Time-Based Language Models

Xiaoyan Li and W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
{xiaoyan,croft}@cs.umass.edu

ABSTRACT

We explore the relationship between time and relevance using TREC ad-hoc queries. Two types of queries are classified as time-based: one favors very recent documents and the other has more relevant documents within a specific period in the past. We propose a time-based language model approach to retrieval for these queries. We show how time can be incorporated into both query-likelihood models and relevance models. The experiments on TREC title queries show time-based language models outperforming baseline language model approaches on both types of time queries.

KEYWORDS

Information retrieval, language models, relevance models, time-based language models, time-based queries

1. INTRODUCTION

The task of information retrieval is to retrieve relevant documents that satisfy the user's information need. Relevance is an abstract measure of how well a document satisfies the user's information need, which is approximated by a query. In the process of approximation, a time-related information need is usually not captured by the query. For example, an old document, which is topically relevant to the query, may not satisfy the information need if the user is only interested in more recent documents. Many news-related queries would fall into this category. It is also possible that a recent document that appears to be topically relevant may not satisfy the user's information need if the user is only interested in documents within a specific period in the past. For example, the query "star wars" could have most of the relevant documents in the Reagan era rather than in recent documents. Most document retrieval systems built for the corporate environment recognize the importance of time and have provided, for many years, default rankings based on recency as well as the ability to specify a time period as a query attribute or field. The problem with these systems is that they are either based on a Boolean retrieval model or the time attribute is combined in a heuristic manner with the document scores to produce a final ranking.

In this paper, we introduce the time-based language model approach that incorporates time as part of the retrieval model. Time-based language models are a simple extension of the language model approaches to retrieval that have been developed over the past few years (e.g. [1-6]). Instead of assuming uniform prior probabilities in these retrieval models, we assign document priors based on creation dates.

In the next section, we explore the relationship between time and relevance on TREC ad-hoc title queries, and identify two types of queries for evaluating the proposed models. The first type of query favors very recent documents and the other has more relevant documents within a specific period in the past. Section 3 describes the time-based language model approaches to retrieval. Section 4 gives the experimental design and results. The experimental results show that time-based language models outperform baseline language model approaches. Related research is discussed in section 5, and the conclusion in section 6 discusses future research directions.

2. TIME AND RELEVANCE

In this section, we explore the relationship between time and relevance based on an analysis of TREC ad-hoc queries. The first part shows the average distribution over time for the TREC relevant documents. The second part highlights the differences between individual queries with respect to time sensitivity.

2.1 Overall Distribution for Different Query Sets

We plotted an overall distribution graph for each of the eight TREC query sets (i.e. q1-50, q51-100, q101-150, q151-200, q201-250, q251-300, q301-350 and q351-400) with the x axis representing time in months (in the past) and the y axis representing the percentage of total relevant documents. The origin corresponds to the most recent date in all the TREC collections. (See Figures 2.1 - 2.8). These averages are affected by a number of factors, such as when the collections were introduced, and which collections were used in a given year, but some trends can be seen.

Figures 2.1 - 2.3 show that, on average, relevant documents are distributed quite evenly across the time line in the period [30, 90]. Figure 2.4 shows that, for these queries, there are more relevant documents in the period [60, 80] on average. Figure 2.5 shows that there are some relevant documents in the period [32, 60] and

some in the period [72, 82]. Figure 2.6 shows a sharp increase in relevant documents in the period [72, 82], although there are relevant documents in the entire time period.. Figure 2.7 and Figure 2.8 show that, for these queries, there are substantially more relevant documents in the recent past.

