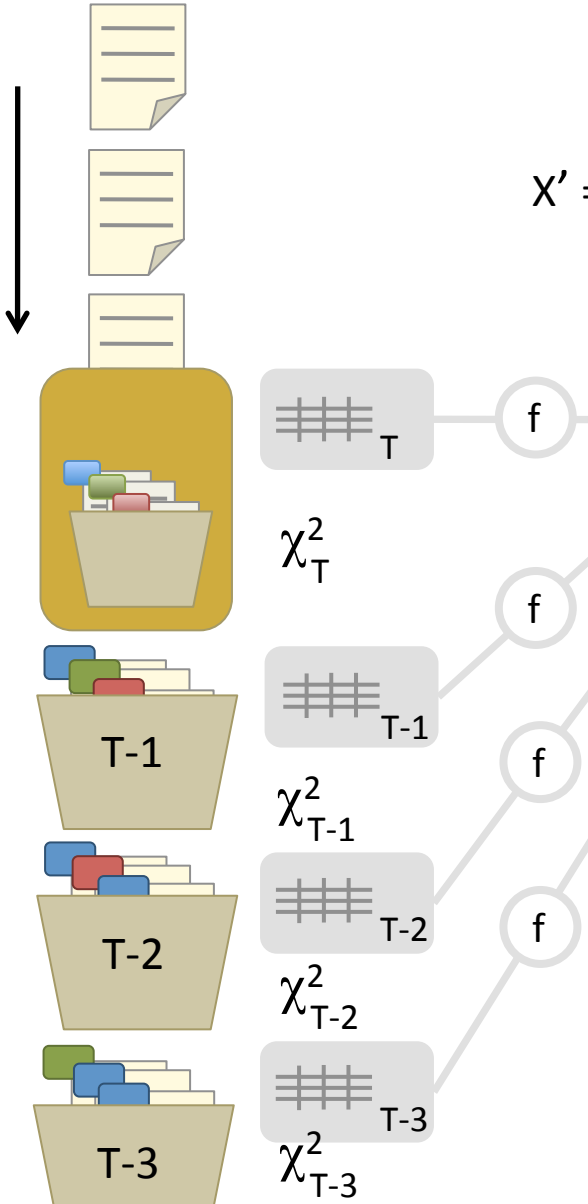
χ^2 over time



Weighted contingencies

	w’	¬w’	
c’	a’	b’	C’
¬c’	c’	d’	¬C’
	W’	¬W’	N’

χ^2 over time



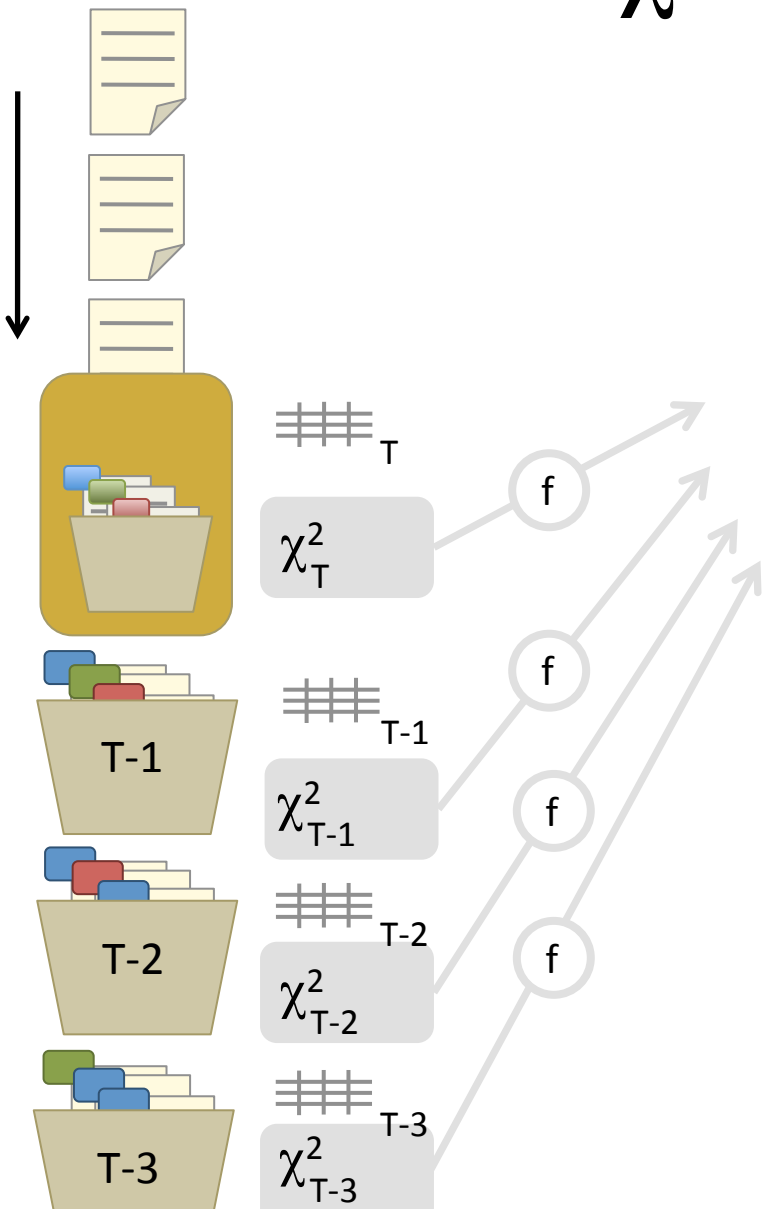
Weighted contingencies

$$X' = \sum_{t=0}^{T} f(T - t) \cdot X_t$$

	w	¬w	
c	a'	b'	C'
¬c	c'	d'	¬C'
	W'	¬W'	N'

