Cognitive Temporal Document Priors

Maria-Hendrike Peetz and Maarten de Rijke

ISLA, University of Amsterdam
{m.h.peetz,derijke}@uva.nl

Abstract. Temporal information retrieval exploits temporal features of document collections and queries. Temporal document priors are used to adjust the score of a document based on its publication time. We consider a class of temporal document priors that is inspired by retention functions considered in cognitive psychology that are used to model the decay of memory. Many such functions used as a temporal document prior have a positive effect on overall retrieval performance. We examine the stability of this effect across news and microblog collections and discover interesting differences between retention functions. We also study the problem of optimizing parameters of the retention functions as temporal document priors; some retention functions display consistent good performance across large regions of the parameter space. A retention function based on a Weibull distribution is the preferred choice for a temporal document prior.

1 Introduction

Every moment of our life we retrieve information from our brain: we remember. We remember items to a certain degree: for a mentally healthy human being retrieving very recent memories is virtually effortless, while retrieving non-salient[1] memories from the past is more difficult [2]. Early research in psychology was interested in the rate at which people forget single items, such as numbers. Psychology researchers have also studied how people retrieve events. [3] let users remember entities, which prove to be better remembered if they recently appeared in a newspaper; the authors propose models of how people retrieve terms based on their findings. Similarly, [4,5] record events and hits of web pages related to an event and fit models of how people remember, the so-called *retention function*.

Modeling the retention of memory has a long history in psychology, resulting in a range of proposed retention functions. In information retrieval (IR), the relevance of a document depends on many factors. If we request recent documents, then how much we remember is bound to have an influence on the relevance of documents. Can we use the psychologists' models of the retention of memory as (temporal) document priors? Previous work in temporal IR has incorporated priors based on the exponential function into the ranking function [6,7,8,9]—this happens to be one of the earliest functions used to model the retention of memory. Many other such functions have been considered by

[1] Salient memories are very emotional memories and traumatic experiences; human retrieval of such memories is markedly different [1].

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 318–330, 2013.

psychologists to model the retention of memory—what about the potential of other retention functions as temporal document priors?

Inspired by the cognitive psychology literature on human memory and on retention functions in particular, we consider seven temporal document priors. We propose a framework for assessing them, building on four key notions: *performance*, *parameter sensitivity*, *efficiency*, and *cognitive plausibility*, and then use this framework to assess those seven document priors. For our experimental evaluation we make use of two (temporal) test collections: newspapers and microblogs. We show that on several data sets, with different retrieval models, the exponential function as a document prior should not be the first choice. Overall, other functions, like the Weibull function, score better within our proposed framework for assessing temporal priors.

2 Related Work

We survey cognitive memory models and temporal information retrieval.

Memory Models. Modeling the retention of memory has been a long studied area of interest in cognitive psychology. [10] hypothesizes that retention decays exponentially and supports his hypothesis with a self-experiment. [11] propose a power law model for retention and learning and [12] fit a power function to 100 participants. [13] analyzes probability distributions for their suitability as retention models. [14] show that the exponential functions fit much better. Finally, [2] perform a study with 14,000 participants and compare state-of-the-art memory models and how they fit the retention data. Fig. 1 shows how much people could remember over time. [5] use large-scale experiments to show that the Weibull function is a much better model and the power law can merely be an approximation.

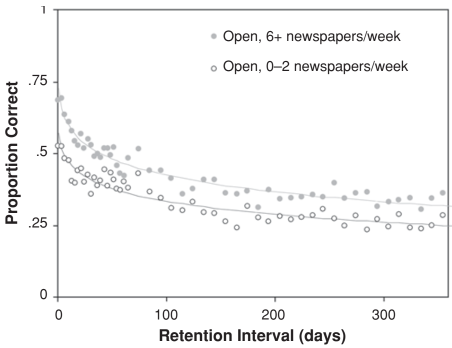


Fig. 1. Retention curves for participants in a study on how they remembered news and fitted retention functions. Plotted separately are participants who read many newspapers ($\geq$ 6/week) and those who read few (0–2/week). Taken from [2].