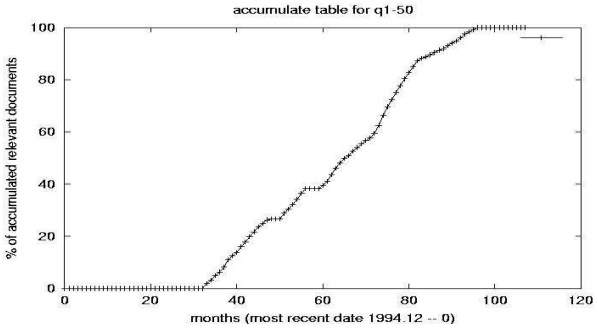

Figure 2.1

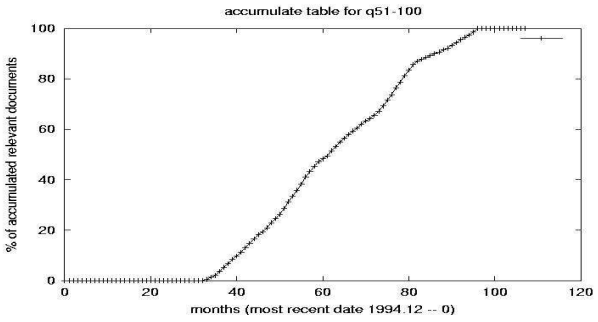

Figure 2.2

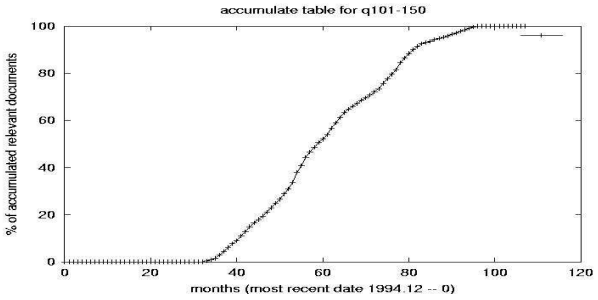

Figure 2.3

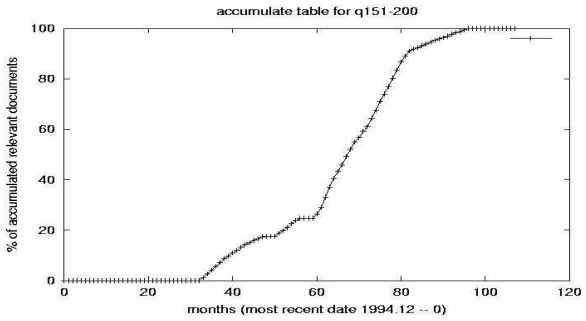

Figure 2.4

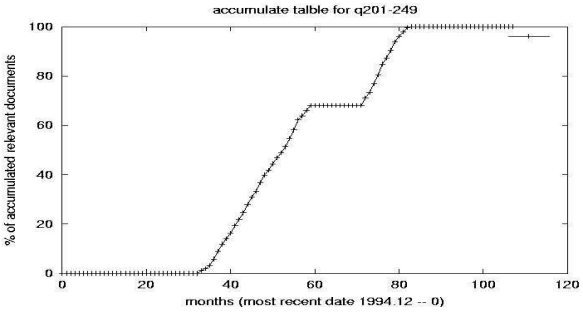

Figure 2.5

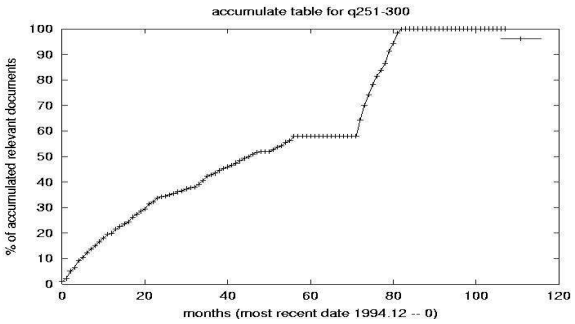

Figure 2.6

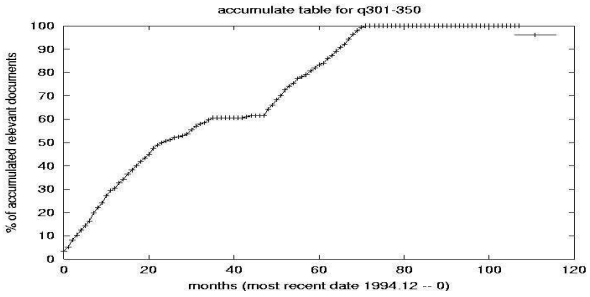

Figure 2.7

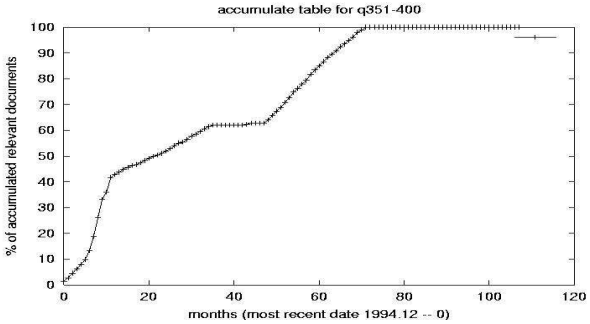

Figure 2.8

2.2 Examples of Different Types of Queries.

Individual queries can show much more time sensitivity than the averages. As we mentioned previously, there are two main types of queries that do not have a "uniform" distribution of relevant documents over time (there are actually many types of distributions but these are more common). The first type of query favors very recent documents and the other has more relevant documents within a specific period in the past. Query 301 is an example of the first type of query. (See figure 2.9). Query 156 is an example of the second type of query, which has more relevant documents within a particular period in the past. (See figure 2.10) Query 165 is an example of a query that has a more uniform distribution of relevant documents along the time line. (See figure 2.11). This group is the most numerous, but there are still a significant number of examples of the first two types. In the 400 TREC queries, there were approximately 80 queries each of these query types. For the experiments described in section 4, we used 50 queries of the first type and 10 of the second.

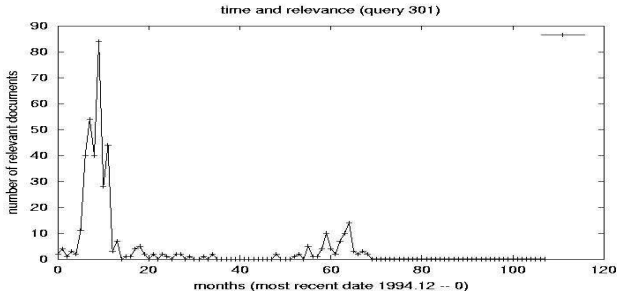

