Temporal Aspects of Web Search

Vassilis Plachouras
Yahoo! Research, Barcelona

Monday, June 18, 2007 - Bertinoro

State of Web Search

- The Web is continuously evolving over time
  - Content is created, removed, modified
  - Search engine indexes are rebuild to make fresh content searchable
- Users employ search engines to satisfy their information needs
  - Informational, Navigational & Transactional tasks (Broder, 2002)
  - Temporal patterns in query logs
- Aim to combine the temporal aspects of *Web content* and *Search Engine usage* to enhance Web search engines

Outline

Introduction

Temporal aspects of

1. Document ranking
2. Advertising
3. Search engine architectures

Wrap-up

Introduction

- Traditional Information Retrieval (IR) has a mostly static view of data sets
  - Document collections
  - Search tasks

Introduction

- TREC Filtering tasks (Robertson & Callan, 2005)
  - Online decisions according to feedback on what is relevant/non-relevant
- TREC Blog Track (Ounis *et al.*, 2006)
  - Crawl of blogs for several weeks
- Searching the future (Baeza-Yates, 2005)
  - Predict events in the future from existing information on the Web
  - Rank documents based on their content and the most probable event in them

Introduction

- Ranking of documents is mostly independent of
  - when a query is issued
  - when a document is created
  - whether the document is about a particular point in time

Introduction

- Time-based Language Models (Li & Croft, 2003)
  - Time information in language models as prior information
  - Ranking according to recency
- Also exploiting temporal information to:
  - Predict query difficulty (Diaz & Jones, 2005)
  - Perform relevance feedback and query clarification (Rode & Hiemstra, 2006)
  - Visualize search results using timelines (Alonso *et al*. 2007)

Introduction

- Temporal information is readily available for some genres of documents

- But a document may also have many associated dates or timestamps

Outline

Introduction

Temporal aspects of

1. *Document ranking*
2. Advertising
3. Search engine architectures

Wrap-up

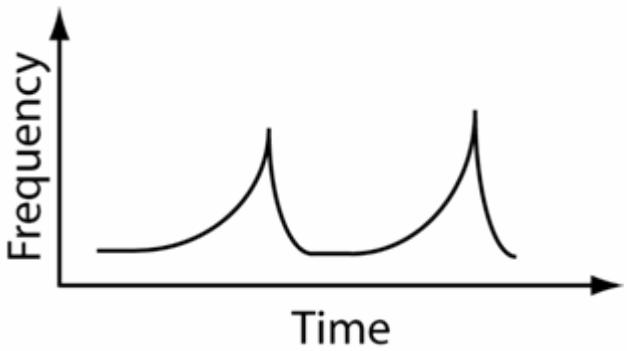
Temporal aspects of document ranking

- Temporal profiles of topics based on:
  - Queries
  - Retrieved documents

Temporal profiles based on queries

- Aggregating information from the submissions of a query

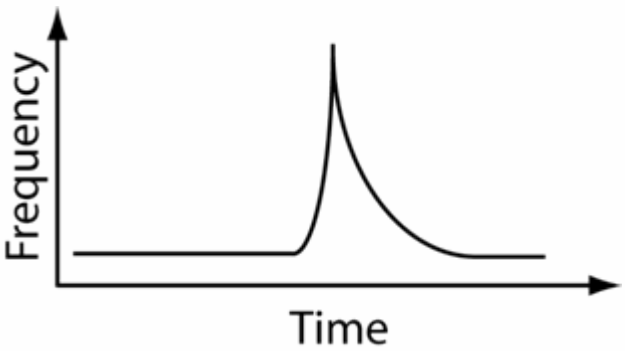
- Periodic profiles
  - Yearly (query: christmas)
  - Even longer (query: soccer World Cup)

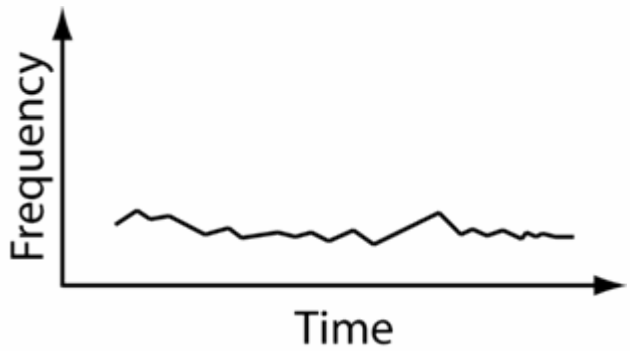


Temporal profiles based on queries

- Aggregating information from the submissions of a query

- Aperiodic profiles
  - One time/unexpected events
  - Noisy always present topics