Temporal Information Retrieval. Temporal IR is a difficult problem. [15] state the main challenges of temporal IR ranging from extracting mentions of time within documents and linking them [16] to spatio-temporal information exploration [17] and temporal querying. We address issues they raise with respect to real-time search. [6] introduce a temporal document prior. This exponential prior imitates the decay of news documents over time and prioritize recent documents. [9] use a similar prior to re-estimate term frequencies. Recent work focusses not only on a recency prior [7] but also on detecting temporally active time periods (salient events) in the temporal distribution of pseudo relevant documents [18,19,20,21]. [19] select top ranked documents in the highest peaks as *pseudo relevant*, while documents outside peaks are considered non-relevant. They use Rocchio's algorithm for relevance feedback based on the top 10 documents. [21] use salient events for query modeling in news and blog data. [22] argues that dynamic functions might be the key. [8] uses document expansion and incorporates a dynamic exponential prior.

We evaluate the effectiveness of the recency priors by incorporating them into query likelihood similar to [6] and query modeling as in [7].

3 Methods

We introduce basic notation and well-known retrieval models into which the temporal document priors that we consider are to be integrated. We then describe several retention functions serving as temporal document priors.

We say that document D in document collection $\mathcal{D}$ has time $\mathit{time}(D)$ and text $\mathit{text}(D)$. A query q has time $\mathit{time}(q)$ and text $\mathit{text}(q)$. We write $\delta_g(q, D)$ as the time difference between $\mathit{time}(q)$ and $\mathit{time}(D)$ with the granularity g. E.g., if $\mathit{time}(q') =$ July 20, 2012 and $\mathit{time}(D') =$ June 20, 2012, we have $\delta_{\text{day}}(q', D') = 30$, $\delta_{\text{month}}(q', D') = 1$, and $\delta_{\text{year}}(q', D') = 0.083$ for a granularity of a day, month, and year, respectively.

Baselines. In order to keep our experiments comparable with previous work, we use the query likelihood model [23,24], both as baseline and as retrieval algorithm for an initially retrieved set of documents. We rank documents by the likelihood $P(D \mid q)$; with Bayes' rule and the assumption that $P(q)$ is uniform, we have $P(D \mid q) \propto P(q \mid D)P(D)$. For query likelihood we set the prior distribution $P(D)$ to be uniform and rank documents by the probability that their model (the multinomial unigram language model) generates the query. Formally, $P(q \mid D) = \prod_{w \in \mathit{text}(q)} P(w \mid D)$. To obtain $P(w \mid D)$, we use Dirichlet smoothing, a linear interpolation between $\hat{P}(w \mid D)$, the maximum likelihood estimate of D, and a document dependent probability of observing w in the background corpus C [24]:

$$P(w \mid D) = \frac{\hat{P}(w \mid D) + \mu\lambda P(w \mid C)}{|D| + \mu}, \tag{1}$$

where μ is the average document length of the collection. A variant of this baseline for recency queries has been proposed by [6]; they use an exponential distribution as an approximation for the prior (see (5)). We use different functions to approximate the prior.

(Temporal) Query Modeling. [7] introduce a query modeling approach that aims to capture the dynamics of topics in Twitter. This model takes into account the dynamic nature of microblogging platforms: while a topic evolves, the language usage around it is expected to evolve as well.

We rank terms according to their temporal and topical relevance selecting the top k:

$$\text{score}(w, q) = \tag{2}$$
$$\log\left(\frac{|\mathcal{D}_{\text{time(q)}}|}{|\{D : w \in D, D \in \mathcal{D}_{\text{time(q)}}\}|}\right) \cdot \sum_{\{D \in \mathcal{D}_{\text{time(q)}} : w_q \in \mathit{text}(q) \text{ and } w, w_q \in \mathit{text}(D)\}} f(D, q, g),$$

where $f(D, q, g)$ is a retention function (introduced below), $\mathcal{D}_{time(q)}$ is the set of documents published before the time of query q, and g is the granularity. The set W_q consists of the top k terms w for query q, sorted by score(w, q). The probability of term t given query q is:

$$P(w \mid q) = \begin{cases} \frac{\text{score}(w,q)}{\sum_{w' \in W} \text{score}(w',q)} & \text{if } w \in W, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We then use KL-divergence [24] to estimate the score of a document D for a query q:

$$\text{Score}(q, D) = -\sum_{w \in V} P(w \mid q) \log P(w \mid D), \tag{4}$$

where V is the vocabulary, i.e., the set of all terms that occur in the collection and $P(w \mid D)$ is the generative probability for a term as specified in (1).

Retention Functions. We introduce a series of retention functions. The *memory chain models* ((5) and (6)) build on the assumptions that there are different memories. The memory model introduced in (5) is equivalent to the exponential prior used in the IR literature. The Weibull functions ((7) and (8)) are of interest to psychologists because they fit human retention behavior well. In contrast, the retention functions *linear* and *hyperbolic* ((10) and (11)) have little cognitive background.