Figure 2.9

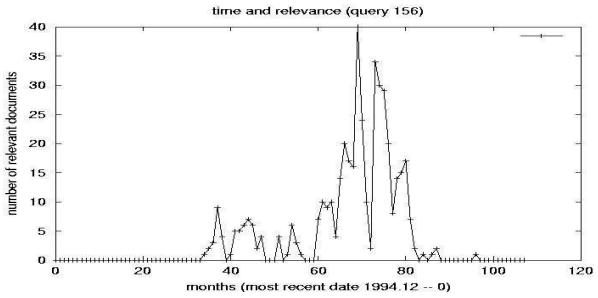

Figure 2.10

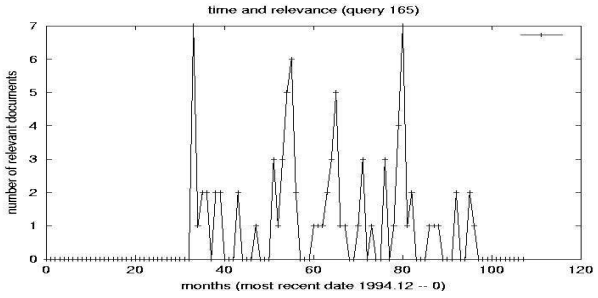

Figure 2.11

3. LANGUAGE MODELS FOR RETRIEVAL

3.1 Query Likelihood Models

Language modeling frameworks were introduced to information retrieval by Ponte and Croft [1], followed by some variations [2,3,4,5] that adopted a similar framework. In the language modeling framework, there are basically three approaches to ranking documents: the query likelihood model, the document likelihood model and comparing query and document language models directly. In the simplest case, the posterior probability of a document given in (3.1) is used to rank the documents in the collection.

$$p(d/q) \propto p(q/d)\,p(d) \tag{3.1}$$

The prior probability of the document $p(d)$ is usually assumed to be uniform and is ignored for ranking. Ponte and Croft treat the query Q as a binary vector over the entire vocabulary and use (3.2) for estimating of the probability of generating query text (the notation M_D is used to indicate that the query is generated by a document language model).

$$P(Q/M_D) = \prod_{w \in Q} P(w/M_{D)} \prod_{w \notin Q} 1 - P(w/M_D) \tag{3.2}$$

Song and Croft [2], Hiemstra [3], and Miller et al [6] treat the query Q as a sequence of independent words instead of a binary vector and use (3.3) for query likelihood (q_w is the number of times the word w occurs in the query).

$$P(Q/M_D) = \prod_{w} P(w/M_D)^{q_w} \tag{3.3}$$

In the present study, the formula specified in equation (3.3) is used as a baseline in the experiments.

