Document Timespan Normalisation and Understanding Temporality for Clinical Records Search

Bevan Koopman
Australian e-Health Research Centre
CSIRO, Brisbane, Australia
bevan.koopman@csiro.au

Guido Zuccon
Faculty of Science & Technology, Queensland
University of Technology, Brisbane, Australia
g.zuccon@qut.edu.au

ABSTRACT

Previous qualitative research has highlighted that temporality plays an important role in relevance for clinical records search. In this study, an investigation is undertaken to determine the effect that the timespan of events within a patient record has on relevance in a retrieval scenario. In addition, based on the standard practise of document length normalisation, a *document timespan normalisation* model that specifically accounts for timespans is proposed. Initial analysis revealed that in general relevant patient records tended to cover a longer timespan of events than non-relevant patient records. However, an empirical evaluation using the TREC Medical Records track supports the opposite view that shorter documents (in terms of timespan) are better for retrieval. These findings highlight that the role of temporality in relevance is complex and how to effectively deal with temporality within a retrieval scenario remains an open question.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Theory, Experimentation

1. INTRODUCTION

Considerable research effort has focused on the application of information retrieval (IR) methods for searching clinical records [9, 8, 7, 10, 11, 12]. However, clinical search presents some specific aspects that makes estimating relevance challenging in this domain [4, ch.2]. One of these aspects is the role that temporality plays in determining relevance. A study of how medical professionals perform relevance assessments highlighted that temporality was a significant and cognitively demanding aspect of relevance [6]. When searching documents related to a patient's hospital admission, temporality can affect relevance in a number of ways: query terms may be found in a patient's past medical history and the assessor needs determine whether the information is still valid (e.g., chronic vs. acute conditions); some conditions are defined with temporal constraints (e.g., chronic back pain is considered to be persisting for at least 3 months). In this paper, we hypothesise that the timespan of events documented in a patient record has an effect on the relevance of that patient (document) to a clinical query. Specifically, we investigate the following research questions:

RQ1: Are patient records covering a shorter timespan of events more or less relevant in a retrieval scenario?

There are two alternative hypotheses underlying RQ1. Firstly, *shorter* timespans may be more relevant in that they represent a more focused patient record that covers a small set of specific conditions; thus they may be more relevant to queries for this condition than patient records with longer timespans where many conditions are represented. In contrast, *longer* timespans may be more relevant in that there is more evidence supporting the condition being chronic or on-going, as opposed to something that is peripheral or less significant within a patient's record.

RQ2: Does a retrieval model that specifically accounts for timespan of events covered in a patient record lead to improvements in retrieval effectiveness?

We draw a parallel with the common IR approach of document length normalisation, used for example in BM25, and instantiate a *document timespan normalisation* component in order to investigate the above research question.

2. RELATED WORK

Previous research has considered the role of temporality in IR. Efforts such as the TREC Temporal Summarization Track have focused on the goal of developing systems for "efficiently monitoring the information associated with an event over time" [1]. Much of the focus of this line of research has been on the development of temporal-aware models that favour more recent events, i.e., up-to-date information. This is in contrast to the problem studied in this paper of how timespans or durations of events effect relevance.

More aligned to this study was the development of retrieval models that included a temporal profile of both queries and documents. Diaz & Jones [2] used these profiles to match "temporal relevance" between the query and document (i.e., did both occur within a similar time period). This differs from our study in that we are not concerned with the actual time period (and do not consider temporal information of the query) but instead are concerned with the timespan of events represented by a single patient record

ADCS'14, November 27–28 2014, Melbourne, VIC, Australia.

http://dx.doi.org/10.1145/2682862.2682879.

(i.e., the temporal length of the document).

Considerable research effort has focused on the application of IR to searching clinical records; many of these works have exploited features that characterise (and often somewhat differentiate) clinical records. Age and gender of a patient are two common examples [9, 8], while specific retrieval methods have also been proposed for a number of others. Limsopatham et al. [7] considered information regarding the hospital department that produced a clinical record (e.g., ER, Oncology department, etc.) and integrated this information within voting models and federated search approaches. Similarly, Zhu et al. [11] exploited the distribution of relevant evidence across multiple records for a patient, along with department source information. The same authors [10] have also proposed exploiting evidence from different medical document collections, such as medical publication repositories, medical image captions, etc., that are then combined using Mixture of Relevance Models to best estimate term likelihoods for improving clinical records retrieval. These methods both highlighted the complex factors influencing relevance in clinical IR, while also demonstrating that accounting for these factors in new retrieval models can lead to more effective clinical IR systems.

