Festschrift for Richard M. Shiffrin

DATA-DRIVEN APPROACHES TO IMPROVING INFORMATION ACCESS

Susan Dumais, Microsoft Research

Overview

- From IU to Industry (Bell Labs 1979, MSR 1997)
- Themes
  - Practical focus of understanding/improving how people retrieve information from external sources, notably computers
  - Simple statistical models, operating over large representative data to solve information access problems
  - Implications for models of human memory
- Examples
  - Latent Semantic Indexing/Analysis
  - Web Search: Personalization; Temporal Dynamics

From IU to Industry

- HCI group at Bell Labs, 1979
- What we did
  - The problem(s)
    - *Human factors in da*
    - *Describing categorie*
    - *Naming commands, information services, etc.*
  - Some solutions
    - Rich aliasing / Adaptive indexing / Latent semantic indexing
  - Closing the loop back to psychology
    - *A solution to Plato's problem* [Psychological Review, 1997]

From Verbal Disagreement to LSI

- Observed: Mismatch between the way that people want to retrieve information from external sources and the way that systems designers or authors describe that information
  - The trouble with UNIX command names
  - This trouble was everywhere – menus, category descriptors, keywords, etc.
- Studied: How people describe objects and operations
  - Text editing operations, systems functionality, common objects, recipes, classified ads, etc.
  - Data:
    - Term x Object matrices
    - Sparse
    - But, no single good name

TABLE I. Word-Object Data

(a) *Sample data from the text-editing study*

Words	Objects: *"Insert"*	*"Delete"*	*"Replace"*	*"Move"*	*"Transpose"*	...
Change	30	22	60	30	41	
Remove	0	21	12	17	5	
Spell	4	14	13	12	10	
Reverse	0	0	0	0	27	...
Leave	10	0	0	1	0	
Make into	0	4	0	0	1	
⋮	⋮	⋮	⋮	⋮	⋮	

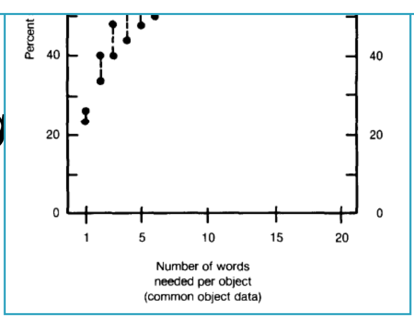
(b) *Sample data from the common object study*

Words	Objects: *"Calculator"*	*"Nectarine"*	*"Lucille Ball"*	*"Pear"*	*"Raisin"*	*"Robin"*	...
Machine	4	0	0	0	0	0	
Green	0	0	0	7	0	0	
Bird	0	0	0	0	0	21	...
Fruit	0	12	0	19	1	0	
Red	0	0	8	0	0	7	
Female	0	0	2	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

From Verbal Disagreement to LSI

- Furnas et al. (1982): *Statistical Semantics: How can a computer use what people name things to guess what things people mean when they name things?*
- Findings:
  - Tremendous (same objects
    - Single keyw
    - Single norm
    - Three aliase
    - Infinite alias

In describing items in a data base, however, system designers are at a disadvantage in that they do not usually get explicit, immediate, and continuous feedback from users. Knowing how people describe common objects and shift their descriptions for audiences of different levels of sophistication may help designers build systems whose information is accessible to the widest possible audience.



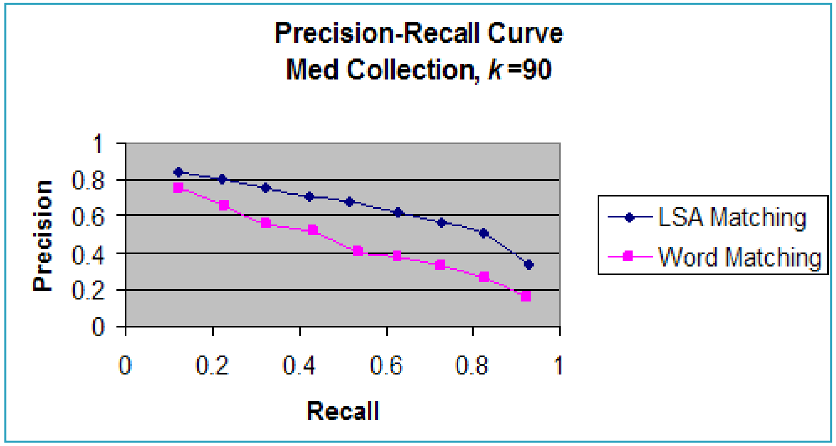
- Interestingly …
  - We referred to this problem as: *verbal disag vocabulary mismatch, statistical semantics*

