Information retrieval using temporally smoothed language models

Craig Willis, Miles Efron
Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign
{willis8, mefron}@illinois.edu

ABSTRACT

This paper presents a novel approach to utilizing temporal information to improve the effectiveness of ad hoc information retrieval (IR). While temporally-informed IR is an active research area, previous work has focused mainly on improving ranking by analysis of the timestamps associated with retrieved documents. In contrast, we integrate time more directly into retrieval via the smoothing of language models. In our approach, document models are smoothed using *temporal language models*–language models corresponding to a period in time (e.g., day, week, or month) in a collection. We present three temporal smoothing methods and evaluate them using seven standard TREC collections. We show that document smoothing using temporal language models improves retrieval effectiveness over state-of-the art approaches. We demonstrate that this method complements established approaches that operate on document timestamp distributions.

1. INTRODUCTION

Temporality is an intrinsic characteristic of information systems. Documents are published at particular points in time, users issue queries over time, collections have temporal constraints, languages and writing styles even change over time. Temporal constraints, such as date ranges and sort-by options, have long been standard features of search interfaces for decades. However, even today general-purpose retrieval algorithms do not directly incorporate temporal information.

Over the past decade, researchers have explored the role of temporal information in a variety of retrieval contexts. Studies have explored techniques for extracting temporal expressions [2, 1, 17, 10], studied the temporal dynamics of language [12], documents [18], queries [7, 16], and users' behaviors [15]. Numerous models have been proposed incorporating temporal information into the retrieval process, generally relying on the analysis of timestamps associated with documents [9, 3, 5, 6, 13]. Closer to the work presented in this paper, [4] used language models for temporal intervals to date historical texts.

In this paper we study how temporal information can be directly incorporated into the language modeling framework through the use of *temporal language models*. Similar to [4], temporal language models are language models corresponding to a period in time (e.g., day, week, or month) in a document collection. We propose that temporal language models can be used to smooth document language models, improving model estimation and retrieval effectiveness. This work can be seen as building on the temporal retrieval models of [9, 3, 5, 6] and the cluster- or LDA-based smoothing methods proposed by [11] and [19].

To our knowledge, temporal language models have never been used in this way. Earlier work in [4] used temporal language models to predict the date of historical documents.

2. RELATED WORK

2.1 Temporal information retrieval

Li and Croft [9] were the first to suggest that documents can be scored based on a combination of lexical and temporal information, which they refer to as "time-based language models". Working within the language modeling framework [14], they incorporate temporal information via an exponential prior modeling the relevancy of documents based on timestamps. This later became known as a "recency model" because it heavily favors recent documents.

This basic model was later extended by Dakka, Gravano, and Ipeirtos [3] to support any time-sensitive query. After an initial retrieval, a histogram is created from document timestamps and bins are re-ordered by descending magnitude. An exponential distribution is used to model the importance of documents in each bin, with a global rate parameter estimated for the collection. Efron, and Golovchinsky [5] further refine this model to allow for query-specific estimation of the exponential rate parameter. Finally, Efron, Lin, He and de Vries [6] replace the histogram approach with a kernel-density estimator (KDE) to estimate the query-specific temporal distribution of results. Each of these methods depends on an initial retrieval to estimate the distribution of document timestamps.

In each of these models, temporal information is assumed to be independent of the lexical information. Standard query likelihood scores are combined with the temporal scores based on this assumption.

Efron and Golovchinsky [5] also explore a novel model using temporal information to smooth the document language model. In this case, the smoothing parameter λ_t in Jelinek-Mercer smoothing is learned based on the temporal distribution of results. The amount of smoothing for each document varies depending on the time of each document d.

2.2 Temporally-estimated relevance models

Each of the above models is also considered in the context of relevance models [8]. In each case, the temporal models are used to score documents in an initial retrieval, the results of which used to construct relevance models. Only [6] re-scores the final results using the temporal model. Peetz, Meij, and Rijke [13] propose an alternative query modeling approach that relies on the identification of temporal "bursts" in pseudo-feedback.

