Determining Time of Queries for Re-ranking Search Results

Nattiya Kanhabua and Kjetil Nørvåg

Database System Group
Norwegian University of Science and Technology
Trondheim, Norway

ECDL'2010, September 6 - 9, Glasgow, Scotland

Outline

Outline

Outline

Outline

Outline

Temporal IR

What is temporal IR?

- searching temporal document collections
- such as digital libraries, web archives and news repositories
- especially historians, librarians, journalists, and students

Temporal IR

What are challenges?

Semantic gaps in temporal IR: lacking knowledge about

1. terminology changes over time
2. possible relevant time of queries

Temporal IR

What are challenges?

Semantic gaps in temporal IR: lacking knowledge about

1. **terminology changes over time**
2. possible relevant time of queries

Terminology changes over time

Queries composed of named entities (people, organization, location)

- very dynamic in appearance, i.e., relationships between terms changes over time
- e.g. changes of roles, name alterations, or semantic shift

Terminology changes over time

Queries composed of named entities (people, organization, location)

- very dynamic in appearance, i.e., relationships between terms changes over time
- e.g. changes of roles, name alterations, or semantic shift

Scenario 1

- Query: **"Pope Benedict XVI"** and written *before 2005*
- Documents about **"Joseph Alois Ratzinger"** are relevant

Terminology changes over time

Queries composed of named entities (people, organization, location)

- very dynamic in appearance, i.e., relationships between terms changes over time
- e.g. changes of roles, name alterations, or semantic shift

Scenario 2

- Query: **"Hillary R. Clinton"** and written *from 1997 to 2002*
- Documents about **"New York Senator"** and **"First Lady of the United States"** are relevant

Terminology changes over time

Queries composed of named entities (people, organization, location)

- very dynamic in appearance, i.e., relationships between terms changes over time
- e.g. changes of roles, name alterations, or semantic shift

Our proposed approaches

"Exploit time-based synonyms in searching document archives" [JCDL'2010]

- Automatically extract synonyms over time from Wikipedia snapshots
- Expand a query using time-based synonyms to improve the accuracy

Temporal IR (cont')

What are challenges?

Semantic gaps in temporal IR: lacking knowledge about

1. terminology changes over time
2. **possible relevant time of queries**

Temporal IR (cont')

What are challenges?

Semantic gaps in temporal IR: lacking knowledge about

1. terminology changes over time
2. **possible relevant time of queries**

Relevant time of query "tsunami"

1900s

- 1960: Valdivia, Chile
- 1964: Alaska, USA
- 1993: Hokkaido, Japan
- 1998: Papua New Guinea

2000s

- 2004: Indian Ocean
- 2007: Solomon Island
- 2009: Samoa, Pacific Ocean
- 2010: Chile

Temporal IR (cont')

What are challenges?

Semantic gaps in temporal IR: lacking knowledge about

1. terminology changes over time
2. **possible relevant time of queries**

Problem

- temporal queries that comprise *only keywords*
- difficult to achieve high accuracy using only keywords
- relevant documents are associated to particular time not given by the queries

Problem statement

- Time-dependent queries exist in both standard collections and the Web [Li and Croft 2003; Diaz and Jones 2004]
  - relevancy is dependent on time
  - documents are about events at a particular time period

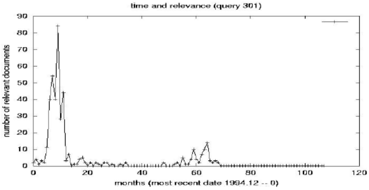


Figure 2.2: Query 301 "International Organized Crime" – A "recency" query.

"Recency query"

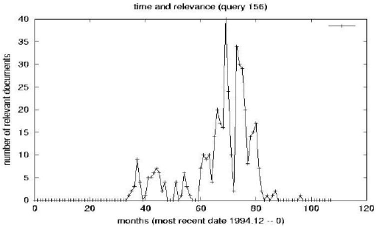


Figure 2.3: Query 156 "Efforts to Enact Gun Control Legislation"- Relevant documents mostly in the past.