Memory Chain Model. The memory chain model [4] assumes a multi-store system of different levels of memory. The probability to store an item in one memory being μ,

$$f_{\text{MCM-1}}(D, q, g) = \mu e^{-a\delta_g(q,D)}. \tag{5}$$

The parameter a indicates how items are being forgotten. The function $f_{\text{MCM-1}}(D, q, g)$ is equivalent to the exponential decay in [6] when the two parameters (μ and a) are equal. As μ is document independent it does not change the absolute difference between document priors if used for query likelihood and $f_{\text{MCM-1}}(D, q, g)$ is equal to the exponential function used in [6]. In the two-store system, an item is first remembered in short term memory with a strong memory decay, and later copied to long term memory. Each memory has a different decay parameter, so the item decays in both memories, at different rates. The overall retention function is

$$f_{\text{MCM-2}}(D, q, g) = 1 - e^{-\mu_1 \left(e^{-a_1\delta_g(q,D)} + \frac{\mu_2}{a_2 - a_1}(e^{-a_2\delta_g(q,D)} - e^{-a_1\delta_g(q,D)})\right)}, \tag{6}$$

where an overall exponential memory decay is assumed. The parameter μ_1 and μ_2 are the likelihood that the items are initially saved in short and long term memory, whereas a_1 and a_2 indicate the forgetting of the items. Again, t is the time bin.

[13] discusses different memory modeling functions. The prefered function is the Weibull function

$$f_{\text{basic Weibull}}(D, q, g) = \left(e^{-\frac{a\delta_g(D,q)}{d}^{d}}\right), \tag{7}$$

and its extension

$$f_{\text{extended Weibull}}(D, q, g) = b + (1 - b)\mu e^{\left(-\frac{a\delta_g(D,q)}{d}\right)^{d}}. \tag{8}$$

Table 1. Summary of collection statistics for AP, LA/FT, and Tweets2011

	AP (disks 1, 2)	LA/FT (disks 4, 5)	Tweets2011
# documents	164,597	342,054	4,124,752
period covered	02/1988–12/1989	04/1991–12/1994	01/24/2011–02/08/2011
topics	101–200	351–450 (test), 301–350 (train)	MB01–MB49
recent queries	20	16 (train), 24 (test)	–

Here, a and d indicate how long the item is being remembered: a indicates the overall volume of what can potentially be remembered, d determines the steepness of the forgetting function; μ determines the likelihood of initially storing an item, and b denotes an asymptote parameter.

The amended power function has also been considered as a rentention function [12]. The power function is ill-behaved between 0 and 1 and usual approximations start at 1. The *amended power function* is

$$f_{\text{power}}(D, q, g) = b + (1 - b)\mu(\delta_g(D, q) + 1)^a, \tag{9}$$

where a, b, and μ are the decay, an asymptote, and the initial learning performance.

baseline is given by the linear function,

$$f_{\text{lin}}(D, q, g) = \frac{-(a \cdot \delta_g(q, D) + b)}{b}, \tag{10}$$

where a is the gradient and b is $\delta_g(q, \operatorname{argmax}_{D' \in \mathcal{D}} \delta_g(q, D'))$. Its range is between 0 and 1 for all documents in $\mathcal{D}$.

discounting function [25] has been used to model how humans value rewards: the later the reward the less they consider the reward worth. Here,

$$f_{\text{hyp}}(D, q, g) = \frac{1}{-(1 + k * \delta_g(q, D))}, \tag{11}$$

where k is the discounting factor.

4 Experimental Setup

We introduce the data sets, detail a framework of requirements for priors and then proceed with a description of our experiments.

Data. A summary of the collection and topic statistics can be found in Table 1. We have 100 topics for TREC-2, of which 20 are selected as "recent queries" in [9]. We have 150 topics for TREC-{6,7,8}. We use a subset of the topics TREC-{7,8}. This query set was selected in [9], based on its recency. Training and testing data are the queries from TREC-6 and TREC-{7,8}, respectively. The Tweets2011 data set consists of 16 million tweets, collected between 24th January and 8th February, 2011. We consider two flavors of the collection: *filtered* and *unfiltered*; only tweets were returned that have

Table 2. Parameter values for document priors based on retention functions, as fitted on the news training data and as fitted on human data (last column). For cells marked with *, the function was fitted to data with a granularity of milliseconds, otherwise months.