2.3 Document-specific smoothing

As noted by Zhai [20], the usual approach of smoothing all documents with the same collection language model seems suboptimal. Document-dependent smoothing methods, such as those proposed in [11] and [19], instead smooth the document language model with document-dependent models constructed using methods such as clustering or topic modeling. Liu and Croft [11] use k-means to cluster documents and create cluster language models. Wei and Croft [19] use latent Dirichlet allocation (LDA) to calculate topic models. Both studies use a two-phase smoothing approach. The cluster or topic language models are smoothed using the collection language model. The resulting model is then used to smooth the document language model. Either Jelinek-Mercer or Dirichlet smoothing can be used in either phase, but both studies reported results using Jelinek-Mercer for the first phase and Dirichlet for the second. This general approach is adopted in the models described below.

3. TEMPORAL LANGUAGE MODELS

We define a *temporal language model* as a unigram model of terms in a collection at time or interval t. We can imagine that the language model for a collection for the year 1900 is different than a language model for 1950. In a stream of news, the language model for 1995 might be different than the language model for 2000. From this, we can define a generative model of either queries or documents. Given some interval t, the probability that the interval generated the query (or document) can be estimated as $p(q|\theta_t)$. The interval could be an hour, day, week, or year, depending on the characteristics of the collection.

3.1 Model 1: Independent evidence

Following [3] we can directly score documents using the document and temporal language models treated independently:

$$p_{TD}(d|q) \propto p(q|\theta_d) \cdot p(d) \cdot p(q|\theta_t) \cdot p(t)$$

Where θ_d is the document language model and θ_t is the temporal language model. We refer to this as Model 1.

3.2 Model 2: Smoothing using the temporal language model based on document time

Following [11] and [19], we can smooth the document language model using the temporal language model and a two-phase smoothing approach:

$$p_{TSM}(w|d) = \frac{n(w,D) + \mu \cdot [(1-\alpha) \cdot p(w|\theta_t) + \alpha \cdot p(w|\theta_c)]}{n(D) + \mu}$$

Where θ_t is the temporal language model for the document time d_t and θ_c is the collection language model. Under this model, temporal evidence is no longer treated as independent. The two-stage smoothing approach can use either Jelinek-Mercer or Dirichlet smoothing for either stage. In their study of smoothing methods, [21] found that Dirichlet smoothing was more effective for explaining unseen words while Jelinek-Mercer was more effective for explaining noise in the query for longer queries.

3.3 Model 3: Smoothing using the most likely temporal language model

Loosening the restriction that a document was generated based on the language model associated with publication time, we can instead smooth using the temporal language model that is most likely to have generated the document. For each document, the temporal language model used for smoothing is selected using Kullback-Liebler divergence $KL(\theta_d||\theta_t)$.

$$p_{BMN}(w|d) = \frac{n(w,D) + \mu \cdot [(1-\alpha) \cdot p(w|\pi_t) + \alpha \cdot p(w|\theta_c)]}{n(D) + \mu}$$

Where π_t is the temporal language model with the smallest KL-divergence from the document language model:

$$p(\pi_i|D) = f(KL(\pi_i|\theta_d))$$

3.4 Model 4: Smoothing using weighted average of all temporal language models

Instead of smoothing by a single language model, following the approach used by [19], we can smooth the document using the weighed combination of all temporal language models.

$$p_{TSA}(w|D) = \sum_{t=1}^{K} p(w|\theta_t) p(\theta_t|D)$$

Where the $p(\theta_t|D) \sim KL(\theta_c||\theta_t)$.

3.5 Estimating temporal language models

Adopting a unigram multinomial language model for θ_t, the next step is to estimate $p(w|\theta_t)$. The simplest approach is to use the maximum likelihood estimator:

$$p_{MLE}(w|\theta_t) = \frac{n(w,t)}{n(t)}$$

Where t is the temporal interval. In our experiments we found the MLE model to be too similar too the collection model, dampening the effect of the available temporal information. We explored two different approaches to enhancing the temporal language model.