3.2 Relevance models

Lavrenko and Croft [5] incorporate relevance feedback and query expansion into language modeling frameworks. They proposed a technique for estimating a relevance model based on the query. The relevance model, $P(w/R)$, is estimated using a joint probability of observing the word w together with query words $q_1, q_2, ..., q_m$.

$$P(w/R) \approx P(w/Q) = \frac{p(w,q_1,...,q_m)}{P(q_1,...,q_m)} = \frac{p(w,q_1,...,q_m)}{\sum_{vocabulary} p(v,q_1,...,q_m)} \tag{3.4}$$

Lavrenko and Croft describe two methods of estimating the joint probability. The two methods differ in the independence assumptions that are being made. The first method assumes that w was sampled in the same way as the query words. The second method assumes that w and the query words were sampled using two different mechanisms. In this paper, we use the first method. If we assume that w and $q_1, q_2, ..., q_m$ are mutually independent once we pick a distribution M, then we get:

$$p(w,q_1,...,q_m) = \sum_{M \in \mathrm{M}} p(M)p(w/M)\prod_{i=1}^{m} p(q_i / M) \quad (3.5)$$

Here P(M) denotes some prior probability which is kept uniform over all distributions *M*.

Lavrenko and Croft [5] calculate the KL divergence between the relevance model and a document model. The KL divergence, which is given in equation (3.6), is used to rank documents. Documents with smaller divergence are considered more relevant.

$$KL(R \parallel M_d) = \sum_w P(w/R) \log \frac{P(w/R)}{P(w/M_d)} \quad (3.6)$$

In the present study, we use equation (3.6) for the baseline relevance model in the experiments.

3.3 Time-Based Language Models

The study of the relationship between time and relevance in section 2 shows that for time-based queries, documents with different document creation dates/timestamps may have different prior probabilities for relevance. Therefore, we propose to replace p(d) in equation (3.1) and P(M) in equation (3.5) with some probability dependent on documents date T, say $p(d / T_d)$ or $p(M / T_D)$. This gives us the time-based language models:

$$p(d / q) \propto p(q / d)\, p(d / T_d) \quad (3.7)$$

and

$$p(w,q_1,...,q_m) = \sum_{M \in \mathrm{M}} p(M/T_D)p(w/M)\prod_{i=1}^{m} p(q_i / M) \quad (3.8)$$

Although $p(d / T_d)$ in the query likelihood language model and $p(M / T_D)$ in the relevance model have somewhat different meanings, we refer to both as $p(D / T_D)$ for simplicity. In the case of the relevance model, the time-based prior will affect the documents that are used to construct the model. When viewed as a form of query expansion, this means that the expansion will be based on the top-ranked documents subject to a time constraint, such as favoring the most recent documents. This property could be exploited to change the interpretation of a query in, for example, systems with user models that change over time.

The next challenge is to estimate the probability $p(D / T_D)$. We suggest some simple methods for estimating this probability for different types of time-based queries.

3.3.1 Exponential Distribution

For queries where recency is a major requirement of a user's information need, we used an exponential distribution for prior probability assignment. The prior $p(D / T_D)$ is given in equation (3.9). Documents with a more recent creation date are assigned higher probability.

$$p(D / T_D) = P(T_D) = \lambda e^{-\frac{(T_C - T_D)^2}{2}} \quad (3.9)$$

Here T_C is the most recent date (in month) in the whole collection and T_D is the creation date of a document.