function	parameter	TREC-6 optimized	Tweets2011 optimized	reported values
MCM-1 (5)	r	0.0013	0.2	0.00142* [12]
	μ	1	0.9	3800* [12]
MCM-2 (6)	μ_1	0.7	0.3	0.49–1.29 [2]
	a_1	0.007	0.004	0.018–0.032 [2]
	μ_2	0.6	0.7	0.01–0.018 [2]
	a_2	0.4	0.4	0–0.0010
basic Weibull (7)	a	0.00301	0.3–0.9	–
	d	0.087	0.4	–
extended Weibull (8)	a	0.009	0.1	0.0017–0.0018 [2]
	d	0.7	0.02–0.04	0.087–0.2 [2]
	b	0.1	0.1	0–0.25 [2]
	μ	0.7	0.7	1 [2]
amended power (9)	a	0.03	0.9	840.56* [12]
	b	0.01	0.02	0.33922* [12]
	μ	0.6	1	17037* [12]
linear (10)	a	0.4	1.0	–
	b	0.05	1.0	–
hyperbolic (11)	k	0.0007–0.0009	0.5	–

a URL, do not have mentions, and do not contain the terms *I, me, my, you,* and *your*. We have 49 topics for this dataset.

A Framework for Assessing Temporal Document Priors. We propose a set of three criteria for assessing temporal document priors. Below, we determine whether the priors meet the criteria.

Performance. A document prior should improve the performance on a set of test queries for a collection of time-aware documents. A well-performing document prior improves on the standard evaluation measures across different collections and across different query sets. We use the *number of improved queries* as well as the *stability of effectiveness* with respect to different evaluation measures as an assessment for performance, where stability refers to that improved or non-decreasing performance over several test collections.

Sensitivity of Parameters. A well-performing document prior is not overly sensitive with respect to parameter selection: the best parameter values for a prior are in a *region* of the parameter space and not a single value.

Efficiency. Query runtime efficiency is of little importance when it comes to distinguishing between document priors: if the parameters are known, all document priors boil down to simple look-ups. We use the *number of parameters* as a way of assessing the efficiency of a prior.

Cognitive Plausibility. We define the cognitive plausibility of a document prior (derived from a retention function) with the goodness of fit in large scale human experiments [2]. This conveys an experimental, but objective, view on cognitive plausibility. We also use a more subjective definition of plausibility in terms of *neurobiological background* and how far the retention function has a biological explanation.

Experiments. To ensure comparability with previous work, we use different models for different datasets. On the news data set, we analyse the effect of different temporal priors on the performance of the baseline, query likelihood with Dirichlet smoothing (D). We optimize parameters for different priors on TREC-6 using grid search. On the Tweets2011 data set, we analyse the effect of different temporal priors incorporated in the query modeling (QM). We do not have a training set and evaluate using leave-one-out cross-validation. Table 3 lists the models whose effectiveness we examine.

We optimize parameters with respect to mean average precision (MAP). MAP, precision at 10 (P@10), R-precision (Rprec) and mean reciprocal rank (MRR) are the quantitative evaluation measures. For the Tweets-2011 collection we do not use the official metric for TREC 2011 (sorting by time and then precision at 30), but the metric to be used for TREC 2012; the previously used metric proved to be sensitive to good cut-off values [26]. The parameter values found are listed in Table 2. For the values based on months, in particular, extended Weibull and MCM-2, we can see that they are in a similar range as the parameters in the literature. We find that using those parameters does not yield very different results from the optimised parameters. We use the Student's t-test to evaluate the significance for all but the small temporal query sets from the news data. We denote significant improvements with $^{\blacktriangle}$ and $^{\vartriangle}$ ($p < 0.01$ and $p < 0.05$, respectively). Likewise, $^{\triangledown}$ and $^{\blacktriangledown}$ denote a decline.

Table 3. Abbreviations of methods and their description

Run id	Description
D	smoothed query likelihood
QM	Query modeling [7]
MCM-1	one store memory chain (5)
MCM-2	two store memory chain (6)
BW	basic Weibull (7)
EW	extended Weibull (8)
AP	amended power (9)
L	linear (10)
HD	hyperbolic discounting (11)

5 Analysis

In this section we seek to understand whether document priors based on retention functions meet the conditions set out above. We examine the retrieval effectiveness of the approaches and then use our framework for assessing the document priors.

Retrieval Effectiveness. We analyze the effectiveness of the priors on the news data, follow-up with the microblog data and conclude with a cross-collection discussion.