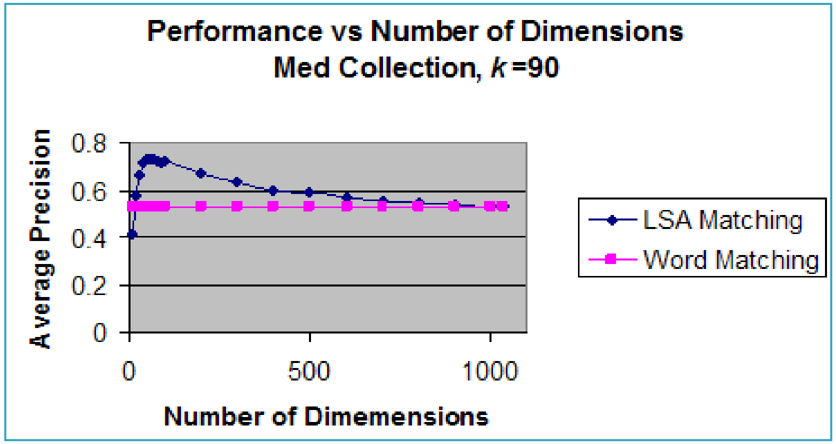
From Verbal Disagreement to LSI

- Some Solutions:
  - Rich aliasing [Gomez et al. 1990]
    - Allow alternative words for the same item
    - From keyword indexing to full-text indexing
  - Adaptive indexing [Furnas 1985]
    - Associate (failed) user queries to destination objects
    - Add these queries as new entries in term-document matrix
    - Quickly reduces failure rate for common requests/tasks
  - Latent Semantic Indexing [Dumais et al. 1988; Deerwester et al. 1990]
    - Model relationships among words, using dimension reduction
    - Especially useful when query and documents are short

LSI and IR

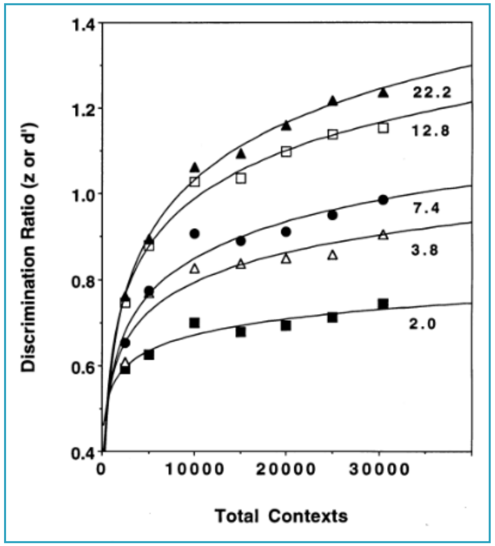
- Learn relations among words indirectly from local co-occurrence data in large collections of text, using dimension reduction (SVD)
- Improves IR
  - Average 30% advantage
  - Widely applicable, incl to cross-language retrieval
- Dimension reduction impt
  - Too many dimensions poor (co-occurrence not enough)
  - Too few dimensions poor





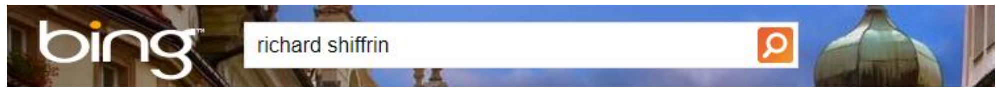
LSA and Human Memory

- Landauer and Dumais (1997): *A solution to Plato's problem: The LSA theory of acquisition, induction and representation of knowledge.*
- Vocabulary tests
  - TOEFL – multiple-choice synonym test
  - Human test takers (64%)
  - Rate of vocabulary acquisition comparable to humans
- Essay scoring
  - $Cor(ETS_1, ETS_2) = 0.87$; $Cor(ETS_i, LSA) = 0.86$
- Semantic priming
- Textual coherence
- Etc.