First, we instead used the normalized pointwise mutual information:

$$npmi(w;t) = \frac{\log \frac{p(w,t)}{p(w)p(t)}}{-\log[p(w,t)]}$$

The resulting value ranges from -1 to 1. For estimating $p(w|t)$, all values less than zero were treated as 0:

$$p_{NPMI}(w|t) = \frac{npmi(w;t)}{\sum_{w \in W} npmi(w;t)}$$

Second, we used the likelihood ratio test to restrict the terms allowed in the model θ_t. Setting $\alpha = 0.01$, only terms w found to be significantly associated with time t are added to the model. This is the final approach used in our reported results.

4. EXPERIMENTS AND RESULTS

4.1 Data

Experiments were conducted using seven TREC test collections. The first five collections were included for comparability to [11] and [19]. These include the Associated Press Newswire (AP) 1988-90 with queries 51-150; Financial Times (FT) 1991-94 with queries 301-400; Los Angeles Times (LA) with queries 301-400; San Jose Mercury News (SJM) 1991 with queries 51-150; and Wall Street Journal (WSJ) 1987-92 with queries 51-100 and 151-200. Additional experiments were conducted using the Tweets2011 test collection with queries 1-49 and 50-110. Queries were constructed using only the "title" field of the TREC topics. The standard relevance judgments were used. Queries with no documents in the judged pool for a collection were removed. Information about each collection is presented in Table 1. All documents and queries are stemmed using the Krovetz [] stemmer and stopwords removed.

[Describe Tweets2011 corpus...]

Table 1: Test collection details.

Collection	# docs	Queries	# queries
AP	242,918	51-150	99
FT	210,158	301-400	95
SJM	90,257	51-150	94
LA	131,896	301-400	98
WSJ	173,252	51-100, 151-200	100
Tweets2011	11,908,899	1-49 50-110	110

4.2 Parameter estimation

Following [11] and [19], the AP collection is used as the training collection to estimate parameters that are then tested on the FT, LA, SJM, and WSJ collections. The Tweets2011 collection (queries 1-49) is used to estimate parameters that are then tested on queries 50-110. Models are optimized for mean average precision (MAP) using an exhaustive search.

4.2.1 Temporal interval

An important parameter in the definition of temporal language models is the temporal interval. Using the AP and Tweets 2011 training collections, we explored hourly, daily, weekly, bi-weekly, and monthly intervals. All documents with timestamps in each interval are candidates for inclusion in the associated temporal language model.

[Chart performance with different temporal intervals.]

4.2.2 LDA-based language models

For the LDA-based language models, we attempted to repeat the experiment described in [19] using off-the-shelf software, specifically Mallet 2.0.7. The Mallet LDA implementation has several parameters including the number of topics, number of sampling iterations, and burn-in. For the number of topics, we used 800 as identified by Wei and Croft. Because the implementations differ significantly, we could use the number of iterations and chains specified in [19]. We explored several combinations of iterations/burn-in including 500/100, the Mallet default 1000/200, and 5000/1000. We found the effectiveness and speed of the Mallet default values to be adequate for this study [1] We accept the original paper's values for the smoothing parameters $\lambda = 0.07$ and $\mu = 1000$.

4.2.3 Temporal smoothing parameters

For each of the temporal models, we must estimate the mixing parameter λ. As noted above, the AP and Tweets2011 collections are used for training. Separate values are estimated for both feedback λ_{RM3} and non-feedback λ_{QL} scenarios.

Table 2

Table 2: Parameter estimates.