News Data. We compare the retrieval performance of our document priors on the TREC-2 and TREC-{7,8} datasets. Table 4 shows the results for the TREC-{7,8} dataset. We observe significant improvements (in terms of MAP and Rprec) for temporal queries using the basic Weibull function (BW) function as a document prior over the baseline without any prior and using MCM-1. We see significant improvements in terms of Rprec using the MCM-2 function, over both the baseline and using MCM-1. There

Table 4. Results on news data, TREC-7 and TREC-8. Significant changes w.r.t. the baseline (D) and the exponential prior (D+MCM-1). The latter is shown in brackets.

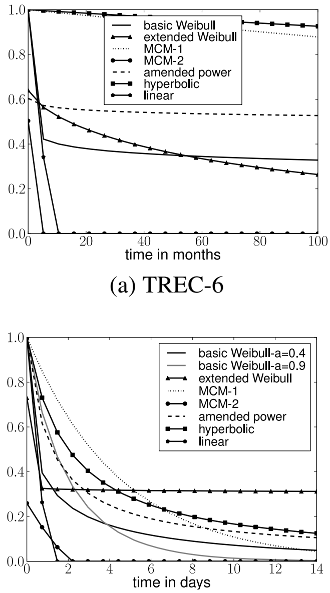
	all queries			temporal queries			non-temporal queries		
Run	MAP	P@10	Rprec	MAP	P@10	Rprec	MAP	P@10	Rprec
D	0.2220	**0.3770**	0.2462	0.2030	**0.3667**	0.2251	0.2281	0.3803	0.2529
D+MCM-1	0.2223	0.3750	0.2473	0.2057$^{\vartriangle}$	0.3625	0.2279	0.2275	0.3789	0.2534
D+MCM-2	0.2253	0.3640$^{\triangledown\,(\triangledown)}$	0.2560	**0.2108**$^{\vartriangle}$	0.3542	**0.2428**$^{\blacktriangle\,(\blacktriangle)}$	0.2299	0.3671$^{\triangledown\,(\triangledown)}$	0.2602
D+BW	**0.2270**	0.3730	0.2603	0.2079$^{\vartriangle\,(\vartriangle)}$	0.3625	0.2339$^{\blacktriangle\,(\vartriangle)}$	**0.2331**	0.3763	0.2687
D+EW	0.2268	0.3720	**0.2611**	0.2086$^{\vartriangle}$	0.3583	0.2346$^{\blacktriangle}$	0.2326	0.3763	**0.2695**
D+AP	0.2222	0.3760	0.2462	0.2032	**0.3667**	0.2251	0.2281	0.3789	0.2528
D+L	0.2157$^{\blacktriangledown}$	0.3740	0.2468	0.1855$^{\blacktriangledown}$	0.3458	0.2123	0.2253	**0.3829**	0.2577
D+HD	0.2224	**0.3770**	0.2462	0.2042	0.3583	0.2261	0.2281	**0.3829**	0.2525

are interesting differences between MCM-1 and MCM-2; first, using MCM-2 yields the worst precision at 10, for temporal and non-temporal queries; second, while using MCM-2 yields the highest MAP for temporal queries, the change is not significant. A per query analysis shows that the changes for MCM-2 are due to changes on very few queries, while for the majority of queries the average precision decreases. Using the basic Weibull function as document prior, however, has very small positive changes for more than half of the queries and, hence, has more stable improvements.

Table 5 shows the results for the TREC-2 data set. The improvements using the temporal priors over the baseline D are not significant. However, functions that work well on the temporal query set (D+MCM-1, D+EW), yield significantly worse performance on the non-temporal set. The only stable performance comes with the use of MCM-1 and basic Weibull.

Using BW as a document prior improves the average precision of few temporal queries, without decreasing the average precision of other temporal queries very much. It improves average precision of the temporal queries without harming non-temporal queries.

Fig. 2a shows the slopes of our document priors. The similarity between MCM-2 and basic Weibull is apparent, both drop to a more or less stable function at the same time. The basic Weibull function, however, features a more gradual change. We also find that the hyperbolic and MCM-1 functions are very similar. The two functions that have a very similar slope to the basic Weibull are the amended power and the extended Weibull, but using them does not change the perfor-



(a) TREC-6

(b) Tweets2011

Fig. 2. The temporal document prior instantiated with parameters optimised on different datasets. Y-axis shows the weight of the prior.