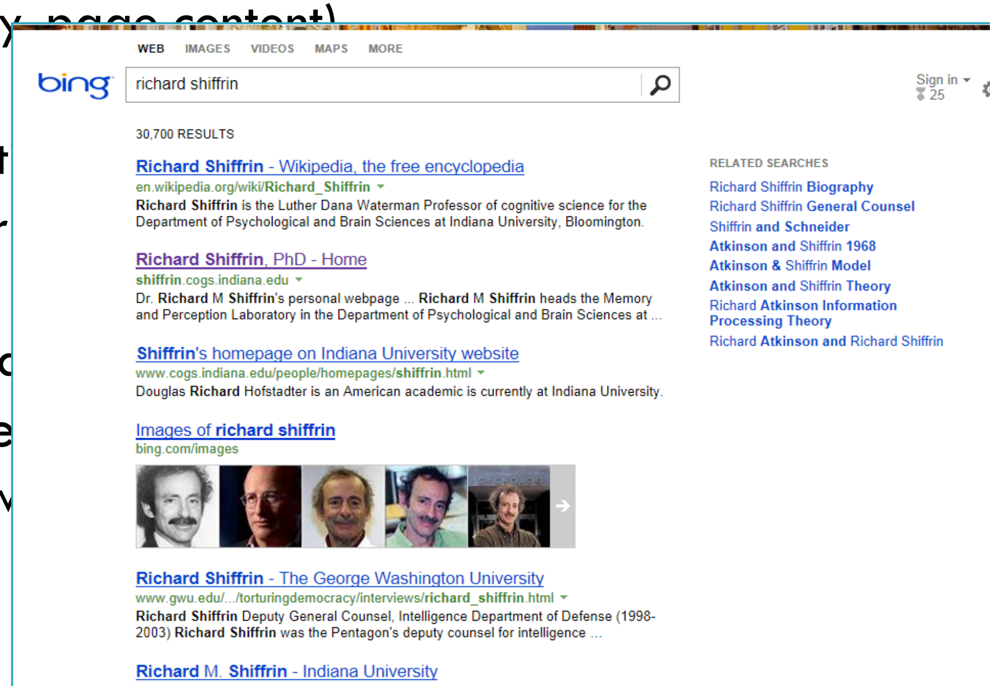
Information Access in the Web Age

- Web Search: How do you go from 2.4 words to anything sensible?



  - Content
    - Match (query, page content)
  - Link structure
    - Used for sett
  - User behavior
    - Anchor text
    - Query-click d
  - Contextual me
    - Who, what, w

g data



- Understanding they are successful?
  - Behavioral logs (and more)

What Are Behavioral Logs?

- Traces of human behavior
  - … seen through the lenses of whatever sensors we have
  - Web search: queries, results, clicks, dwell time, etc.
- Actual, real-world (*in situ*) behavior
  - Not …
    - Recalled behavior
    - Subjective impressions of behavior
    - Controlled experimental tasks
- Large-scale and real-time

