A Probabilistic Framework for Time Sensitive Search

MPII at the NTCIR-12 Temporalia-2 Task

Dhruv Gupta
dhgupta@mpi-inf.mpg.de

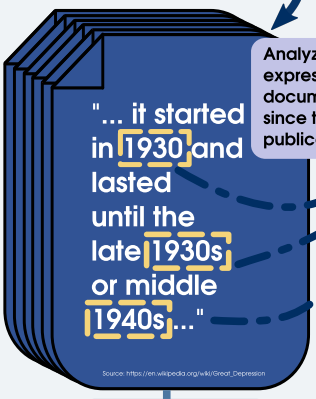
Klaus Berberich
kberberi@mpi-inf.mpg.de

Objective | For a given keyword query, identify a discrete probability distribution over different temporal classes and use this for diversifying documents.

economic depression

Temporal Intent Disambiguation | Given, classes $C = past, recent, future, atemporal$ and keyword query q_{text}, estimate $P(C|q_{text})$.

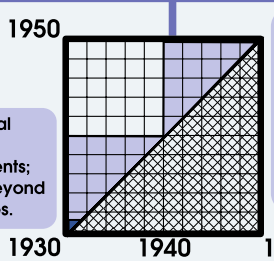
Step 1. Retrieve pseudo-relevant documents.



Analyze temporal expressions in document contents; since they go beyond publication dates.

Pseudo-relevant Documents (R)

Time Model - Understand uncertainty behind temporal expressions, e.g. 1930s, by associating lower & upper bounds on begin and end.



Uncertainty-Aware Time Model

Step 2. Identify interesting time intervals to query (q_{time}) by analyzing the temporal expressions.

$$P(C = past|q) = \frac{1}{|\hat{q}_{time}|} \sum_{[b,e]\in\hat{q}_{time}} \mathbf{1}(t_{issue} > e)$$

$$P(C = recent|q) = \frac{1}{|\hat{q}_{time}|} \sum_{[b,e]\in\hat{q}_{time}} \mathbf{1}(b \le t_{issue} \le e)$$

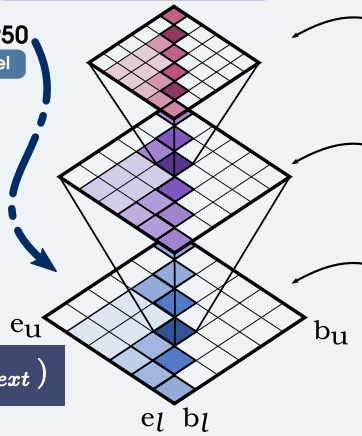
$$P(C = future|q) = \frac{1}{|\hat{q}_{time}|} \sum_{[b,e]\in\hat{q}_{time}} \mathbf{1}(t_{issue} < b)$$

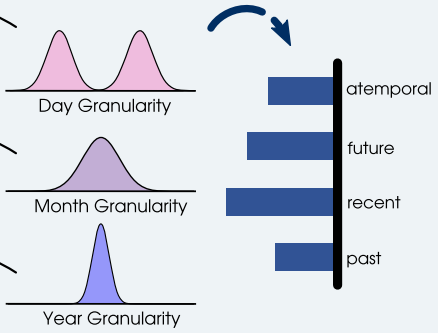
$$P(C = atemporal|q) = \sqrt{|\hat{q}_{time}|} \max_{[b,e]\in\hat{q}_{time}} |P([b,e]|q) - P([b,e]|D_{time})|$$

Time Intervals of Interest - Two step generative model. A time interval is deemed *interesting* if it's referred frequently by highly relevant documents.

$P(d_{text} \mid q_{text})$ - Estimates document likelihoods.

$P([b,e] \mid d_{time})$ - Likelihood of generating time interval from temporal expressions in document.





Generate a discrete probability distribution over the four different temporal classes by analyzing the unit length time interval probability distribution w.r.t. query issue time.

$$P([b,e] \mid q_{text}) = \sum_{d\in R} P([b,e] \mid d_{time}) P(d_{text} \mid q_{text})$$

Generative Model

economic depression 1930

Temporally Diversified Retrieval | Given, keyword query q_{text} and document collection D, estimate $P(d|q, C)$.

Step 3. Re-rank documents for different temporal categories.

$$P(q|d) = P(q_{text}|d_{text}) \cdot P(q_{time}|d_{time})$$

$$P(q_{time}|d_{time}) = \prod_{[b,e]\in q_{time}} P([b,e]|d_{time})$$

Temporal Language Model

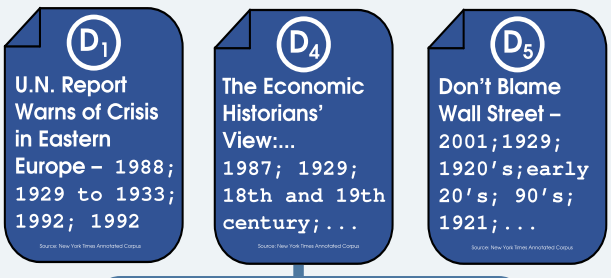
- Use the temporal language model to re-rank documents.
- For *recent*, expand query with query issue time.
- For *past*, expand query with time intervals that lie before query issuing time.
- For *future*, expand query with time intervals that lie after query issuing time.
- For *atemporal*, use the pseudo-relevant set of documents.

Step 4. Diversify documents using q_{time} as aspects.

$$\sum_{[b,e]\in q_{time}} \left(P([b,e] \mid q_{text}) \left(1 - \prod_{d\in S} \left(1 - P(q_{text} \mid d_{text}) P([b,e] \mid d_{time})\right)\right)\right)$$

Objective Function

[1990,1995] [1987,1987] [1930,1930]



Diversified Documents (S)

Time-Diversify - The objective is to maximize the probability that the user sees at least one result relevant to her time interval of interest.

$P([b,e] \mid q_{text})$ - Salience of time interval.

$P(q_{text} \mid d_{text}) P([b,e] \mid d_{time})$ - Relevance of document for the time interval.

MAX-PLANCK-GESELLSCHAFT

max planck institut informatik