Temporal profiles based on queries

- Detecting bursts in query time series
  - Using the Fourier coefficients of the query time series to represent and detect bursts (Vlachos *et al*., 2004)

- Quantifying similarity between queries
  - Correlation coefficient between different query time series (Chien & Immorlica, 2005)
  - Similarity of queries using click-though data and temporal information (Zhao *et al*., 2006)
  - Highly similar queries according to temporal features were not similar using session features (Liu *et al*., 2006)

Temporal profiles based on matched documents

- Distribution of retrieved document timestamps for a query (Diaz & Jones, 2004)

- Automaton-based approach to detect bursts and their intensity in streams of documents (Kleinberg, 2002)

Match profiles based on queries and documents

- How to refine the ranking of documents for a query by matching query and document profiles

- Rank higher documents that are:
  - strongly associated with the most important dates of the topic
    - Not necessarily the most recent dates
  - strongly associated with the date/time the query was submitted
    - Considering the periodicity of the query
- Combine the above two options

Outline

Introduction

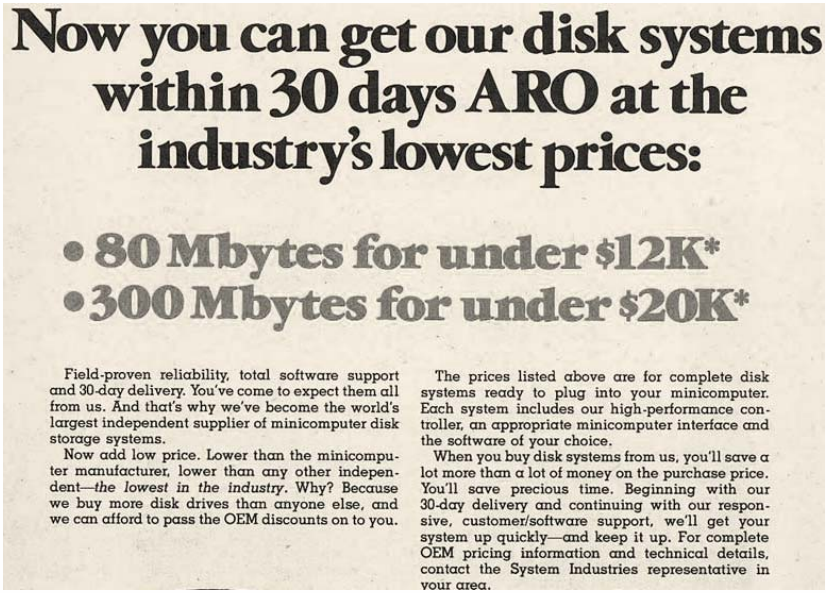
Temporal aspects of

1. Document ranking
2. *Advertising*
3. Search engine architectures

Wrap-up

Temporal aspects of advertising

- Real-world advertising can be highly temporal
  - IT ad from 1977



Now you can get our disk systems within 30 days ARO at the industry's lowest prices:

- **80 Mbytes for under $12K***
- **300 Mbytes for under $20K***

Field-proven reliability, total software support and 30-day delivery. You've come to expect them all from us. And that's why we've become the world's largest independent supplier of minicomputer disk storage systems.

Now add low price. Lower than the minicomputer manufacturer, lower than any other independent—*the lowest in the industry.* Why? Because we buy more disk drives than anyone else, and we can afford to pass the OEM discounts on to you.

The prices listed above are for complete disk systems ready to plug into your minicomputer. Each system includes our high-performance controller, an appropriate minicomputer interface and the software of your choice.

When you buy disk systems from us, you'll save a lot more than a lot of money on the purchase price. You'll save precious time. Beginning with our 30-day delivery and continuing with our responsive, customer/software support, we'll get your system up quickly—and keep it up. For complete OEM pricing information and technical details, contact the System Industries representative in your area.