"Time-dependent query"

Problem statement

- Time-dependent queries exist in both standard collections and the Web [Li and Croft 2003; Diaz and Jones 2004]
  - relevancy is dependent on time
  - documents are about events at a particular time period

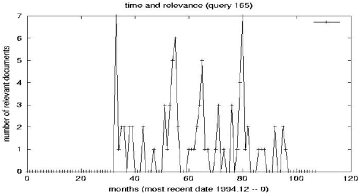


Figure 2.4: Query 165 "Tobacco Company Advertising and the Young" - More uniform distribution

"Time-independent query"

Problem statement

- 1.5% of web queries are explicitly provided with temporal expression [Nunes et al. 2008]
  - time is a part of query, “U.S. Presidential election **2008**”
- about 7% of web queries have temporal intent implicitly provided [Metzler et al. 2009]
  - time is not a part of query, “Germany World Cup”

Outline

Contributions

1. Formal models
   - temporal document models
   - temporal query models
   - temporal language models
2. Proposed approaches
   - determining the time of queries when *no* temporal criteria provides
   - re-ranking search results using the determined time
3. Experiments
   - evaluating our approach to determining the time of queries
   - evaluating our approach to re-ranking search results

Contributions

1. **Formal models**
   - temporal document models
   - temporal query models
   - temporal language models
2. **Proposed approaches**
   - determining the time of queries when *no* temporal criteria provides
   - re-ranking search results using the determined time
3. **Experiments**
   - evaluating our approach to determining the time of queries
   - evaluating our approach to re-ranking search results

Contributions

1. Formal models
   - temporal document models
   - temporal query models
   - temporal language models
2. Proposed approaches
   - determining the time of queries when *no* temporal criteria provides
   - re-ranking search results using the determined time
3. Experiments
   - evaluating our approach to determining the time of queries
   - evaluating our approach to re-ranking search results

Outline

Formal models

- Collection contains corpus documents $C = \{d_1, \ldots, d_n\}$
- Document d_i consists of bag-of-words and a creation date
  - $d_i = \{\{w_1, \ldots, w_n\}, Time(d_i)\}$, where $Time(d_i)$ is timestamp
  - $[t_k, t_{k+1}]$ is the associated time partition of d_i

Example

- partition the collection C with the *1-month* granularity
- the document timestamp $Time(d_i)$ is 05/03/2010
- the associated time partition of d_i is $Time(d_i) \in$ [01/03/2010,31/03/2010]

Formal models

- Collection contains corpus documents $C = \{d_1, \ldots, d_n\}$
- Document d_i consists of bag-of-words and a creation date
  - $d_i = \{\{w_1, \ldots, w_n\}, Time(d_i)\}$, where $Time(d_i)$ is timestamp
  - $[t_k, t_{k+1}]$ is the associated time partition of d_i

Example

- partition the collection C with the *1-month* granularity
- the document timestamp $Time(d_i)$ is 05/03/2010
- the associated time partition of d_i is $Time(d_i) \in$ [01/03/2010,31/03/2010]

Formal models

- Collection contains corpus documents $C = \{d_1, \ldots, d_n\}$
- Document d_i consists of bag-of-words and a creation date
  - $d_i = \{\{w_1, \ldots, w_n\}, Time(d_i)\}$, where $Time(d_i)$ is timestamp
  - $[t_k, t_{k+1}]$ is the associated time partition of d_i

Example

- partition the collection C with the *1-month* granularity
- the document timestamp $Time(d_i)$ is 05/03/2010
- the associated time partition of d_i is $Time(d_i) \in$ [01/03/2010,31/03/2010]

Formal models

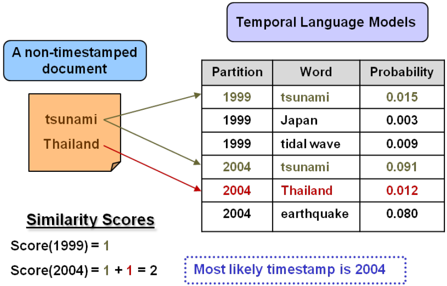
- Temporal query q composed of two parts:
  - keywords $q_{word} = \{w_1, \ldots, w_m\}$
  - temporal criteria $q_{time} = \{t'_1, \ldots, t'_l\}$, where $t'_j = [t_j, t_{j+1}]$