$$\chi^2 = \frac{N'(a'd' - b'c')^2}{C' \cdot \neg C' \cdot \neg W' \cdot W'}$$

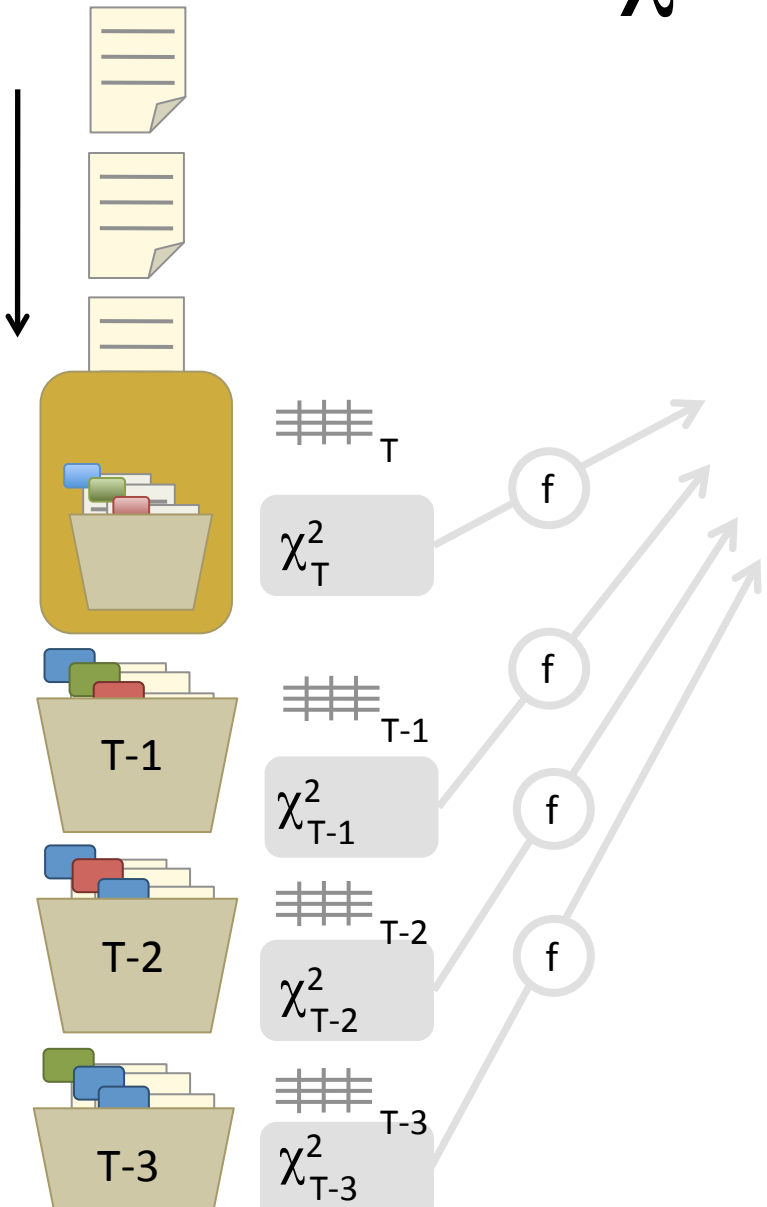
χ^2 over time



Weighted χ^2 values

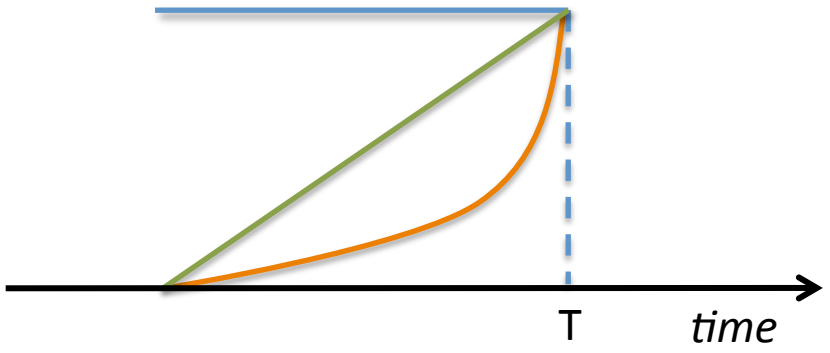
$$\chi^2 = \sum_{t=0}^{T} f(T-t) \cdot \chi^2_t$$

χ^2 over time



Weighted χ^2 values

$$\chi^2 = \sum_{t=0}^{T} f(T-t)\cdot\chi^2_t$$



f(·) = decay function

Experimental setup

- Difference between weighted contigencies and weighted χ^2 values
- Various bucket sizes
- Different decay functions

Difference between two approaches

Weighted contingencies

- Losing distribution of terms per time period
- Better χ^2 estimates
- Might have drag effect
- Can handle small buckets

Weighted χ^2 values

- Preserving saliency information over time
- Less data to work from
- Can adapt quicker
- Suffers from small bucket size

Future work

- Find out which examples to learn from
- Support ‘garbage’ class
- Adaptive feature space

Publications

Tom Kenter, David Graus, Edgar Meij, and Maarten de Rijke. 2013. **Time-aware chi-squared for document filtering over time**. In Proceedings of the SIGIR 2013 Workshop on Time-aware Information Access.

http://ilps.science.uva.nl/sites/ilps.science.uva.nl/files/TimeAwareChiSquared.pdf

Full SIGIR 2013 TAIA workshop proceedings:

http://research.microsoft.com/en-us/people/milads/taia2013.proceedings.final.pdf

TREC KBA CCR 2013 notebook paper:

To be published