3.3.2 Normal Distribution

For queries where the user favors a particular time period, we propose to use a normal distribution for prior probability assignment. The prior $p(D / T_D)$ is given in equation (3.10). Documents closer to the mean are assigned higher probability according to normal distribution.

$$p(D / T_D) = P(T_D) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(T_C - T_D - \mu)^2}{2\sigma^2}} \quad (3.10)$$

Here T_C is the most recent date (in month) in the whole collection and T_D is the creation date of a document.

The training of the parameters in equation (3.9) and (3.10) and experimental results are detailed in section 4.

4. EXPERIMENTAL DESIGN AND RESULTS

4.1 Data

The training data consists of two sets: 25 queries from TREC queries 301-350 over collections from TREC volumes 4 and volume 5, and 5 queries from TREC queries 151-200 over collections from TREC volumes 1, 2 and 4. The first training set is the first type of time-based queries, which has more relevant documents in the recent past. This set is used for determining the parameters for the exponential distribution. The second training set is the second type of time-based queries, which has more relevant documents in a specific period in the past. This is a smaller set since the specific time periods of interest will vary from query to query. For this reason, the experiments involving this query type are more indicative in nature rather than being comprehensive. The training set is used to set the parameters of the normal distribution.

The test data consists of three sets: 25 queries from TREC queries 351-400 over collections from TREC volumes 4 and volume 5, 5 queries from TREC queries 151-200 over collections from TREC volumes 1, 2 and 4, and 5 queries from TREC queries 251-300 over collections from TREC volumes 2 and 4. The first test set is to test the performance of time-based language models with exponential distributions. The second test set and the third test set are to test the performance of time-based language models with normal distributions. The specific queries used in these sets are listed in the appendix.

4.2 Experimental Design

Four sets of training experiments and six sets of testing experiments were performed.

The first set of experiments was used to determine the best value of λ in the exponential distribution on time-based query likelihood language models. The second set of experiments was to determine the best value of λ in the exponential distribution on

time-based relevance language models. The third set of experiments is to determine the best value of σ in the normal distribution on time-based query likelihood language models. The fourth set of experiments is to determine the best parameter of σ in the normal distribution on time-based relevance language models. For each set of training experiments, a number of different parameter values were tested. The parameter value with highest performance in terms of average precision was chosen as best parameter value for later experiments.

Table 1 shows results using the exponential distribution in the time-based query likelihood language model and the first category of queries. Only three values of λ are shown here, although more were tried. The best value in terms of performance was .01. The same result was obtained with the time-based relevance model.

Table 1. Training query set 1 (query likelihood)

	LM	**TB1-.005**	**% Chg**	**TB1-.01**	**% Chg**	**TB1-.02**	**% Chg**
Rel	3546	3546		3546		3546	
Rret	1187	1167	-1.7	1162	-2.1.	1162	-2.1
0.00	0.382	0.412	+7.9	0.414	+8.4	0.414	+8.4
0.10	0.284	0.292	+2.8	0.294	+3.5	0.294	+3.5
0.20	0.206	0.206		0.209	+1.5	0.209	+1.5
0.30	0.144	0.149	+3.5	0.151	+4.9	0.151	+4.9
0.40	0.109	0.113	+3.7	0.115	+5.5	0.114	+5.5
0.50	0.097	0.097		0.099	+2.1	0.099	+2.1
0.60	0.086	0.084	-2.3	0.084	-2.3	0.084	-2.3
0.70	0.067	0.068	+1.5	0.068	+1.5	0.068	+1.5
0.80	0.017	0.017		0.017		0.017	
0.90	0.005	0.005		0.005		0.005	
1.00	0	0		0		0	
Avg	0.112	0.116	+3.6	0.117	+4.5	0.117	+4.5

LM: query–likelihood language model.

TB1-a: time-based query-likelihood model with λ = a in the exponential distribution.

Table 2 shows the results using different values of σ with the time-based relevance model and the second category of time queries. In this case the value of 15 for σ produced the best results. In the case of the time-based query likelihood model, a value of 20 for σ was the best.

4.2 Empirical Results

The six sets of test experiments use parameters determined from the training experiments. The performance is compared with the performance of appropriate baseline query likelihood result or relevance model result.

