NTNU
Norwegian University of Science and Technology

Exploiting Temporal Information in Retrieval of Archived Documents

Nattiya Kanhabua
(Supervised by Prof. Kjetil Nørvåg)

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

Agenda

- Motivation
- Research Questions
- Proposed Approaches
- Previous and Current work
- Issues for Discussion

Records Management over Decades

- Initiated by *Det Norske Veritas* with initial motivations:
  - Long-term* digital storage of ship drawings
  - Validation and notary services (trusted third-party roles taken by DNV)
  - Certification of maturity wrt. long-term management of data
  - Funding: Norwegian Research Council 1.12 M€, partners 3.36 M€

*A domain where documents are preserved for more than 10 years

- Partners including:
  - The National Library of Norway
  - The National Archives of Norway
  - Ministry of Foreign Affairs
  - Fast Search & Transfer (a Microsoft subsidiary)
  - StatoilHydro

- Objective:
  - Persistent, reliable and trustworthy long-term archival of digital information records, with emphasis on availability and use of the information
    - Enable the digital representation as the original version
    - Enable long-term usability over decades
    - Explore the potential for commercial products/services in this area

Research Questions

NTNU
Norwegian University of Science and Technology

The main research question:

"How to improve the quality of search in a document archive using temporal information?"

Q1. How to handle large number of documents retrieved?

- This decreases the accuracy of search since users have to spend more time in exploring the information needed.

Q2. How to search with awareness of language changes?

- With time, terms might change semantics, e,g, "awesome", or original terms are obsolete, e.g. "Siam" – *terminology evolution*.

Q3. How to rank search results wrt. temporal information?

- In other research, hit lists are listed *chronologically* (newer pages are more important). In some cases, a chronological order is *not* always needed.

Approach I: Handling Large Number of Documents Retrieved

Problem: The accuracy of search is decreased since users have to spend more time in exploring the information needed.

Proposed approach : Re-ranking for presentation or including a temporal relationship with respect to a query, i.e., extending keyword search with a creation or update date of documents – *temporal criteria*

strongly time-related
e.g. "tsunami" or "presidential election"

Two ways to obtain
1) provided by users, or
2) determined by the system

Challenge :

1. **How to find time related to a query and attach time to the query implicitly, and retrieve results created within that time.**
2. **Note: users have *no clue* regarding possible time of a query, thus no time can be explicitly provided for search.**

Approach II: Searching with Awareness of Language Changes

Problem: With time, terms might change semantics, e.g. “awesome”, or original terms are obsolete, e.g., “Siam” – *terminology evolution*.

Proposed approach : Use a dictionary linking **concepts and entities** ***based on time***, example for the concept of current Thailand can be defined as follows:

Thailand → Siam[0,1939]

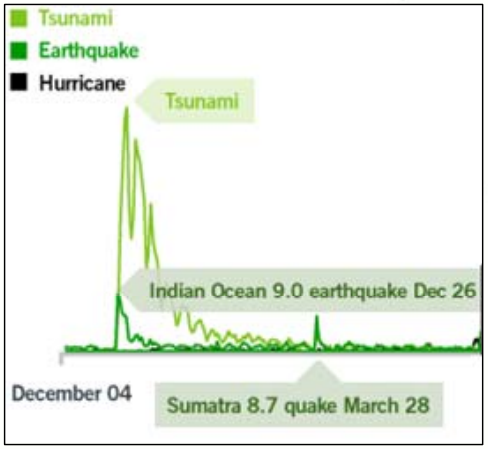
Challenge :

1. **For a query for “Thailand”, the query might be expanded to “Thailand or Siam”.**
2. **For a query for documents written at a certain time (before 1939) the query might be *rewritten* from “Thailand” to “Siam”.**
3. **How can such a time-concept table be machine-generated?**

Approach III: Temporal Ranking of Search Results

Problem: In other research, hit lists are listed *chronologically* (newer pages are more important). A chronological order is *not* always needed, e.g. taking into account topical trend.