- Highly temporal ad campaigns may be related to the promotion of:
  - Events : explicitly associated with dates (*e.g.* a movie)
  - Seasonal products (*e.g.* Christmas decorations)

Temporal aspects of advertising

- Online advertising could benefit by leveraging temporal correlations between user queries and customer trends

- Example: ‘Christmas decorations’
  - High interest before Christmas
  - Decreases sharply just after Christmas

- Bias the ranking of ads according to the intensity in the temporal profile of the queries
  - If users search more about a topic, bias ads towards that topic

Outline

Introduction

Temporal aspects of

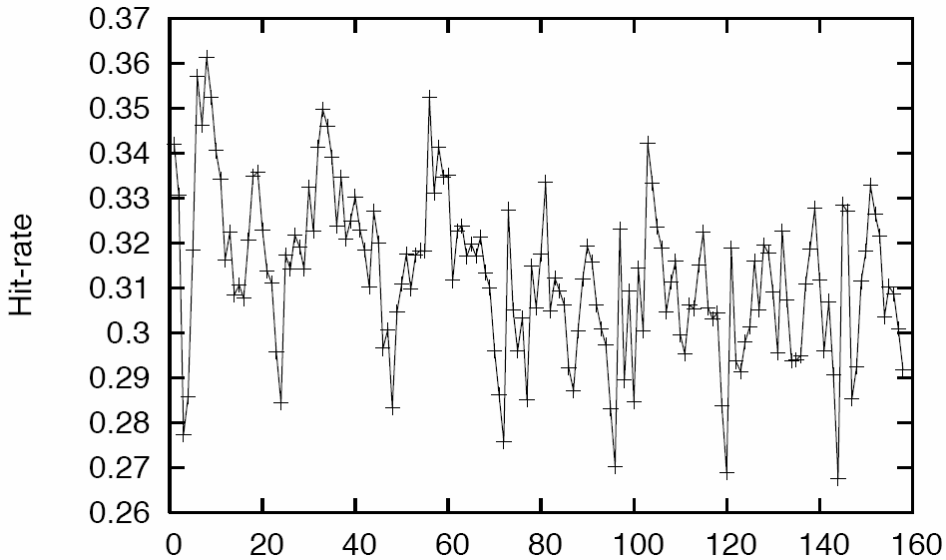
1. Document ranking
2. Advertising
3. *Search engine architectures*

Wrap-up

Temporal aspects of Web search system architectures

- System issues are mostly studied irrespectively of the temporal fluctuations of load

- Cache hit ratios depend on the time of the day (Baeza-Yates *et al.*, 2007b)



Temporal aspects of Web search system architectures

- System issues are mostly studied irrespectively of the temporal fluctuations of load

- In a fully distributed search engine (Baeza-Yates *et al*., 2007a), route queries according to the expected workload of other nodes
  - Computed from past temporal information

Outline

Introduction

Temporal aspects of

1. Document ranking
2. Advertising
3. Search engine architectures

Wrap-up

Wrap-up

- Several aspects of the engine's operation can employ temporal information
- Challenges
  - Find efficient and effective ways to incorporate temporal information
  - Accurately detect temporal information when it is not explicit

References

- O. Alonso, M. Gertz and R. Baeza Yates. Search Results using Timeline Visualizations. Demonstration in SIGIR, 2007.
- R. Baeza-Yates. Searching the Future. In ACM SIGIR Workshop on MF/IR, 2005.
- R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras and F. Silvestri. Challenges on Distributed Web Retrieval. In ICDE, 2007a.
- R. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras and F. Silvestri. The Impact of Caching on Web Search Engines. In Proceedings of SIGIR, 2007b.
- A. Broder: A taxonomy of web search. SIGIR Forum 36(2): 3-10, 2002.
- S. Chien and N. Immorlica. Semantic Similarity Between Search Engine Queries Using Temporal Correlation. In Proceedings of WWW, 2004.
- F. Diaz and R. Jones. Using Temporal Profiles of Queries for Precision Prediction. In Proceedings of SIGIR, 2004.
- J. Kleinberg. Burst and Hierarchical Structure in Streams. In Proceedings of KDD, 2002.
- X. Li and B. Croft. Time-based Language Models. In Proceedings of CIKM, 2003.
- B. Liu, R. Jones and K. Klinkner. Measuring the Meaning in Time Series Clustering of Text Search Queries. In CIKM, 2006.
- I. Ounis, M. de Rijke, C. Macdonald, G. Mishne and I. Soboroff. Overview of the TREC-2006 Blog Track, In Proceedings of TREC, 2006.
- S. Robertson and J. Callan. Routing and Filtering. In TREC: Experiment and Evaluation in Information Retrieval, (eds. E. Voorhees and D. Harman), 2005.
- H. Rode and D. Hiemstra. Using Query Profiles for Clarification. In Proceedings of ECIR, 2006.
- Identifying Similarities, Periodicities, and Bursts for Online Search Queries. In Proceedings of SIGMOD, 2004.
- Q. Zao, S.C.H. Hoi, T.-Y. Liu, S. Bhowmick, M.R. Lyu and W.-Y. Ma. Time-Dependent Semantic Similarity Measure of Queries Using Historical Click-Through Data. In Proceedings of WWW, 2006.