Within the clinical domain, Edinger et al. [3] provided a valuable failure analysis of systems participating in the TREC Medical Records Track (the same test collection used in this study). Their analysis revealed temporality to be a significant factor adversely affecting retrieval, with 18 out of the 35 queries identified as temporally affected. A separate study analysed how cognitively demanding relevance assessment was for clinical professionals [6]. Results from this study highlighted that temporality played an import role in determining relevance and that temporality made relevance assessment more demanding. The findings of both the above mentioned studies motivate the research questions proposed in this study: to further understand how timespan effects relevance and investigate new retrieval models to account for timespans.

3. TIMESPAN RETRIEVAL MODEL

A document D may contain a number of time points, $\langle t_0, \ldots, t_n \rangle$. The timespan of a document, denoted T_D, is calculated as:

$$T_D = \max(\langle t_0, \ldots, t_n \rangle) - \min(\langle t_0, \ldots, t_n \rangle), \quad (1)$$

i.e., the duration from the earliest to the latest time point.

To incorporate this information into a retrieval function, we adapt the Lemur variant of tf-idf[1], which uses BM25 term weighting, and calculates a retrieval status value (RSV) for document D in response to query Q as:

$$\text{RSV}(D,Q) = \sum_{q \in Q} \frac{tf_{q,D}(k_1+1)}{tf_{q,D} + k_1(1 - b + b\frac{|D|}{|D^{\text{avg}}|})} \log \frac{N}{n_q}, \quad (2)$$

where $tf_{q,D}$ is the term frequency of q within the document D, N is the total number of documents in the collection and n_q is the number of documents containing the query term q.

In the same vein as the document length normalisation component, we add a *document timespan normalisation* component that includes: the timespan of the document T_D, the average timespan of all documents in the collection T^{avg} and a new parameter $b_t \in [0,1]$, controlling the influence of timespan normalisation. The new retrieval function is:

$$\text{RSV}(D,Q) = \sum_{q \in Q} \frac{tf_{q,D}(k_1+1)}{tf_{q,D} + k_1(1 - b + b\frac{|D|}{|D^{\text{avg}}|}) + b_t \frac{T_D}{T^{\text{avg}}}} \log \frac{N}{n_q}. \quad (3)$$

This retrieval model favours documents of shorter timespan and is denoted tf-idf-short (favours shorter). We also investigate an alternative model that favours longer documents:

$$\text{RSV}(D,Q) = \sum_{q \in Q} \frac{tf_{q,D}(k_1+1)}{tf_{q,D} + k_1(1 - b + b\frac{|D|}{|D^{\text{avg}}|}) - b_t \frac{T_D}{T^{\text{avg}}}} \log \frac{N}{n_q} \quad (4)$$

This retrieval model is denoted tf-idf-long (favour longer).

4. EMPIRICAL EVALUATION

This section details our experimental setup, evaluation methodology and results; discussion of results is reserved for the next section.

4.1 Clinical Documents and Test Collection

The TREC Medical Records Track, a collection of 100,866 clinical record documents from U.S. hospitals, was used to empirically investigate the influence of timepans on retrieval. Documents belonging to a single patient's admission were treated as sub-documents and were concatenated together into a single document called a patient *visit* document. This was done because the unit of retrieval in TREC MedTrack was a patient visit rather than an individual report. Collapsing reports to patient visits was a common practise among many TREC MedTrack participants [9, 8]. The corpus then contained 17,198 patient visit documents. Query topics (81 in total) and relevance judgements were combined from the 2011 and 2012 TREC tracks to have a single, larger query set for more powerful statistical analysis.