Example

- “Boxing Day tsunami” q_{time} = {[01/01/2004,31/12/2004]}
- “the U.S. presidential election”
  q_{time} = {[01/01/2000, 31/12/2000],[01/01/2004, 31/12/2004],[01/01/2008, 31/12/2008]}

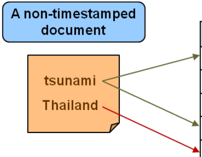
Model for dating documents

- Temporal Language Models in [de Jong, Rode and Hiemstra 2005]
- Assign a probability to a time partition according to word usage/statistics over time
- *The determined time* is a partition maximizes a score (mostly overlaps in terms)



Model for dating documents

- Temporal Language Models in [de Jong, Rode and Hiemstra 2005]
- Assign a probability to a time partition according to word usage/statistics over time
- *The determined time* is a partition maximizes a score (mostly overlaps in terms)



Temporal Language Models

Partition	Word	Probability
1999	tsunami	0.015
1999	Japan	0.003
1999	tidal wave	0.009
2004	tsunami	0.091
2004	Thailand	0.012
2004	earthquake	0.080

Similarity Scores

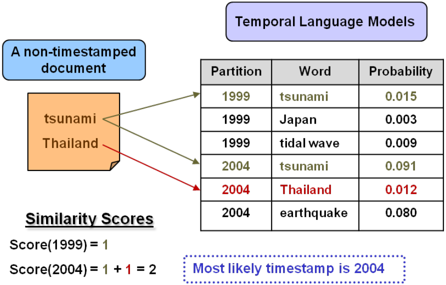
Score(1999) = 1

Score(2004) = 1 + 1 = 2

Most likely timestamp is 2004

Model for dating documents

- Temporal Language Models in [de Jong, Rode and Hiemstra 2005]
- Assign a probability to a time partition according to word usage/statistics over time
- *The determined time* is a partition maximizes a score (mostly overlaps in terms)



Partition	Word	Probability
1999	tsunami	0.015
1999	Japan	0.003
1999	tidal wave	0.009
2004	tsunami	0.091
2004	Thailand	0.012
2004	earthquake	0.080

Compute a similarity score

- Normalized log-likelihood ratio [Kraaij 2005]
  - a normalized variant of Kullback-Leibler divergence
  - measure similarity between two language models: non-timestamped document and a reference corpus

$$Score(d_i, p_j) = \sum_{w \in d_i} P(w|d_i) \times \log \frac{P(w|p_j)}{P(w|C)}$$

  - C is the background model estimated on the collection
  - linear interpolation smoothing to avoid the zero probability of unseen words

Compute a similarity score

- Normalized log-likelihood ratio [Kraaij 2005]
  - a normalized variant of Kullback-Leibler divergence
  - measure similarity between two language models: non-timestamped document and a reference corpus

$$Score(d_i, p_j) = \sum_{w \in d_i} P(w|d_i) \times \log \frac{P(w|p_j)}{P(w|C)}$$

  - C is the background model estimated on the collection
  - linear interpolation smoothing to avoid the zero probability of unseen words

Compute a similarity score

- Normalized log-likelihood ratio [Kraaij 2005]
  - a normalized variant of Kullback-Leibler divergence
  - measure similarity between two language models: non-timestamped document and a reference corpus

$$Score(d_i, p_j) = \sum_{w \in d_i} P(w|d_i) \times \log \frac{P(w|p_j)}{P(w|C)}$$

  - C is the background model estimated on the collection
  - linear interpolation smoothing to avoid the zero probability of unseen words

Outline

Proposed approaches

- Approach I. Dating query using keywords
- Approach II. Dating a query using *top-k* documents
  - in general, queries are short
  - inspired by pseudo-relevance feedback
- Approach III. Using timestamp of top-k documents
  - no temporal language models are used