Model	AP	Tweets2011
KDE	$\alpha_{QL} = 0.35, \alpha_{RM3} = 0.15$	$\alpha_{QL} = 0.45, \alpha_{RM3} = 0.05$
TSM	$\lambda_{QL} = 0.05, \lambda_{RM3} = 0.05$	$\lambda_{QL} = 0.25, \lambda_{RM3} = 0.05$
BMN	$\lambda_{QL} = 0.05, \lambda_{RM3} = 0.15$	$\lambda_{QL} = 0.05, \lambda_{RM3} = 0.05$
TSA	$\lambda_{QL} = 0.05, \lambda_{RM3} = 0.15$	$\lambda_{QL} = 0.15, \lambda_{RM3} = 0.10$
LDA	$\lambda = 0.7$	$\lambda = 0.7$

5. RESULTS

Results are presented in Table 4 and 5 and 6 and 7.

Table 3: Symbols indicating statistically significant changes.

Symbol	Description
▲	$p < 0.01$ improvement against the QL baseline
▼	$p < 0.01$ degradation against the QL baseline
△	$p < 0.05$ improvement against the QL baseline
▽	$p < 0.05$ degradation against the QL baseline
↑	$p < 0.10$ improvement against the QL baseline
↓	$p < 0.10$ degradation against the QL baseline

5.1 Conclusion and future work

6. DISCUSSION

7. REFERENCES

[1] I. Arikan, S. Bedathur, and K. Berberich. Time will tell: Leveraging temporal expressions in ir. In *In WSDM*, 2009.

[1]We did find that increasing the number of iterations improved effectiveness, sometimes significantly. However, since LDA-based language models are not the focus of this study, we accepted the default values.

Table 4: Comparison of query likelihood (QL) with temporal models.

Model	AP	FT	LA	SJM	WSJ	TWEETS2011	TWEETS2012
KL	0.1885	0.2571	0.2373	0.1951	0.2724	0.2503	0.2049
KDE	0.1894	0.2539 ↓	0.2393	0.1939	0.2722	0.2623 ▲	0.2146 ▲
TSM	0.1884	0.2563	0.2360 ▽	0.1941 ↓	0.2717 ↓	0.2334	0.1580 ▼
BMN	0.1883	0.2591	0.2388	0.1901 ▽	0.2712	0.2000 ▼	0.1565 ▼
TSA	0.1889	0.2589 ▲	0.2395 △	0.1948	0.2726	0.2229 ▼	0.1692 ▼
LDA	0.2155 ▲	0.2666	0.2572 ↑	0.2157 ▲	0.2977 ▲	0.2436	0.1957

Table 5: Comparison of relevance models (RM3) with temporal models.

Model	AP	FT	LA	SJM	WSJ	TWEETS2011	TWEETS2012
RM3	0.2098	0.2684	0.2485	0.2145	0.2975	0.2895	0.2406
KDE	0.2102	0.2680	0.2477	0.2137	0.2954 ▽	0.2890	0.2427
TSM	0.1884 ▼	0.2563 ▼	0.2360 ▼	0.1941 ▼	0.2717 ▼	0.2370 ▽	0.1583 ▼
BMN	0.2102	0.2717	0.2524 ↑	0.2092 ▽	0.2970	0.2167 ▼	0.1707 ▼
TSA	0.2123 ↑	0.2721 ↑	0.2551 △	0.2172	0.2971	0.2448 ▼	0.1896 ▼
LDA	0.2349 ▲	0.2650	0.2721 ↑	0.2299 △	0.3234 ▲	0.2588 ▽	0.2139 ▽

[2] L. Childs and D. Cassel. Extracting and normalizing temporal expressions. *Proceedings of ACL Workshop*, pages 51–56, 1999.

[3] W. Dakka, L. Gravano, and P. Ipeirotis. Answering General Time-Sensitive Queries. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):220–235, Feb. 2012.

[4] F. de Jong, H. Rode, and D. Hiemstra. Temporal language models for the disclosure of historical text. pages 1–9, 2005.

[5] M. Efron and G. Golovchinsky. Estimation methods for ranking recent information. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, page 495, 2011.