Benefits of Behavioral Logs

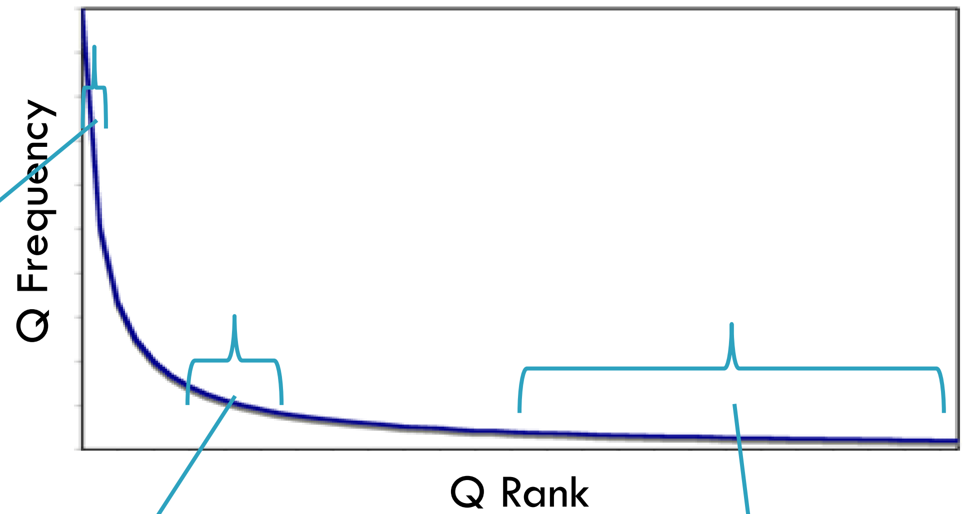
- Real-world
  - Portrait of real behavior, warts and all
- Large-scale
  - Millions of people, tasks, behaviors
  - Diversity of behaviors and information needs (the “long tail”)
  - Subtle differences in behavior
- Real-time
  - Dynamics of information needs
- Practical improvement of Web services
- Broader influence on understanding information needs and impacting policies and society

Surprises In (Early) Query Logs

- Early log analysis …
  - E.g., Jansen et al. 1998, Silverstein et al. 1999, Broder 2002
- Web search != library search
  - Queries are very short, 2.4 words
  - Advanced operators not used or misused
  - Lots of people search for *sex*
  - "Navigational" behavior common, 30-40%
    - Getting to places vs. finding out about things
  - Amazing diversity and dynamics of information needs

Query Frequency Is Not Uniform

- Excite 1999 data
  - ~2.5mil queries <
  - ~1.3mil unique queri
  - ~950k occur exactly
- Zipf Distribution



Top 10 Q

- *sex*
- *yahoo*
- *chat*
- *horoscope*
- *pokemon*
- *hotmail*
- *games*
- *mp3*
- *weather*
- *ebay*

Navigational queries, one-word queries

Query Freq = 10

- bahia AND brazil
- Playstation codes
- breakfast or brunch menus
- cambridge uk telecenter
- www.att.com

Multi-word queries, specific URLs

Query Freq = 1

- *'coren, s'*
- *UNC neuroscience*
- *hormones in memory loss*
- *electronic roladex memory*
- *email address for paul allen the seattle seahawks owner*

Complex queries, rare info needs, misspellings, URLs

One Size Does Not Fit All

- Queries are difficult to interpret in isolation

- Easier if we can model: <u>who</u> is asking, <u>where</u> they are, <u>what</u> they have done in the past, etc.

  Searcher: (*SIGIR* | Susan Dumais … an information retrieval researcher)
  vs. (*SIGIR* | Stuart Bowen Jr. … the Special Inspector General for Iraq Reconstruction)

  Pr

ctio 

(SIGIR | infor

etri

  vs. (… al provi … a

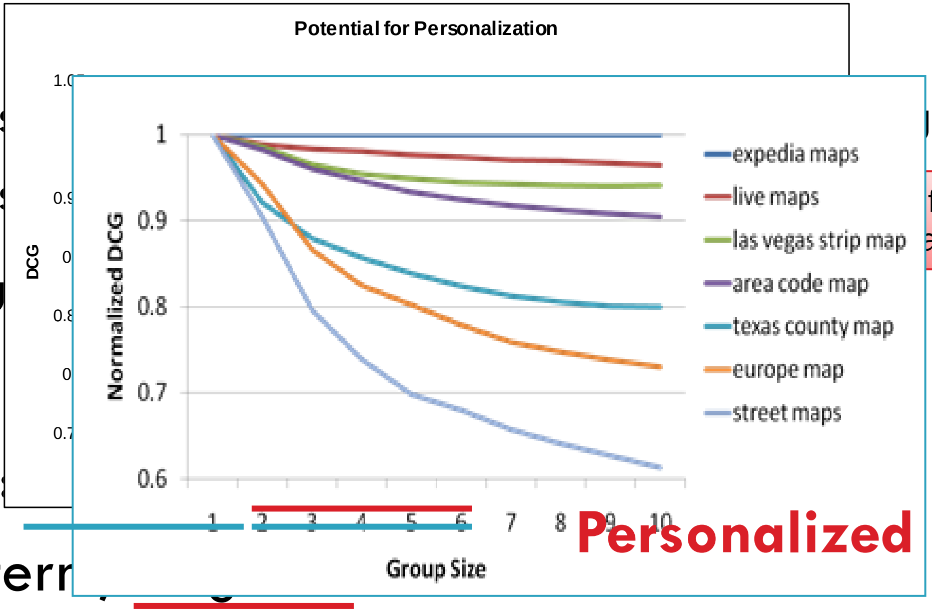
  Locati … G … onference … I … gton DC)