Proposed approach : Analyze a document collection to obtain a ***topical trend*** – the trend of a topic – that can be represented as the weight of a topic over time.



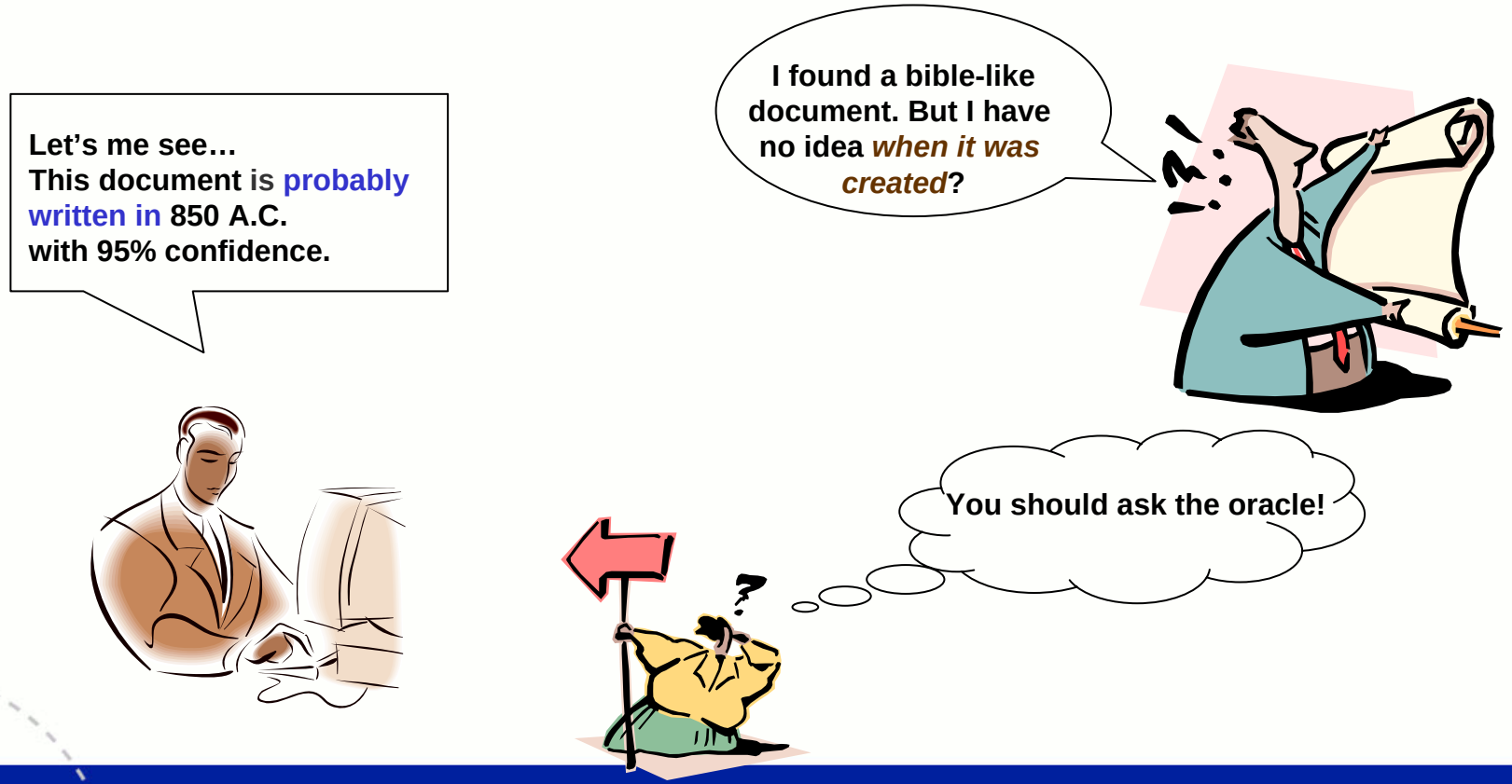
Challenge :