4.2 Extracting Timespans from Documents

Before the TREC MedTrack collection was made publicly available, the documents were de-identified to remove any Personal Health Identifiers, including names, places, organisations and dates.[2] The de-identification of dates is done by first recognising references to a date and then adjusting dates within a document by a random offset (e.g., shift all dates for a single patient forward 3 days). This allows for comparison between dates for a patient but removes the ability to know exactly when the events occurred. The de-identification algorithm used for TREC MedTrack also annotated all dates according to a specific format (e.g., `**DATE[Jan 21 2007]`). We could, therefore, easily extract all the mentions of dates (time points $\langle t_0, \ldots, t_n \rangle$, in our retrieval model) for a given document. The timespan of a document T_D, could be determined as the number of days between the earliest and latest time point. This was computed prior to retrieval and read at retrieval time. The entire timespan for the collection (earliest time point vs. latest time point for any document) was 3,512 days (9.62 years).

4.3 Baselines and Evaluation Measures

The evaluation measures used in MedTrack 2011 were bpref and precision @ 10 (P@10). However, in MedTrack

[1]Lemur's tf-idf model was chosen as it proved to be the most effective among a number of baselines used in previous studies [4, ch.4] and on the same test collection used here.

[2]Dates can, in some cases, be used to reveal the identify of a patient.

Model	Bpref	P@10	b_t
tf-idf	0.3839	0.4481	0.00
tf-idf-short	0.3857	0.4494	0.08
(cross-eval)	0.3843	0.4483	-
tf-idf-long	$0.3763^{\dagger}$	$0.3309^{\dagger}$	0.08
(cross-eval)	0.3840	0.4483	-

Table 1: Retrieval results on TREC MedTrack using timespan normalisation. Statistical significant results using paired t-test indicated with †.

2012 inferred measures and P@10 were used. Inferred measures required specific relevance assessments (prels) not available for 2011, but bpref and P@10 could be used for 2012 as qrels were available. Therefore, in this paper, we use bpref and P@10 to evaluate the two proposed retrieval models: tf-idf-short from Eq. 3, which favours shorter timespans and tf-idf-long from Eq. 4, which favours longer timespans. Both models add a timespan normalisation component, controlled by the additional parameter b_t; thus, a baseline retrieval method, representing the standard Lemur tf-idf model, is included by setting $b_t = 0.00$ and is denoted tf-idf. For the timespan model, we consider a full exploration of the parameter space of b_t from 0.0 to 1.0 in 0.01 increments. In addition, we also report the results of a 10-fold cross validation analysis (based on bpref).

5. RESULTS

RQ1: Are longer (or shorter) patient record timespans more relevant?

Using the relevance assessments from TREC MedTrack, we consider, on a per-query basis, the median timespan for relevant documents vs. the median timespan for non-relevant documents; this is shown in Figure 1. The large number of positive values suggests that, on average, relevant documents tend to have longer timespans than non-relevant documents (70% of queries display longer timespans for relevant documents). This result would support the use of the tf-idf-long model that favours documents with above average timespans. It is worth also noting that documents covering longer timespans are not necessarily greater in length: the correlation between timespan in days and length is words was only 0.12.

RQ2: Does accounting for the clinical record timespan lead to improvements in retrieval effectiveness?

Retrieval results of the three models (two timespan and a baseline) are reported in Table 1. The best performance was observed for $b_t = 0.08$ (representing an "oracle" tuned system). In addition, the results of the 10-cold cross validation are also reported. Small but not significant improvements are found for tf-idf-short while the tf-idf-long model is significantly less effective than the baseline.

The sensitivity of the parameter b_t that controls the influence of timespan is shown in Figure 2 ($b_t = 0.00$ represents the baseline tf-idf system). The results show that only small value of b_t, where timespan normalisation has less effect, leads to any improvement in retrieval effectiveness.

6. DISCUSSION

There are two opposing rationales around how timespan may influence relevance in a retrieval scenario. One intuition is that clinical records with *longer* timespans should be favoured because there is more evidence supporting the condition being chronic or on-going, as opposed to something that is peripheral or less significant within a patient's record. This reasoning is supported by the results for RQ1 that compared the median timespan of relevant vs. non-relevant documents shown in Figure 1 and where, in general, relevant documents tended to have longer timespans.