  Time: (*SIGIR* | Aug confernece) vs. (*SIGIR* | Iraq news)

- Using a <u>single ranking for everyone</u>, in every context, at every point in time <u>limits how well a search engine can do</u>

Potential For Personalization

- A single ranking for everyone limits search quality
  - Model the “potential for personalization”
  - Personalized search
    - 46% potential increas
    - 70% potential increas
- Construct and evalu different
  - Sources of evidence
  - Time frames: Short-term
  - Who: Individual, group



for
ation

Personalized Nav

Adaptive Ranking

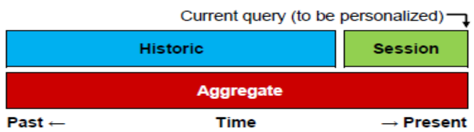
Example 1: Personal navigation

- Re-finding is common in Web search
  - 33% of queries are repeat queries
  - 39% of clicks are repeat clicks
- Many of these are navigational queries
  - E.g., *nytimes*-> www.nytimes.com
- Personal navigational queries
  - Different intents across individuals, but consistently the same intent for an individual
    - E.g., *SIGIR* (for Dumais) -> www.sigir.org
    - E.g., *SIGIR* (for Bowen Jr.) -> www.sigir.mil
  - Very high prediction accuracy (~95%)
  - High coverage (~15% of queries)

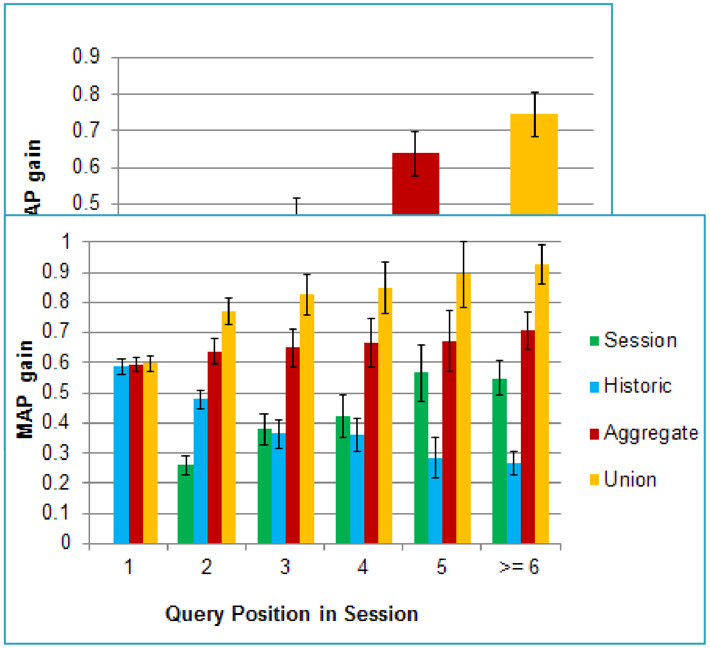
		Repeat Click	New Click
Repeat Query	**33%**	29%	4%
New Query	**67%**	10%	57%
		39%	**61%**

Example 2: Adaptive ranking

- Represent search activities
  - Features
    - Specific queries/URLs
    - Generalizations and specializations
    - Topic (and reading level) distributions
  - Temporal extent
    - Session, Historical, Combinations
    - Temporal weighting



