Answering General Time-Sensitive Queries

Wisam Dakka
Columbia University
wisam@cs.columbia.edu

Luis Gravano
Columbia University
gravano@cs.columbia.edu

Panagiotis G. Ipeirotis
New York University
panos@stern.nyu.edu

ABSTRACT

Time is an important dimension of relevance for a large number of searches, such as over blogs and news archives. So far, research on searching over such collections has largely focused on locating topically similar documents for a query. Unfortunately, topic similarity alone is not always sufficient for document ranking. In this paper, we observe that, for an important class of queries that we call *time-sensitive queries*, the publication time of the documents in a news archive is important and should be considered in conjunction with the topic similarity to derive the final document ranking. Earlier work has focused on improving retrieval for "recency" queries that target recent documents. We propose a more general framework for handling time-sensitive queries and we automatically identify the important time intervals that are likely to be of interest for a query. Then, we build scoring techniques that seamlessly integrate the temporal aspect into the overall ranking mechanism. We extensively evaluated our techniques using a variety of news article data sets, including TREC data as well as real web data analyzed using the Amazon Mechanical Turk. We examined several alternatives for detecting the important time intervals for a query over a news archive and for incorporating this information in the retrieval process. Our techniques are robust and significantly improve result quality for time-sensitive queries compared to state-of-the-art retrieval techniques.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.m [Miscellaneous]: Processing Time-Sensitive Queries.

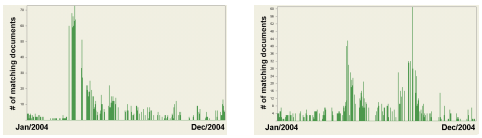
General Terms: Algorithms, Performance, Experimentation.

Keywords: Time-Sensitive Search.

1. TIME-SENSITIVE QUERIES

Time is an important dimension of relevance for a large number of searches, such as over blogs and news archives. So far, research on searching over such collections has largely focused on retrieving topically similar documents for a query. Unfortunately, ignoring or not fully exploiting the time dimension can be detrimental for a large family of queries for which we should consider not only the document topical relevance but the publication time of the documents as well, as demonstrated by the following example:

EXAMPLE 1. Consider the query *[Madrid bombing]* over the Newsblaster [6] news archive. Figure 1 zooms in on a portion of the histogram for the query results, reporting the number of matching documents in the news archive for each day between January and December 2004. This histogram reveals particular time intervals that are likely to be of special interest for the query, such as the month of March 2004, when a terrorist group bombed trains in Madrid. The same figure shows an analogous histogram for query *[Google IPO]*: the "peaks" in the histogram coincide with two important events, namely the announcement of the Google IPO and, a few months later, the actual IPO. □



[Madrid bombing] *[Google IPO]*

Figure 1: Histograms for queries *[Madrid bombing]* and *[Google IPO]*, showing the number of documents with all query words for each day from January to December 2004 in a news archive.

These examples motivate two observations on searching over news archives. First, topic similarity ranking does not model time explicitly, so the important dimension of time is not considered when deciding on the results that are returned for a query. The various "peaks" in the Figure 1 histograms, which reveal important information for the queries, are thus not leveraged to produce high-quality query results. Second, a topic similarity ranking of the query results often does not reflect the distribution of relevant documents over time. In fact, for many queries, users have a general –but often vague and unspecified– idea about the relevant time periods. For example, the query *[Madrid bombing]* might (implicitly) be after articles from March and April 2004. So perhaps a better formulation of the query would be *[Madrid bombing* ***prefer****: 03/11/2004–04/30/2004]*, indicating the relevant time interval for the query.

Earlier work [4] has focused on improving retrieval for "recency" queries that target recent documents. We propose a more general framework for handling all *time-sensitive queries* (such as *[Madrid bombing]*) that target relevant documents that are are not spread uniformly over time but rather tend to concentrate in restricted time intervals. Next, Section 2 presents our strategy for handling general time-sensitive queries, beyond recency queries, with language models and Section 3 briefly discusses our findings and results.

2. ANSWERING TIME-SENSITIVE QUERIES

We propose a general framework for answering time-sensitive

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

queries by incorporating time into language models[1] in a principled manner. For a given time-sensitive query over a news archive, our approach automatically identifies important time intervals for the query. These intervals are then used to combine temporal relevance and topic similarity, to adjust the document relevance scores by boosting the scores of documents published within the important intervals. The goal is to return results that are both topically relevant to the queries and are also from the most "important" time periods for the queries.