Proposed approaches

- Approach I. Dating query using keywords
- Approach II. Dating a query using *top-k* documents
  - in general, queries are short
  - inspired by pseudo-relevance feedback
- Approach III. Using timestamp of top-k documents
  - no temporal language models are used

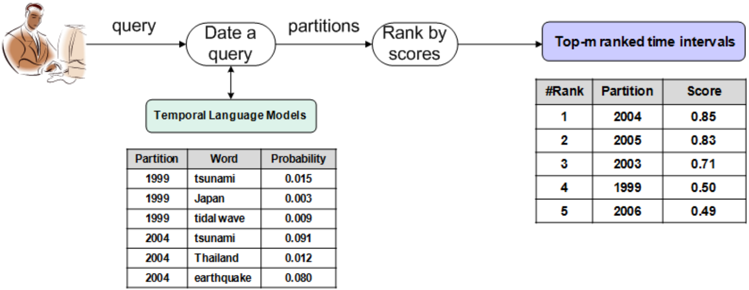
Proposed approaches

- Approach I. Dating query using keywords
- Approach II. Dating a query using *top-k* documents
  - in general, queries are short
  - inspired by pseudo-relevance feedback
- Approach III. Using timestamp of top-k documents
  - no temporal language models are used

Proposed approaches

- Approach I. Dating query using keywords
- Approach II. Dating a query using *top-k* documents
  - in general, queries are short
  - inspired by pseudo-relevance feedback
- Approach III. Using timestamp of top-k documents
  - no temporal language models are used

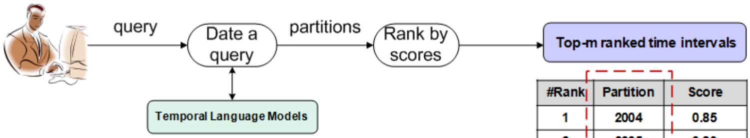
Approach I. Dating query using keywords



Partition	Word	Probability
1999	tsunami	0.015
1999	Japan	0.003
1999	tidal wave	0.009
2004	tsunami	0.091
2004	Thailand	0.012
2004	earthquake	0.080

#Rank	Partition	Score
1	2004	0.85
2	2005	0.83
3	2003	0.71
4	1999	0.50
5	2006	0.49

Approach I. Dating query using keywords

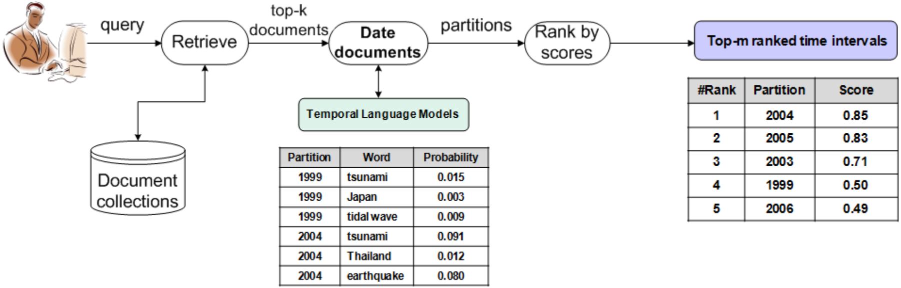


Partition	Word	Probability
1999	tsunami	0.015
1999	Japan	0.003
1999	tidal wave	0.009
2004	tsunami	0.091
2004	Thailand	0.012
2004	earthquake	0.080

#Rank	Partition	Score
1	2004	0.85
2	2005	0.83
3	2003	0.71
4	1999	0.50
5	2006	0.49

Query's temporal profile

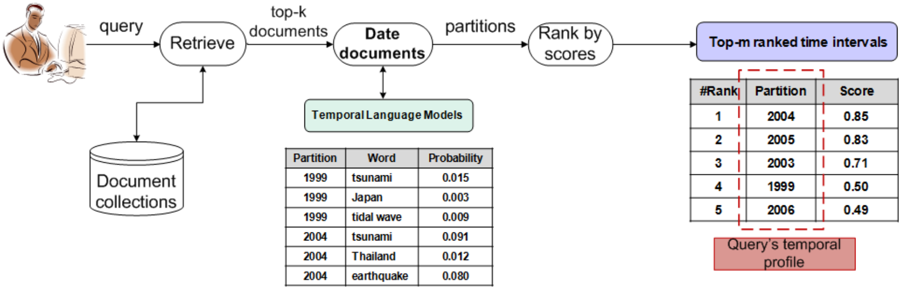
Approach II. Dating a query using *top-k* documents