1. **When a query is "tsunami", documents written in 2004 should receive a higher weight than those written in other years, e.g. 2006, according to stronger relevance in the past.**
2. **The ranking of documents retrieved wrt. temporal criteria are similarity scores (term frequency) *plus* weights to topical trend.**

Overview of the ECDL 2008 Paper

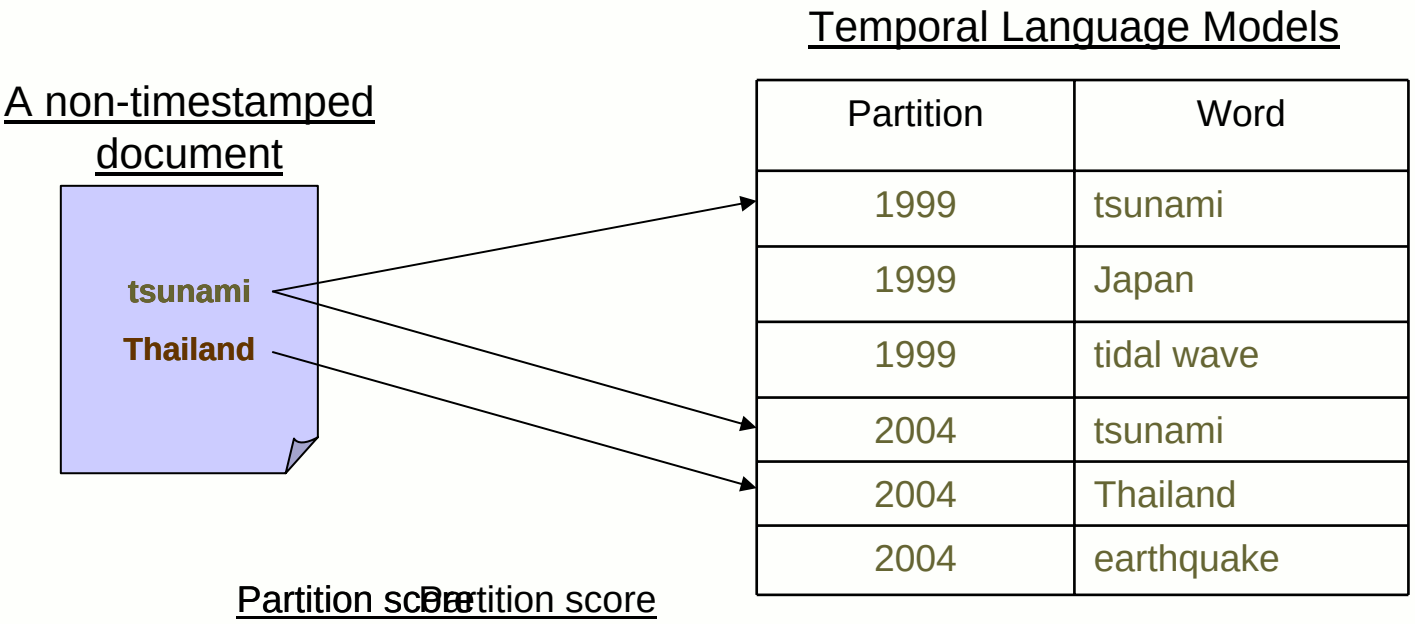
Problem: Due to decentralized nature and the lack of standards for date/time, it is difficult to find accurate and trustworthy timestamp for web documents.

“ For a given document with uncertain timestamp, can the contents be used to determine the timestamp with a sufficiently high confidence? ”



A Model for Dating Documents

- Temporal Language Models proposed by *de Jong et al.* in *AHC*'05
- Based on the statistic usage of words over time.
- Compare a non-timestamped document with a reference corpus.
- A reference time partition *mostly overlaps* in term usage -- the tentative timestamp.