The results of the first two experiments with test set 1 are shown in Table 3. The exponential distribution given in figure 4.1 is used to assign prior probability in both time-based query likelihood language models and time-based relevance language models. Table 3 shows that time-based query likelihood language models outperform the baseline query likelihood language models by 6.2% in average precision and time-base relevance models outperform baseline relevance models by 6.9% (despite the much higher baseline).

Table 2. Training query set 2 (relevance model)

	RM	**TB2-10**	**% Chg**	***TB2-15**	**% Chg**	**TB2-20**	**% Chg**
Rel	2020	2020		2020		2020	
Rret	365	456	+24.9	480	+31.5	451	+23.6
0.00	0.402	0.028	-33.3	0.494	+22.9	0.446	+10.9
0.10	0.108	0.119	+10.2	0.126	+16.7	0.117	+8.3
0.20	0.082	0.10 4	+26.8	0.108	+31.7	0.100	+22.0
0.30	0.022	0.093	+322.7	0.086	+290.9	0.015	-31.8
0.40	0.019	0.017	-10.5	0	-100.0	0	-100.0
0.50	0	0		0		0	
0.60	0	0		0		0	
0.70	0	0		0		0	
0.80	0	0		0		0	
0.90	0	0		0		0	
1.00	0	0		0		0	
Avg	0.033	0.039	+18.2	0.041	+24.2	0.035	+6.1

RM: relevance model

TB2-b: time-based relevance model with σ = b in the normal distribution

Table 3. Testing query set 1

	LM	**TB1-.01**	**% Chg**		**RM**	**TB2-.01**	**% Chg**
Rel	2804	2804			2804	2804	
Rret	1043	1088	+4.3		1529	1552	+1.5
0.00	0.588	0.645	+9.8	0.00	0.582	0.595	+2.1
0.10	0.286	0.312	+9.2	0.10	0.443	0.460	+4.0
0.20	0.244	0.265	+8.7	0.20	0.397	0.412	+4.5
0.30	0.206	0.226	+9.6	0.30	0.321	0.340	+6.0
0.40	0.140	0.159	+12.9	0.40	0.250	0.270	+8.1
0.50	0.107	0.115	+7.5	0.50	0.193	0.213	+10.5
0.60	0.058	0.057	+1.3	0.60	0.143	0.154	+7.5
0.70	0.037	0.042	+12.1	0.70	0.110	0.116	+4.7
0.80	0.025	0.026	+4.5	0.80	0.084	0.083	-0.4
0.90	0.018	0.020	+12.6	0.90	0.045	0.051	+12.7
1.00	0.010	0.010	-3.9	1.00	0.009	0.011	+23.9
Avg	0.134	0.142	+6.2		0.220	0.235	+6.9

The results of the third and fourth experiments with test set 2 are shown in Table 4. The normal distribution with σ = 20 given in Figure 4.2 is used to assign prior probability in time-based query likelihood models. For the relevance models, a normal distribution with σ = 15 is used. The normal distribution favors documents in a specific period, around 38 (Oct. 1991) in the case of Figure 4.2. Documents with document date closer to Oct. 1991 are assigned higher prior probabilities. Table 4 shows that the time-based query likelihood model outperforms the baseline query likelihood model by 12.0% and time-based relevance models outperform baseline relevance models by 13.4% in terms of average precision on these queries. Again, the relevance model was a much higher baseline but the inclusion of time-based priors still made a significant difference.