In contrast, the other intuition is that clinical records with *shorter* timespans should be favoured because these records are more cohesive, more likely based on a single condition and therefore more likely to be relevant if they contain the query terms. In this situation, a patient record with a longer timespan may contain significant references to past medical history that may no longer apply to the patient's current state. For example, a patient may have had Hepatitis C in the past, was treated and the condition was resolved; this patient would *not* be relevant to the TREC query of "Patients with Hepatitis C and HIV" as they no longer suffer from Hepatitis C. The longer the timespan of the patient record, the greater the risk that the record may be *temporally* not-relevant. This reasoning is supported by the results from the retrieval experiments that showed the tf-idf-short retrieval model was more effective. In addition, there is some support in the literature that indicates that past medical history can adversely affect clinical IR [4, 9, 8].

An important consideration and possible limitation of this study is what the actual timespan of the patient record represents. Initially, the study was proposed for the timespan to capture the duration that the patient was admitted to hospital — shorter admissions are more likely to represent acute conditions and longer admission are more likely represent chronic conditions. However, given the data, it was not possible to determine the specific admission date and the discharge date. Therefore, all dates were extracted from the record (including references to past admissions and possible future procedures); the timespan then represented the whole range of a patient's illness, as documented in that admission. This artefact of the collection may have limited the conclusions possible in this study; future work would therefore be directed towards determining the specific admission and discharge dates.

7. CONCLUSION & FUTURE WORK

Previous research has highlighted that temporality impacts relevance in clinical IR. This paper has contributed an initial investigation into whether the timespan of events mentioned in a patient's records is an indicator of relevance in a retrieval scenario. Initial analysis of the relevance assessments (qrels) revealed that relevant patient records tended to have longer timespans. However, when a document timespan normalisation component was incorporated into the retrieval function the results support favouring shorter patient records. Overall, the retrieval results suggest that accounting for temporality has a minor impact on retrieval effectiveness, at least when the proposed timespan normalisation retrieval model is used. However, qualitative findings from previous research has shown that temporality plays a key role for specific queries [3, 6]; this suggests that alternative retrieval methods that consider temporality may further improve retrieval effectiveness.

In this paper, we focus on dealing with temporality from

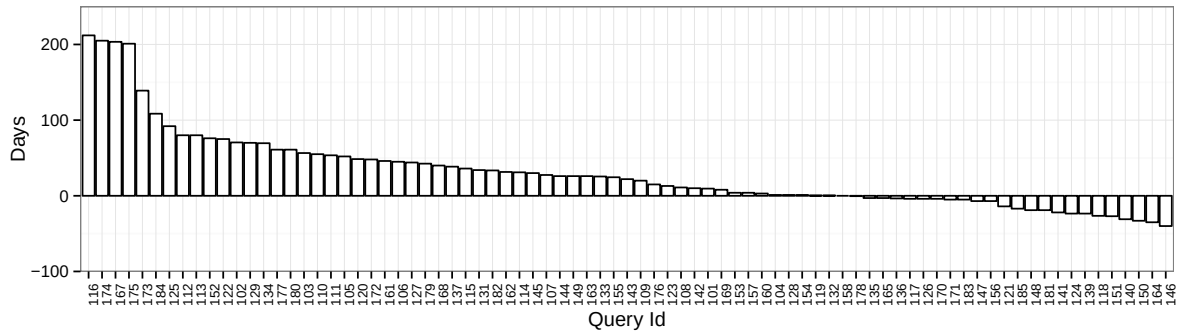
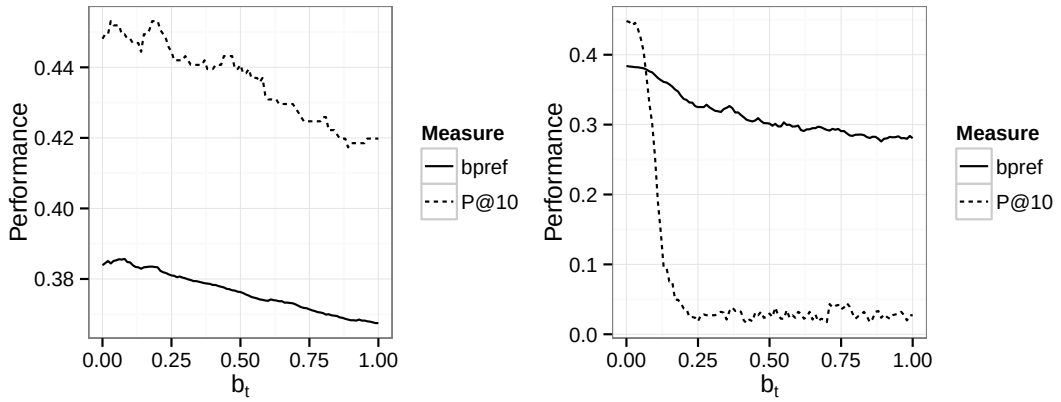