Table 5. Results on news data, TREC-2. Significant differences w.r.t. D+MCM-1.

	all queries			temporal queries			non-temporal queries		
Run	MAP	P@10	Rprec	MAP	P@10	Rprec	MAP	P@10	Rprec
D	0.1983	0.3430	0.2287	0.2719	0.4000	0.2913	**0.1799**	0.3287	0.2130
D + MCM-1	**0.1985**	0.3400	**0.2289**	0.2730	0.4050	0.2937	**0.1799**	0.3238▽	0.2127
D + MCM-2	0.1961	0.3330	0.2240▽	0.2731	**0.4150**	**0.2952**	0.1769▼	0.3125▼	0.2063▼
D + BW	0.1984	0.3420	0.2287	0.2727	0.4050	0.2915	0.1798	0.3263	0.2130
D + EW	0.1983	0.3400	0.2277	**0.2749**	**0.4150**	0.2927	0.1792	0.3213▽	0.2114
D + AP	0.1983	**0.3430**	0.2283	0.2717	0.4050	0.2915	0.1799	**0.3275**	0.2125
D + L	0.1961▽	0.3410	0.2288	0.2671	0.3950	0.2902	0.1783▽	**0.3275**	**0.2135**
D + HD	0.1984	0.3410	0.2284	0.2730	0.4050	0.2915	0.1798	0.3250	0.2127

mance much. The main difference between the slope of the functions and basic Weibull is close to 0: the steeper the function at the beginning, the better the performance.

Fig. 3 shows the temporal distribution of the top 100 retrieved documents for different approaches on the TREC-{7,8} test set. The topmost distribution shows the distribution for all relevant documents, which has only very few documents. The baseline, D, ranks older documents high. Using a linear retention function as document prior (D+L), the system retrieves even more old documents and fewer recent documents and it does not outperform the baseline for queries with recent documents. The distribution for D+MCM2 is the opposite and performs well for very recent queries, while D+MCM1 and D+BW reduce the number of old retrieved documents.

Microblog Data. We compare the retrieval performance of the different priors on the Tweets2011 dataset. Table 6 shows the results for the Tweets2011 dataset. Query modeling (QM) with the MCM-1 function does not yield significant improvements. QM with basic Weibull (BW), amended power (AP), linear (L) and hyperbolic discounting (HD) does yield significant improvements in MRR over the baseline QM. The increase is up to 15% for AP and BW. MAP improves as well, but not significantly. Filtering improves the results for all approaches and while MRR increases by over 7%, this is not significant. There are similar effects on the filtered results: the prior does not act a filter.

When we perform a query analysis of the differences between QM and QM+BW, we see that, in the unfiltered condition, QM+BW outperforms QM on 17 (out of 49) queries, while QM outperforms QM+BW on 6 queries; for the filtered condition, the numbers are 11 and 6, respectively. The comparisons are similar for the other functions.

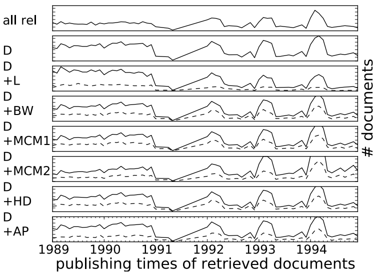


Fig. 3. Distribution of retrieved (cut-off: 100) documents. The solid line is the distribution for all, the dashed line for documents retrieved for improved queries.

Fig. 2b shows the slope of the different functions for the optimized parameters.

The functions that help significantly are the functions that share the same rapid decrease on the first day with a continuous, slower, decrease on the second and third day. For the other functions, on the one hand MCM-2 decreases similarly on the first day, but not on the following days: QM+MCM-2 even decreases the MAP and P@10. MCM-1 decreases slowly and continues to decrease. The changes in performance with respect to the metrics used are therefore not as visible as, e.g., using QM-HD: here, the slope of HD decreases similarly to MCM-1, but then settles, while MCM-1 continues to fall. Queries for which the HD function increases average precision are queries submitted in the second week of the collection period with more days of tweets to return and to process. QM+BW and QM+AP display significant increases in MRR, but neither of them decreases MAP and P@10; the two models have a very similar slope.