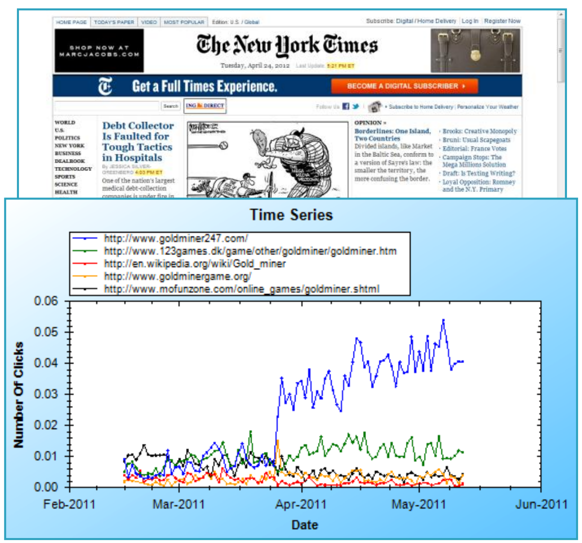
- Learn predictive model
- Re-rank results, given model
  - Session (short-term): +25%
  - Historic (long-term): +45%
  - Combinations: +65-75%
  - 60% of sessions involve multiple queries
    - E.g., (*Rich Shiffrin* | *memory vs. lawyer*)
    - By 3rd query in session, short-term features more important than long-term



Time Changes Everything

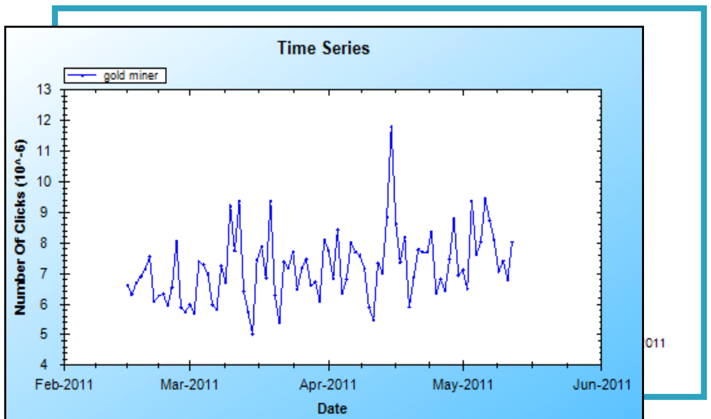
- Content changes over time
  - New documents appear
  - Existing documents change
- User interaction changes over time
  - Queries and query volume non-uniform
  - Clicks, anchor text, "likes", social networks are constantly evolving

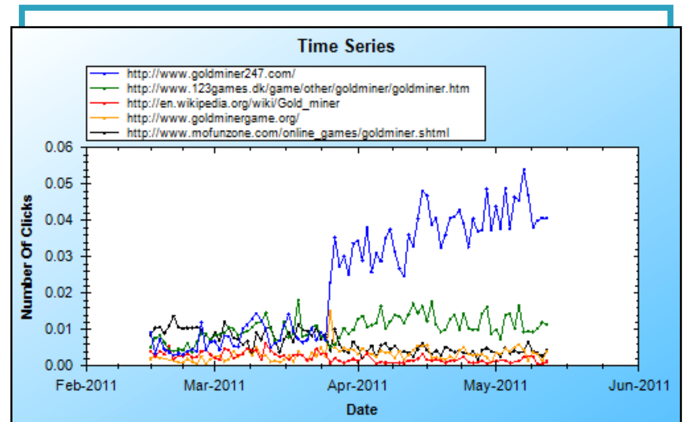
- Diff-IE: Making change more visible
- Temporal retrieval models: Making results more relevant by modeling temporal dynamics of behavior