Figure 1: Median timespan of relevant − Median timespan of non-relevant. Positive values indicate relevant documents have longer timespans; negative values indicate relevant documents have shorter timespans.



(a) tf-idf-short (favour shorter timespans) (b) tf-idf-long (favour longer timespans)

Figure 2: Retrieval effectiveness (bpref and precision @ 10) for different values of the timespan normalisation parameter b_t. (Note that $b_t = 0.00$ represents the baseline tf-idf system.)

the document perspective (by extracting timespans and adding the document timespan normalisation component). Another approach to handle temporality from a document perspective is to divide the document up into past, present and possibly future content, then treat each of these content types as separate in the retrieval model. Dealing with these separately is supported by the failure analysis of Edinger et al. [3] and such methods have previously been shown to work on negated and family history content in medical IR [5]. An alternative to the document perspective is to deal with temporality from the term perspective. This could be done by assigning a measure of *'temporal volatility'* to a term to determine its effect. For example, for each disease, assign some measure of how long it lasts and therefore how long it would be valid. In fact, some diseases, for example appendicitis, are acute and therefore their presence in a patient record is less of an indicator of relevance than chronic diseases, for example diabetes, that are likely to affect a patient over many years. Future work will be directed toward methods to determine temporal volatility of terms and the incorporation of this within a retrieval model.

8. REFERENCES

[1] J. Aslam, M. Ekstrand-Abueg, V. Pavlu, F. Diaz, and T. Sakai. Trec 2013 temporal summarization. In *TREC*, 2013.

[2] F. Diaz and R. Jones. Using temporal profiles of queries for precision prediction. In *SIGIR*, pg. 18–24, Sheffield, UK, July 2004.

[3] T. Edinger, A. M. Cohen, S. Bedrick, K. Ambert, and W. Hersh. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC medical records track. In *AMIA*, volume 2012, pg. 180–188, Washinton D.C., USA, 2012.

[4] B. Koopman. *Semantic Search as Inference: Applications in Health Informatics*. PhD thesis, Queensland University of Technology, Brisbane, 2014.

[5] B. Koopman and G. Zuccon. Understanding negation and family history to improve clinical information retrieval. In *SIGIR*, Gold Coast, Australia, July 2014.

[6] B. Koopman and G. Zuccon. Why assessing relevance in medical IR is demanding. In *MedIR*, Gold Coast, Australia, July 2014.

[7] N. Limsopatham, C. Macdonald, and I. Ounis. Aggregating Evidence from Hospital Departments to Improve Medical Records Search. In *ECIR*, pg. 279–291, 2013.

[8] E. M. Voorhees and W. Hersh. Overview of the TREC 2012 Medical Records Track. In *TREC*, 2012.

[9] E. M. Voorhees and R. M. Tong. Overview of the TREC 2011 Medical Records Track. In *TREC*, Gaithersburg, Maryland, USA, Nov. 2011.

[10] D. Zhu and B. Carterette. Combining multi-level evidence for medical record retrieval. In *Workshop on Smart Health and Wellbeing*, pg. 49–56, 2012.

[11] D. Zhu and B. Carterette. An adaptive evidence weighting method for medical record search. In *SIGIR*, pg. 1025–1028, Dublin, Ireland, 2013.

[12] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt. Exploiting Medical Hierarchies for Concept-based Information Retrieval. In *ADCS*, pg. 111–114, 2012.