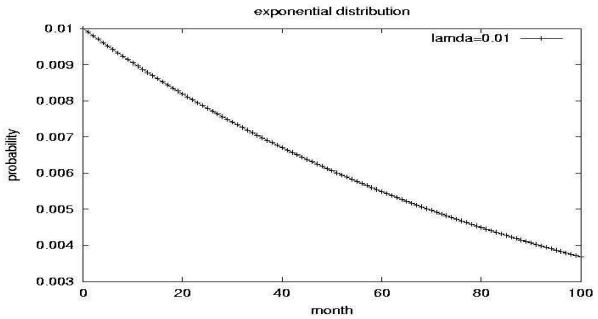

Figure 4.1: Exponential distribution used for priors

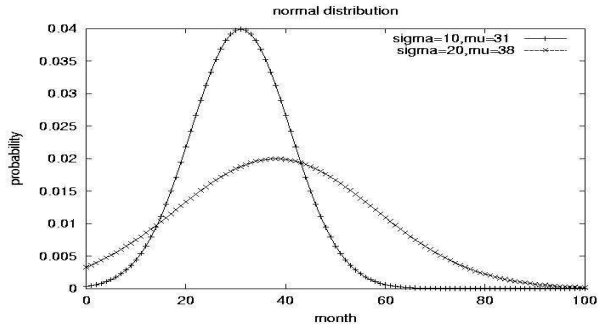

Figure 4.2: Normal distributions used for priors

Table 4. Testing query set 2

	LM	TB1-20	% Chg		RM	TB2-15	% Chg
Rel	1567	1567			1567	1567	
Rret	626	669	+6.9		783	779	-0.5
0.00	0.525	0.560	+6.7	0.00	0.633	0.8	+26.4
0.10	0.468	0.510	+9.0	0.10	0.579	0.752	+29.9
0.20	0.449	0.495	+10.2	0.20	0.575	0.686	+19.3
0.30	0.334	0.472	+41.3	0.30	0.562	0.600	+6.8
0.40	0.288	0.433	+50.3	0.40	0.526	0.569	+8.2
0.50	0.126	0.338	+168.3	0.50	0.474	0.502	+5.9
0.60	0.036	0.216	+500	0.60	0.412	0.430	+4.4
0.70	0	0.035		0.70	0.320	0.330	+3.1
0.80	0	0		0.80	0.221	0.137	-38.0
0.90	0	0		0.90	0.114	0.097	-14.9
1.00	0	0		1.00	0	0	
Avg	0.241	0.270	+12.0		0.395	0.448	+13.4

Queries in test set 2 and queries in training set 2 have similar characteristics. Each query has more relevant documents within a 20-month period from Jan. 1991 to Dec. 1992. Therefore, the best parameter values determined from training set 2 are used in the testing experiments with test set 2.

In the fifth and sixth test experiments with test set 3, each query has more relevant documents within a 10-month period from Jan. 1992 to Oct. 1992. It has a shorter period than queries in training set 2. Therefore, the parameter values determined from training set 2 cannot be directly used in test set 3. However, we assumed that there is a relationship between the value of σ and the length of the specific time period by choosing parameter values in proportion with the length of the specific period. That is, because the time period of interest in test set 3 was half the length of the time period in training set 2, we set $\sigma = 10$, which was half the best value for training set 2. The normal distribution with $\sigma = 10$ shown in figure 4.2 is used to assign prior probability in time-based query likelihood models. Similarly, the value $\sigma = 7.5$ was used for the time-based relevance models.

The results are given in Table 5. This shows that time-base query likelihood models outperform the baseline query likelihood language models by 60% and that time-based relevance models outperform the baseline relevance models by 31.4% in terms of average precision, although in this case the overall results are lower than with test set 2.

Table 5. Testing query set 3

	LM	TB1-10	% Chg		RM	TB2-7.5	% Chg
Rel	1003	1003			1003	1003	
Rret	122	124	+1.6		294	409	+39.1
0.00	0.540	0.587	+8.7	0.00	0.585	0.572	-2.3
0.10	0.045	0.084	+86.7	0.10	0.231	0.262	+13.4
0.20	0.018	0.023	+27.8	0.20	0.182	0.196	+7.7
0.30	0.010	0.019	+90.0	0.30	0.121	0.157	+29.8
0.40	0.010	0.014	+40.0	0.40	0.054	0.122	+125.9
0.50	0	0.014		0.50	0.028	0.114	+307.1
0.60	0	0		0.60	0.024	0.027	+12.5
0.70	0	0		0.70	0.013	0.015	+15.4
0.80	0	0		0.80	0	0	
0.90	0	0		0.90	0	0	
1.00	0	0		1.00	0	0	
Avg	0.020	0.032	+60.0		0.086	0.113	+31.4

5. RELATED RESEARCH

As mentioned previously, the creation date of a document has long been recognized as an important attribute in commercial IR systems. In terms of research, the role of time in retrieval has been somewhat neglected, although recency is often mentioned in discussions of relevance and utility. There has been work on constructing timelines automatically from time-tagged retrieved documents as a visualization and discovery tool (e.g. [7, 8]). There has also been research that exploits the temporal aspect of news streams to improve topic tracking and the detection of novel information [9]. Other related work includes efforts to improve the extraction of time tags for question answering [10].