Partition	Word	Probability
1999	tsunami	0.015
1999	Japan	0.003
1999	tidal wave	0.009
2004	tsunami	0.091
2004	Thailand	0.012
2004	earthquake	0.080

#Rank	Partition	Score
1	2004	0.85
2	2005	0.83
3	2003	0.71
4	1999	0.50
5	2006	0.49

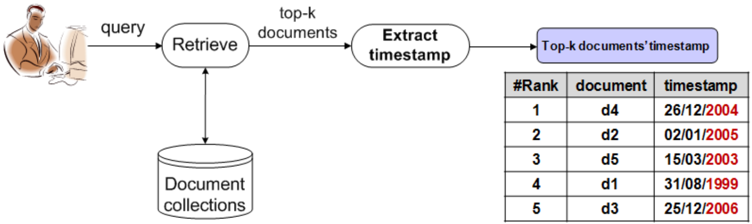
Approach II. Dating a query using *top-k* documents



Partition	Word	Probability
1999	tsunami	0.015
1999	Japan	0.003
1999	tidal wave	0.009
2004	tsunami	0.091
2004	Thailand	0.012
2004	earthquake	0.080

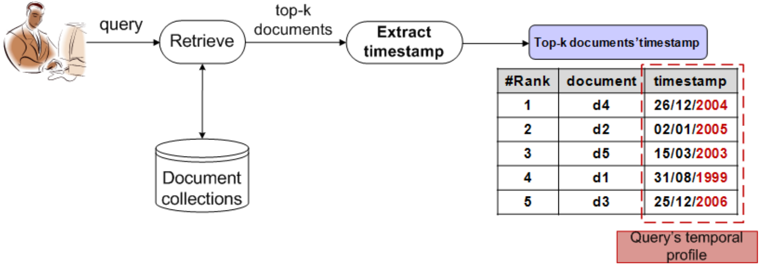
#Rank	Partition	Score
1	2004	0.85
2	2005	0.83
3	2003	0.71
4	1999	0.50
5	2006	0.49

Approach III. Using timestamp of top-k documents



#Rank	document	timestamp
1	d4	26/12/2004
2	d2	02/01/2005
3	d5	15/03/2003
4	d1	31/08/1999
5	d3	25/12/2006

Approach III. Using timestamp of top-k documents



#Rank	document	timestamp
1	d4	26/12/2004
2	d2	02/01/2005
3	d5	15/03/2003
4	d1	31/08/1999
5	d3	25/12/2006

Outline

Re-ranking search result using the determined time

Intuition: documents with creation dates that closely match with the *implicit* time of queries are more relevant

- a mixture model of a keyword score and a time score

Definition

$$S(q,d) = (1-\alpha) \cdot S'(q_{word}, d_{word}) + \alpha \cdot S''(q_{time}, d_{time})$$

α underlining the importance of a keyword score and a time score

Re-ranking search result using the determined time

Intuition: documents with creation dates that closely match with the *implicit* time of queries are more relevant

- a mixture model of a keyword score and a time score

Definition

$$S(q, d) = (1 - \alpha) \cdot S'(q_{word}, d_{word}) + \alpha \cdot S''(q_{time}, d_{time})$$

α underlining the importance of a keyword score and a time score

Re-ranking search result using the determined time

Definition

$$S(q,d) = (1-\alpha)\cdot S'(q_{word}, d_{word}) + \alpha \cdot S''(q_{time}, d_{time}) \quad (1)$$

$$\begin{aligned} S''(q_{time}, d_{time}) &= P(q_{time}|d_{time}) \\ &= P(\{t'_1, \ldots, t'_n\}\,|d_{time}) \\ &= \frac{1}{|q_{time}|} \sum_{t'_j \in q_{time}} P(t'_j|d_{time}) \end{aligned} \quad (2)$$

where q_{time} is a set of time intervals and $(t'_1 \cap t'_2 \cap \ldots \cap t'_n) = \emptyset$