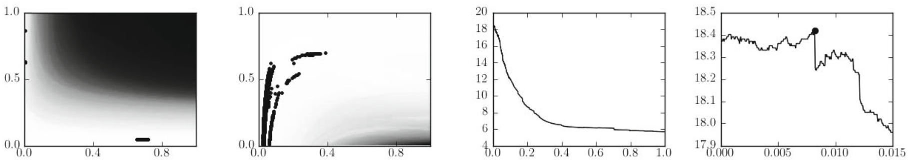
Table 6. Results on microblog data, Tweets2011

	unfiltered			filtered		
Run	MAP	P@10	MRR	MAP	P@10	MRR
QL	0.2731	0.3898	0.6133	0.2873	**0.5408**	0.7264
QM	0.2965	0.4061	0.6624	**0.3140**	0.5367	0.7559
QM+MCM-1	0.3101	0.4143	0.7682	0.3062	0.5306	0.7944
QM+MCM-2	0.2903	0.4102	0.7192	0.2912	0.5265	0.7675
QM+BW	0.3058	**0.4286**	**0.7801**$^{\triangle}$	0.3057	**0.5408**	0.7971
QM+EW	0.3038	0.4224	0.7251	0.3024	0.5224	0.7644
QM+AP	0.3100	0.4327	**0.7801**$^{\triangle}$	0.3103	**0.5408**	0.8046
QM+L	**0.3129**	0.4245	0.7700$^{\triangle}$	0.3082	0.5286	**0.8144**
QM+HD	0.3080	0.4286	0.7698$^{\triangle}$	0.3081	**0.5408**	0.7944

Assessing the Document Priors. We step back to assess the temporal document priors based on the framework introduced in §4.

Performance. Using the BW retention function as prior performs significantly better, better, or similar to MCM-1 over three data sets. Other retention functions either do not show significant improvements or improve on one subset while decreasing on others. BW, EW, and HD improve the greatest number of queries over MCM-1.

Parameter Sensitivity. We first examine parameter sensitivity on news data. Fig. 4 shows heatmaps for the different functions for parameter optimisation TREC-6. Fig. 4d shows that D+ MCM-1 is very unstable with respect to the optimal value for r, especially when we look at the surrounding parameters. D+BW and D+AP have more optimal points and are more stable with respect to those points. We observe similar effects for D+EW. When we examine parameter sensitivity on Tweets2011, we look at the optimal parameters selected for each fold in a cross-validation. We find stable



(a) Basic Weibull (b) Amended Power (c) MCM-1 (d) MCM-1, subset

Fig. 4. Optimisation of parameters on MAP on TREC-6. The lighter the color, the higher the MAP. Black dots indicate the parameter combination with highest MAP.

Table 7. Assessing temporal document priors; # improved queries is w.r.t. MCM-1

Condition	MCM-1	MCM-2	BW	EW	AP	L	HD
# impr. queries (temp.)	n/a	14 (58%)	5 (20%)	16 (67%)	5 (20%)	2 (8%)	6 (25%)
# impr. queries (non-temp.)	n/a	27 (35%)	35 (46%)	26 (34%)	38 (50%)	36 (47 %)	33 (43%)
# impr. queries (Tweets2011)	n/a	16 (32%)	17 (34%)	22 (44%)	0 (0%)	17 (34 %)	21 (42%)
MAP	+	−	+	0	0	−	0
P10	−	−	0	−	0	0	0
Rprec	0	±	+	±	0	0	0
MRR	0	0	+	0	+	+	+
Sensitivity of parameters	−	−	+	−	+	+	+
Efficiency: # parameters	2	4	2	4	3	2	1
Plausibility: fits human behav.	+	++	+	++	+	n/a	n/a
Plausibility: neurobiol. expl.	+	+	−	+	−	−	−

parameters for all priors but the Weibull function. The Weibull function fluctuates mildly between 0.3 and 0.4, with one exception being 0.9 (Fig. 2b).

Efficiency. The only difference in efficiency between the priors is the number of parameters needed for prior optimization. A sweep for four parameters (for MCM-2 and EW) is feasible but time-consuming: ideally, the minimal number of parameters (MCM-1, BW, L, and HD) should be optimized.

Cognitive Plausibility. Previous work [2] fitted retention functions to how participants remember news (see Fig. 1). They report that the MCM-2 and EW functions fit best while MCM-1, as a less general case of MCM-2, obviously fits worse. The AP retention function does not fit well enough to be more than an approximation [5]. The linear and hyperbolic discounting function have so far not been fitted on retention data. Table 7 summarizes how the priors fulfill the requirements listed in §4. Priors using the BW, AP, and HD retention functions show stable performance across collections, on a query level as well as on a general level, with BW performing well and being stable. All three functions have a stable parameter selection process for at least the news dataset. AP with three parameters is too inefficient, while BW and HD with two and one parameter converge to a result much faster. We know that BW has a neurobiological explanation and fits humans fairly well. The exponential function (MCM-1) as prior does not fulfill the requirements as well as other functions. This prior does have good results, but is not particularly stable when it comes to parameter optimization; significant results from the news data set do not carry over to the microblog data set. In sum, we propose to use the basic Weibull retention function for temporal document priors.