[6] M. Efron, J. Lin, J. He, and A. D. Vries. Temporal Feedback for Tweet Search with Non-Parametric Density Estimation. *umiacs.umd.edu*, pages 33–42, 2014.

[7] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Transactions on Information Systems*, 25(3):14–es, July 2007.

[8] V. Lavrenko and W. B. Croft. Relevance based language models. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01*, pages 120–127, 2001.

[9] X. Li and W. B. Croft. Time-based language models. *Proceedings of the twelfth international conference on Information and knowledge management - CIKM '03*, page 469, 2003.

[10] S. Lin, P. Jin, X. Zhao, and L. Yue. Exploiting temporal information in Web search. *Expert Systems with Applications*, 41(2):331–341, Feb. 2014.

[11] X. Liu and W. B. Croft. Cluster-based retrieval using language models. *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, page 186, 2004.

[12] J.-b. Michel, Y. K. Shen, A. P. Aiden, A. Veres, K. Matthew, T. Google, B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. (March), 2011.

[13] M.-H. Peetz, E. Meij, and M. Rijke. Using temporal bursts for query modeling. *Information Retrieval*, 17(1):74–108, July 2013.

[14] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281, 1998.

[15] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. *Proceedings of the 21st international conference on World Wide Web - WWW '12*, page 599, 2012.

[16] M. Shokouhi. Detecting Seasonal Queries by Time-Series Analysis Categories and Subject Descriptors. In *SIGIR 11*, pages 1171–1172, 2011.

[17] J. Strötgen, O. Alonso, and M. Gertz. Identification of top relevant temporal expressions in documents. In *Proceedings of the 2nd Temporal Web . . .*, pages 33–40, 2012.

[18] J. Teevan, S. T. Dumais, and J. L. Elsas. The Web Changes Everything : Understanding the Dynamics of Web Content. 2009.

[19] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, page 178, 2006.

[20] C. Zhai. *Statistical Language Models for Information Retrieval*. Morgan & Claypool, Jan. 2008.

[21] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, Apr. 2004.

Table 6: Number of queries with increased/decreased MAP over baseline QL. Values in parenthesis are number of queries with increased/decrease MAP > 0.05.

Model	AP	FT	LATIMES	SJM	WSJ	TWEETS2011	TWEETS2012
KDE	52/44 (0/0)	34/46 (1/3)	42/45 (2/1)	38/45 (0/0)	42/52 (1/0)	33/16 (5/1)	36/23 (5/1)
TSM	26/63 (1/0)	24/51 (0/0)	23/59 (0/0)	26/58 (0/0)	36/59 (0/0)	26/22 (10/12)	20/40 (4/21)
BMN	44/52 (1/0)	28/42 (2/0)	31/50 (3/0)	27/59 (1/6)	29/64 (1/0)	11/32 (0/18)	7/47 (1/23)
TSA	46/48 (0/0)	40/22 (1/0)	40/35 (0/0)	39/46 (0/0)	35/56 (0/0)	12/31 (2/12)	8/46 (1/17)

Table 7: Number of queries with increased/decreased MAP over baseline RM3. Values in parenthesis are number of queries with increased/decrease MAP > 0.05.

Model	AP	FT	LATIMES	SJM	WSJ	TWEETS2011	TWEETS2012
KDE	44/50 (0/0)	33/46 (2/0)	32/53 (0/0)	34/54 (1/0)	38/59 (0/0)	23/20 (0/1)	30/28 (2/0)
TSM	20/77 (0/12)	22/67 (3/11)	32/63 (3/13)	22/71 (2/17)	15/86 (1/20)	17/33 (5/20)	14/46 (3/29)
BMN	52/44 (0/0)	42/39 (2/2)	43/45 (4/1)	30/61 (1/5)	33/63 (1/0)	12/38 (3/23)	13/47 (2/31)
TSA	53/45 (2/1)	53/26 (2/1)	44/45 (7/2)	45/47 (5/4)	51/50 (0/1)	14/33 (1/17)	13/46 (1/24)