Re-ranking search result using the determined time

Definition

1. $P(t'_j|d_{time})$ with uncertainty-ignorant:

$$P(t'_j|d_{time}) = \begin{cases} 0 & \text{if } d_{time} \neq t'_j, \\ 1 & \text{if } d_{time} = t'_j. \end{cases} \quad (1)$$

2. $P(t'_j|d_{time})$ with uncertainty-aware:

$$P(t'_j|d_{time}) = DecayRate^{\lambda \cdot |t'_j - d_{time}|} \quad (2)$$

DecayRate and λ are constants, $0 < DecayRate < 1$ and $\lambda > 0$

Outline

Overview of experiments

Our experimental evaluation is divided into two parts:

1. Determining the time of queries
2. Re-ranking search results using the determined time

Determining the time of queries

Temporal document collection:

- New York Time Annotated Corpus contains over 1.8 million articles from January 1987 to June 2007

Tools:

- Oracle Berkeley DB version 4.7.25

Queries:

- randomly selected 30 strongly time-related queries from the Robust2004

Parameters: $m = 5$, g and k are varied

Measurement: precision, recall and F_2

Re-ranking of search results

Data collection:

- TREC Robust Track (2004)
  - 30 strongly time-related topics
- New York Time Annotated Corpus
  - 24 queries from the Google zeitgeist

Tools:

- Terrier – an open source search engine developed by University of Glasgow
- BM25 probabilistic model with Generic Divergence From Randomness (DFR) weighting
- Alter scores for retrieved documents by giving prior scores

$$S''(q_{time}, d_{time}) = P(q_{time}|d_{time})$$

Parameters: $DecayRate = 0.5$, $\lambda = 0.5$, $\alpha = 0.05$ for *uncertainty-ignore*, $\alpha = 0.10$ for *uncertainty-aware*

Measurement: MAP, R-precision, P@5, P@10, and P@15

Re-ranking of search results

Examples of the Google zeitgeist queries and associated time periods

Query	Time	Query	Time
diana car crash	1997	madrid bombing	2005
world trade center	2001	pope john paul ii	2005
osama bin laden	2001	tsunami	2005
london congestion charges	2003	germany soccer world cup	2006
john kerry	2004	torino games	2006
tsa guidelines liquids	2004	subprime crisis	2007
athens olympics games	2004	obama presidential campaign	2008

Outline

Performance of query dating methods

Table: Query dating performance using precision, recall and F-score

Method	Precision		Recall		F_2	
	6-month	*12-month*	*6-month*	*12-month*	*6-month*	*12-month*
QW	.56	.67	.34	.64	.37	.65
PRF *(k=5)*	.55	.63	.47	**.79**	.48	**.75**
PRF *(k=10)*	.56	.60	.46	.74	.48	.71
PRF *(k=15)*	.54	.60	.42	.70	.44	.68
NLM *(k=5)*	.92	**.97**	.35	.44	.40	.49
NLM *(k=10)*	.90	.95	.48	.56	.53	.61
NLM *(k=15)*	.89	.93	.56	.63	.61	.67

- **QW** determines time using keywords *plus* uncertainty-ignorant re-ranking
- **PRF** determines time using top-k retrieved documents *plus* uncertainty-ignorant re-ranking
- **NLM** assumes creation dates of top-k documents (no language models) *plus* uncertainty-ignorant re-ranking

Performance of re-ranking methods

Table: Re-ranking performance with the baseline performance 0.3568 and 0.3909 respectively (the Robust2004 collection)