6 Conclusion

We have proposed a new perspective on functions used for temporal document priors used for retrieving recent documents. We showed how functions with a cognitive motivation yield similar, if not significantly better results than others on news and microblog datasets. In particular, the Weibull function is stable, easy to optimize, and motivated by psychological experiments. For future work we propose to analyze the effect of using temporal functions in more retrieval models, in particular in adaptive query models.

Acknowledgments. We thank Jessika Reissland for her inspiration. This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 727.-011.005, 612.001.116, HOR-11-10, the Center for Creation, Content and Technology (CCCT), the BILAND project funded by the CLARIN-nl program, the Dutch national program COMMIT, by the ESF Research Network Program ELIAS, and the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences.

References

1. Porter, S., Birt, A.R.: Is traumatic memory special? Appl. Cogn. Psych. 15, 101–117 (2001)
2. Meeter, M., Murre, J.M.J., Janssen, S.M.J.: Remembering the news: modeling retention data from a study with 14,000 participants. Memory & Cognition 33, 793–810 (2005)
3. Hertwig, R., et al.: Fluency heuristic: a model of how the mind exploits a by-product of information retrieval. J. Exp. Psych.: Learning, Memory, and Cogn. 34, 1191–1206 (2008)
4. Chessa, A.G., Murre, J.M.: A memory model for internet hits after media exposure. Physica A Statistical Mechanics and its Applications (2004)
5. Chessa, A.G., Murre, J.M.: Modelling memory processes and internet response times: Weibull or power-law? A Statistical Mechanics and its Applications (2006)
6. Li, X., Croft, W.B.: Time-Based Language Models. In: CIKM 2003 (2003)
7. Massoudi, K., Tsagkias, M., de Rijke, M., Weerkamp, W.: Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 362–367. Springer, Heidelberg (2011)
8. Efron, M., Organisciak, P., Fenlon, K.: Improving retrieval of short texts through document expansion. In: SIGIR 2012 (2012)
9. Efron, M., Golovchinsky, G.: Estimation Methods for Ranking Recent Information. In: SIGIR 2011 (2011)
10. Ebbinghaus, H.: Memory: a contribution to experimental psychology. Teachers College, Columbia University (1913)
11. Schooler, L.J., Anderson, J.R.: The role of process in the rational analysis of memory. Cognitive Psychology 32, 219–250 (1997)
12. Rubin, D.C., Hinton, S., Wenzel, A.: The precise time course of retention. Journal of Experimental Psychology: Learning, Memory, and Cognition 25, 1161–1176 (1999)
13. Wickens, T.D.: Measuring the time course of retention. On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson–Shiffrin model (1999)
14. Heathcote, A., Brown, S., Mewhort, D.J.: The power law repealed: the case for an exponential law of practice. Psychonomic Bulletin & Review 7, 185–207 (2000)
15. Alonso, O., Strötgen, J., Baeza-Yates, R., Gertz, M.: Temporal Information Retrieval: Challenges and Opportunities. In: TWAW 2011, pp. 1–8 (2011)

16. Verhagen, M., Pustejovsky, J.: Temporal processing with the TARSQI toolkit. In: COLING 2008 (2008)
17. Odijk, D., de Rooij, O., Peetz, M.-H., Pieters, T., de Rijke, M., Snelders, S.: Semantic Document Selection. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) TPDL 2012. LNCS, vol. 7489, pp. 215–221. Springer, Heidelberg (2012)
18. Keikha, M., Gerani, S., Crestani, F.: Time-based relevance models. In: SIGIR 2011 (2011)
19. Amodeo, G., Amati, G., Gambosi, G.: On relevance, time and query expansion. In: CIKM 2011. ACM (2011)
20. Dakka, W., Gravano, L., Ipeirotis, P.G.: Answering General Time Sensitive Queries. In: CIKM 2008, pp. 1437–1438 (2008)
21. Peetz, M.-H., Meij, E., de Rijke, M., Weerkamp, W.: Adaptive Temporal Query Modeling. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 455–458. Springer, Heidelberg (2012)
22. Efron, M.: Query-specific recency ranking. In: SIGIR 2012 Workshop on Time-aware Information Access (2012)
23. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: SIGIR 1998 (1998)
24. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
25. Ainslie, G., Haslam, N.: Hyperbolic discounting. In: Choice over time. Russell Sage Foundation (1992)
26. Amati, G., et al.: FUB, IASI-CNR, UNIVAQ at TREC 2011. In: TREC 2011, NIST (2011)