Improving temporal language models

Three ways:

- Data preprocessing
  - Semantic-based preprocessing, i.e. part-of-speech tagging, collocation extraction, word sense disambiguation, concept extraction, word filtering
- Word interpolation
  - Different smoothing (*zero probability*) for 2 classes of terms depending on characteristics in time: recurring & non-recurring
- Similarity score
  - A term weighting concerns temporality, Temporal Entropy
    - How well a term is suitable for separating time partitions?
    - How important a term is in a specific time partition?
  - Taking into account external search statistics (Google Zeitgeist)
    - Integrated as an additional score to increase probabilities of time partitions.

Handling Semantic Gaps in Temporal Search

Problem Statement:

- Synonyms are *alternative words* referring to the same thing.
- When searching with a named entity (i.e. person, location or company), synonyms should be considered to improve *recall*.

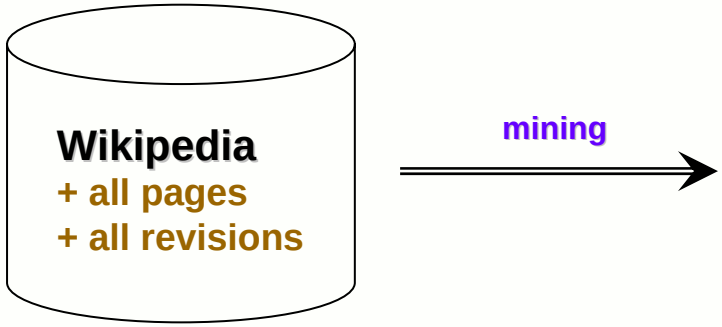
Example:

- In a temporal search, when a query is "US president", only documents exactly containing "US president" are retrieved.
- Documents about "Bush" OR "Obama" should also be retrieved since they refer to the same person.

Proposed approach :

1. **Create a time-dependent dictionary of synonyms using Wikipedia.**
2. **Use the dictionary to expand a query for improving a search quality.**

A Time-dependent Dictionary of Synonyms



Word	Synonyms	Time
US president	Clinton	1993 - 2000
	Bush	2001 - 2008
	Obama	2009 - present

How this information can be used?

- **A query "US president" can be expanded with synonyms (all former presidents) of all the times we know**
  - E.g., "US president" → "US president" **OR** "Clinton" **OR** "Bush" **OR** "Obama"
- **Alternatively, if a query is for a particular time, it can be expanded into its synonyms list as it was at that particular time**
  - E.g., "US president 1997" → "US president" **OR** "Clinton"

Issues for Discussion

Our dataset of the ECDL paper:

- **Manually crawled from the Internet Archive**
- **Web pages of news history from 15 sources, e.g. ABC News, CNN, NewYork Post.**
- **8 years on averaged for each source**
- **~ 9,000 documents (one page in every 5 days)**

1. The lack of proper dataset for evaluating the effect of changing languages
   - ➢ A corpus collection that satisfies two properties:
     1) covering more than ten years period
        - ✓ *the need of a long time-spanned dataset*
     2) evenly spread over time span
        - ✓ *the need of a well-distributed dataset*.
2. No standard test set for evaluating temporal ranking
   - ➢ This issue is critical. What we need are:
     - ✓ A set of sampled queries
     - ✓ Associated documents to each query
     - ✓ Relevance judgments of these documents

NTNU
Norwegian University of Science and Technology

References:

- K. Berberich, S. J. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. In Proceedings of SIGIR'2007, 2007.
- F. Diaz and R. Jones. Using temporal profiles of queries for precision prediction. In Proceedings of the 27th SIGIR, 2004.
- X. Li and W. B. Croft. Time-based language models. In Proceedings of CIKM, 2003.
- O. Alonso, M. Gertz and R. Baeza-Yates. On the value of temporal information in information retrieval. ACM SIGIR Forum, 41(2):35–41, 2007.
- R. Swan and D. Jensen. Timemines: Constructing timelines with statistical models of word usage. In Proceedings of KDD-2000 Workshop on Text Mining, 2000.
- P. S. Yu, X. Li, and B. Liu. On the temporal dimension of search. In Proceedings of the 13th WWW on Alternate track papers & posters. ACM, 2004.