Learning to Select a Time-aware Retrieval Model
Nattiya Kanhabua*, Klaus Berberich* and Kjetil Nørvåg†

*L3S Research Center, Max-Planck Institute for Informatics, †Norwegian University of Science and Technology

Motivation

Problem Statement:
- The retrieval effectiveness of temporal queries can be significantly improved by modeling time into ranking [4].
- Time-aware retrieval models exploit two time dimensions; either publication time (when a document was published) or content time (i.e., what time a document refers to).
- Which time dimension is factored into ranking results makes a huge difference for a given query, as observed empirically.

Contributions:
- A machine learning approach to select the most suitable time-aware retrieval mode using three classes of features.
- Experiments on real-world data with crowdsourced relevance assessments show the potential of our approach.

Features

Temporal KL-divergence [3] measures the difference of the distribution over time of top-k results $D_q$ and the collection $C$.

$$KL(T_d|C,q) = \sum_{t \in T} P(t|q) \cdot \log \frac{P(t|q)}{P(t|T_c)}$$

$$P(t|q) = \sum_{d \in D_q} P(t|d) \cdot \sum_{d \in D_q} P(|q|p|d) \cdot P(|d|) = \frac{c(t,d)}{\sum_{t \in T_{d}} c(t,d)}$$

- PT(t|T_d) is the probability of $t$ in $C$. $T_d$ are all temporal expressions in $C$.
- PT(t|q) is the probability of generating $t$ given $q$.
- P(q|d) is a retrieval score of $q$ wrt a particular retrieval model.
- P(t|d) is the probability of generating $t$ in $d$. $c(t,d)$ is the frequency of $t$ in $d$.

Clarity score [2] measures the difference of the distribution over terms of top-k results $D_q$ and the collection $C$.

$$Clarity = \sum_{w \in V} P(w|q) \cdot \log \frac{P(w|q)}{P(w|C)}$$

- $P(w|q)$ is the probability of generating $w$ given $q$ and $w$ is a term from the vocabulary $V$. $P(w|C)$ is the probability of $w$ in $C$.

Retrieval scores [6] analyze/compare the scores of a term-based baseline model that is not time-aware with PT-Rank and CT-Rank.

1. average retrieval scores of the baseline, PT-Rank or CT-Rank.
2. the divergence of scores of PT-Rank and CT-Rank from the baseline.

$$JS(S_q,S_d) = \sum_{d \in D_q} S_q(t,d) \cdot \log \frac{S_q(t,d)}{\frac{1}{2} \cdot S_q(t,d) + \frac{1}{2} \cdot S_d(t,d)}$$

$$S_q(t,d)$$ is the retrieval score of $d$ according to the baseline $S_q$.

$$S_d(t,d)$$ is the retrieval score of $d$ when ranked using a time-aware retrieval model $S_d$ (PT-Rank, CT-Rank).

Table 1: Accuracy of query classification.

<table>
<thead>
<tr>
<th>Feature</th>
<th>inclusive</th>
<th>exclusive</th>
<th>inclusive</th>
<th>exclusive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarity</td>
<td>0.84</td>
<td>0.83</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>$KL_T$</td>
<td>0.53</td>
<td>0.53</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>$KL_C$</td>
<td>0.53</td>
<td>0.53</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>AVG$C$</td>
<td>0.53</td>
<td>0.53</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>AVG$T$</td>
<td>0.53</td>
<td>0.53</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>AVG$C_T$</td>
<td>0.53</td>
<td>0.53</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.54</td>
<td>0.54</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>$JS_{PT}$</td>
<td>0.42</td>
<td>0.42</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>$JS_{CT}$</td>
<td>0.42</td>
<td>0.42</td>
<td>0.65</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Conclusions:
- Selecting the right time-aware retrieval model can have a significant impact on the effectiveness of temporal queries.
- We proposed a machine learning method for automatically selecting a retrieval model and evaluated through extensive experiments.