Incorporating Time into LM: The query likelihood model (QL) [8] estimates the relevance of a document d to a query q by computing the conditional probability $p(d|q)$ that d is topically relevant to q, which is proportional to $p(d) \cdot p(q|d)$. To answer general time-sensitive queries such as *[Madrid bombing]*, we want to identify not just the relevant documents for the query, but also the relevant time periods. Craswell et al. [1] introduced a framework to complement the topical relevance of a document for a query with additional evidence (e.g., ClickDistance [1]). We build on this framework and conceptually "decouple" each document d into a content component c_d as well as a temporal component t_d. We can then express $p(d|q)$ as $p(c_d, t_d|q)$, which represents the probability that c_d is topically relevant to q and that t_d is a time period relevant to q, where c_d is the content of document d and t_d is the time when d was published. Assuming that topic similarity is conditionally independent of temporal relevance, given query q, we have:

$$p(d|q) = p(c_d, t_d|q) \quad \propto \quad p(q|c_d) \cdot p(c_d) \cdot p(q|t_d) \cdot p(t_d)$$

Note that c_d is what we traditionally refer to as d in language models; our use of c_d is to emphasize that a document in our modified model consists of the traditional textual content component c_d and the temporal information t_d. So, the document prior $p(c_d)$ is typically assumed to be uniform for all documents, considering that there is no document that is more likely to be relevant *across all possible queries*. The time prior $p(t_d)$ can be defined proportionally to the *total* number of documents published at time t_d. The term $p(q|c_d)$ corresponds to the likelihood of generating query q from document c_d and can be computed using existing techniques, such as the QL model or the relevance language model (RM) [3]. Finally, $p(q|t_d)$ corresponds to the probability of "observing" q in the documents published in time t_d. We call this probability the temporal relevance of t_d and we discuss how to estimate it next.

Computing Temporal Relevance: We estimate $p(q|t)$ for query q and time t by analyzing the number of documents matching query q over time. Our conjecture is that certain patterns of matching frequencies over time might help identify relevant time periods for the query. For example, an abrupt change in match frequency between consecutive days might signal a relevant event for the query. To incorporate time into the language models, we then estimate $p(q|t)$ based on the distribution of query matches over time. Specifically, we propose to arrange all time periods into *bins*, such that each bin represents a "priority level." We then assign estimated relevance values to the time periods in these bins accordingly. We have explored alternate binning techniques based on different underlying hypotheses on how to identify the important time intervals. For example, our "running mean" technique considers the average daily match frequency across the archive, to calibrate the "popularity" of a query, in terms of its document matches in the archive over time. For this, we "reduce" the match frequency of a day by subtracting the average daily match frequency computed up to that day. We use the reduced frequencies to sort times into bins, so that bin b_0 will contain the days with the largest reduced frequency, b_1 will correspond to the second-largest reduced frequency, and so on. We define the $p(q|t)$ values based on the assignment of times to bins $b_0, \ldots, b_\ell$ and decay the estimated relevance of bins exponentially, as Li and Croft [4] did for recency queries, with their distance to the time(s) of interest: we define $p(q|t) \propto \lambda \cdot \exp(-\lambda \cdot bin(t))$, where $bin(t)$ returns the index of the time t bin and λ is the parameter of the exponential distribution, often called the rate parameter.

[1]We have also incorporated time into BM25 [7], but we omit the discussion because of space constraints.

3. DISCUSSION

We built a general framework for processing time-sensitive queries over a news archive, with techniques for identifying important time periods for a query. (We omit further details because of space constraints.) We performed a thorough evaluation over multiple data sets, including TREC data[2] as well as six years' worth of Newsblaster news articles analyzed using the Amazon Mechanical Turk[3]. We have implemented our system on top of Indri and Lemur, state-of-the-art search engines [4]. We compared several alternatives to compute the temporal relevance $p(q|t_d)$, including several variations of the binning techniques, recent work by Jones and Diaz [2] that defines $p(q|t)$ as the normalized sum of the relevance scores of documents that are published at time t for query q, and Li and Croft's work on recency queries [4], which we discussed above. We integrated these alternatives into QL, RM, and BM25. Overall, we showed that our techniques improve the quality of search results for time-sensitive queries, compared to the existing state-of-the-art algorithms. Our time-sensitive techniques tend to significantly improve precision at the top recall cutoff levels relative to the baseline techniques. However, the precision drops for higher recall cutoff levels. We also noticed that, for certain queries, the time-sensitive techniques introduce relevant documents that the baseline techniques could not capture. In summary, integrating time in the retrieval task can improve the quality of the retrieval results. These results motivate further research such as inferring the temporal relevance of a document by analyzing its contents [5] and not only relying on the publication time, or introducing time-based diversity in query results by grouping the results into clusters of relevant time ranges, enabling users to be aware of and interact with time information.

References

[1] N. Craswell, S. E. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, 2005.

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

[3] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, 2001.

[4] X. Li and W. B. Croft. Time-based language models. In *Proceedings of the 12th ACM Conference on Information and Knowledge Management (CIKM 2003)*, 2003.

[5] I. Mani, J. Pustejovsky, and R. Gaizauskas. *The Language of Time: A Reader*. Oxford University Press, 2005.

[6] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the 2nd International Conference on Human Language Technology (HLT 2002)*, 2002.

[7] S. E. Robertson. The probability ranking principle in IR. *Readings in information retrieval*, pages 281–286, 1997.

[8] F. Song and W. B. Croft. A general language model for information retrieval. In *Proceedings of the 8th ACM Conference on Information and Knowledge Management (CIKM 1999)*, 1999.

[2]`http://trec.nist.gov/data/test_coll.html`

[3]`http://www.mturk.com`

[4]`http://www.lemurproject.org`