Method	MAP		R-precision	
	6-month	*12-month*	*6-month*	*12-month*
QW	.3565	.3576	.3897	.3924
QW-U	.3556	.3573	.3925	.3943
PRF *(k=5)*	.3564	.3570	.3885	.3926
PRF *(k=10)*	.3568	.3570	.3913	.3919
PRF *(k=15)*	.3566	.3567	.3912	.3921
PRF-U *(k=5)*	.3548	.3574	.3903	.3950
PRF-U *(k=10)*	.3538	.3576	.3904	.3935
PRF-U *(k=15)*	.3538	.3572	.3893	.3940
NLM *(k=5)*	.3585	.3589	.3924	.3917
NLM *(k=10)*	.3586	.3591	.3918	.3925
NLM *(k=15)*	.3584	.3596	.3898	.3934
NLM-U *(k=5)*	.3604	.3608	.3975	**.3978**
NLM-U *(k=10)*	.3604	.3610	.3953	.3961
NLM-U *(k=15)*	.3606	**.3620**	.3943	.3967

QW-U, **PRF-U**, **NLM-U** determines time using *uncertainty-aware* re-ranking

Performance of re-ranking methods

Table: Re-ranking performance using P@5, P@10, and P@15 with the baseline performance 0.35, 0.30 and 0.27 (the NYT collection)

Method	P@5		P@10		P@15	
	6-month	*12-month*	*6-month*	*12-month*	*6-month*	*12-month*
QW	.42	.45	.37	.39	.32	.33
QW-U	.40	.42	.35	.36	.30	.32
PRF *(k=15)*	.42	.46	.38	.42	.35	.39
PRF-U *(k=15)*	.41	.45	.36	.40	.33	.37
NLM *(k=15)*	.50	.52	.47	.49	.42	.44
NLM-U *(k=15)*	.53	**.55***	.48	**.50***	.45	**.46***

Note: * indicates statistically improvement over the baselines using *t-test* ($p < 0.05$)

Outline

Conclusions and future work

- Study implicit temporal queries (no temporal criteria)
- Determine the implicit time of the queries
- Employ the determined time to re-rank the search results
- Conduct extensive experiments and show the improvement in retrieval effectiveness
- Future work:
  - The quality of the query dating is limited when aiming at further increase in effectiveness
  - Improvement on the query dating based on external knowledge from sources like Wikipedia

Conclusions and future work

- Study implicit temporal queries (no temporal criteria)
- Determine the implicit time of the queries
- Employ the determined time to re-rank the search results
- Conduct extensive experiments and show the improvement in retrieval effectiveness
- Future work:
  - The quality of the query dating is limited when aiming at further increase in effectiveness
  - Improvement on the query dating based on external knowledge from sources like Wikipedia

Conclusions and future work

- Study implicit temporal queries (no temporal criteria)
- Determine the implicit time of the queries
- Employ the determined time to re-rank the search results
- Conduct extensive experiments and show the improvement in retrieval effectiveness
- Future work:
  - The quality of the query dating is limited when aiming at further increase in effectiveness
  - Improvement on the query dating based on external knowledge from sources like Wikipedia

Conclusions and future work

- Study implicit temporal queries (no temporal criteria)
- Determine the implicit time of the queries
- Employ the determined time to re-rank the search results
- Conduct extensive experiments and show the improvement in retrieval effectiveness
- Future work:
  - The quality of the query dating is limited when aiming at further increase in effectiveness
  - Improvement on the query dating based on external knowledge from sources like Wikipedia

Conclusions and future work

- Study implicit temporal queries (no temporal criteria)
- Determine the implicit time of the queries
- Employ the determined time to re-rank the search results
- Conduct extensive experiments and show the improvement in retrieval effectiveness
- Future work:
  - The quality of the query dating is limited when aiming at further increase in effectiveness
  - Improvement on the query dating based on external knowledge from sources like Wikipedia

Conclusions and future work

- Study implicit temporal queries (no temporal criteria)
- Determine the implicit time of the queries
- Employ the determined time to re-rank the search results
- Conduct extensive experiments and show the improvement in retrieval effectiveness
- Future work:
  - The quality of the query dating is limited when aiming at further increase in effectiveness
  - Improvement on the query dating based on external knowledge from sources like Wikipedia

Thank you. Question?