6. CONCLUSIONS AND FUTURE WORK

In this paper, we studied the relationship between time and relevance based on TREC ad-hoc title queries. We proposed time-based language model frameworks, which incorporate time into

both query likelihood language models and relevance-based language models. Exponential distributions or normal distributions are used to replace the uniform prior probability in these models. Our empirical results show that, for two important classes of queries, time-based language models outperform baseline query likelihood language models and relevance-based language models respectively using TREC standard measures. The main contribution of this work is to show that contextual features such as time constraints can be incorporated into the underlying retrieval model without resorting to heuristic approaches.

In future work, we will develop techniques to automatically classify time-based queries and set parameters. We have also started using these techniques for time-based question answering. A number of questions, such as "Who is the prime minister of Australia?", have time-dependent answers. We are attempting to use the time-based language models to change the ranking of answer passages and the subsequent answers that are extracted. Our goal is to have a time "slide bar" that would change the answer as it is moved. For this work, we are using extracted dates in addition to document dates.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by SPAWARSYSCEN-SD grant numbers N66001-99-1-8912 and N66001-02-1-8903, and in part by Advanced Research and Development Activity under contract number MDA904-01-C-0984.

Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] J. Ponte and W. B. Croft, "A Language Modeling Approach to information retrieval". *Proceedings of the 21st annual international ACM SIGIR conference*, 275-281, 1998.

[2] F. Song and W. B. Croft. "A general language model for information retrieva". *Proceedings of the 22nd annual international ACM SIGIR conference*, 279-280, 1999

[3] D. Hiemstra. *Using language models for information retrieval*. PhD thesis, University of Twente, 2001.

[4] J. Lafferty and C. Zhai. "Document language models, query models, and risk minimization for information retrieval". *Proceedings of the 24th annual international ACM SIGIR conference*, 111-119, 2001.

[5] V. Lavrenko and W. B. Croft. "Relevance-based language models". *Proceedings of the 24th annual international ACM SIGIR conference*, 120-127, 2001.

[6] D. Miller, T. Leek, and R. Schwartz. "A Hidden Markov Model information retrieval system". *Proceedings of the 22nd annual international ACM SIGIR conference*, 214-221, 1999.

[7] Swan, R. and Allan, J. "Automatic Generation of Overview Timelines". *Proceedings of SIGIR 2000 Conference*, Athens, 49-56, 2000.

[8] Swan, R. and Jensen, D. "TimeMines: Constructing Timelines with Statistical Models of Word Usage". *Proceedings of KDD 2000 Conference*, 73-80, 2000.

[9] J. Allan, R. Gupta, and V. Khandelwal. "Temporal Summaries of News Topics". *Proceedings of ACM SIGIR 01 conference*, 10-18, 2001.

[10] J. Pustejovsky, "TERQAS: Time and Event Recognition for Question Answering Systems", ARDA Workshop, MITRE, Boston (2002). (http://www.cs.brandeis.edu/~jamesp/arda/time/index.html)

APPENDIX: QUERIES USED IN EXPERIMENTS

(1) Training set 1 consists of following TREC queries:

301, 302, 304, 306, 307, 311, 313, 314, 316, 318, 319, 321, 326,

327, 329, 331, 333, 334, 437, 340, 341, 343, 345, 346, 347

(2) Training set 2 consists of following queries:

151, 156, 161, 163, 177

(3) Test set 1 consists of following queries:

351, 352, 353, 355, 356, 357, 359, 360, 365, 367, 370, 372, 373,

376, 378, 381, 382, 384, 385, 387, 389, 391, 395, 396, 400

(4) Test set 2 consists of following queries:

180, 182, 185, 189, 191

(5) Test set 3 consists of following queries:

264, 266, 273, 284, 297