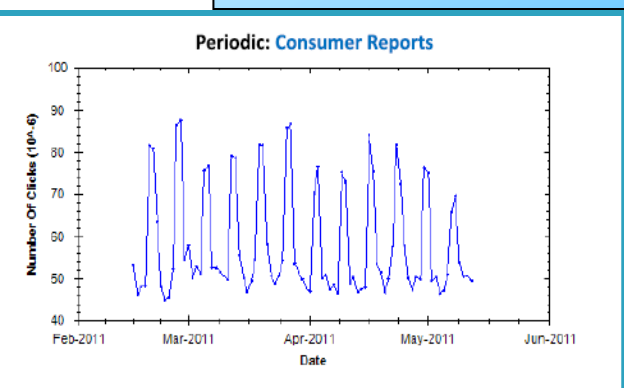


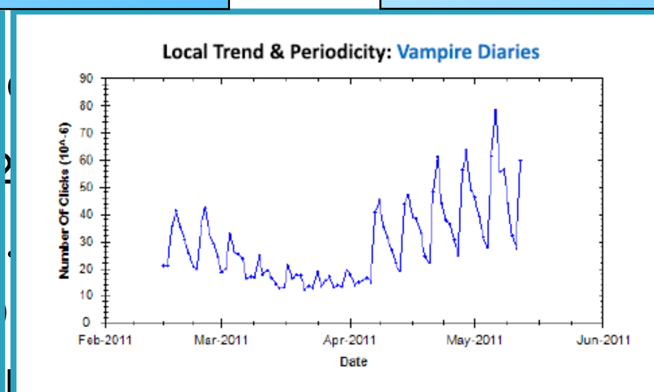
Temporal Retrieval Models

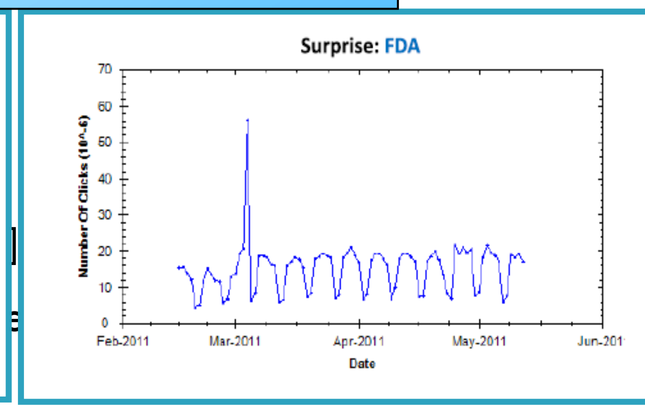
- Queries are not uniformly distributed over time











  - During event: Real-time scores, e.g., espn, cbssports
  - After event: General sites, e.g., wikipedia, usga

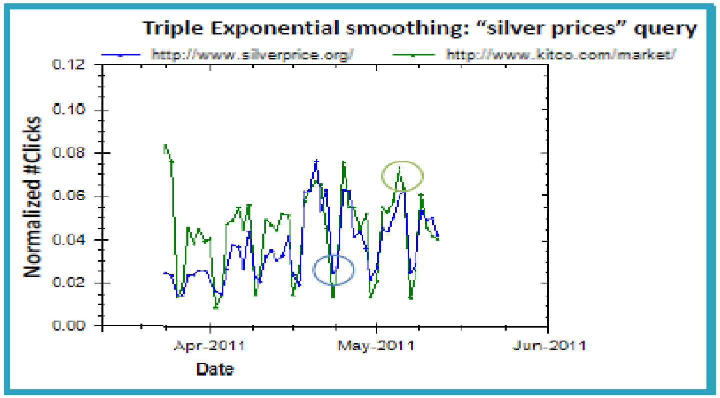
Temporal Retrieval Model

- Model search behavior as time series
  - Assume that the series of observations $Y_1 \ldots Y_n$ is generated sequentially based on some underlying structure
- Linear State Space Model
  - X_t is state vector at time t; a state space model is defined by:

$$Y_t = W(\theta)\, X_t + \epsilon_t \qquad \text{(observation eqn.)}$$

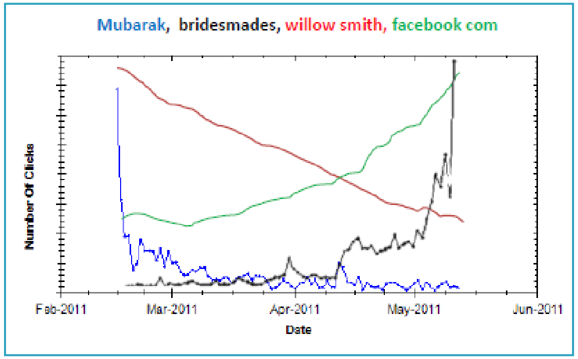
$$X_t = F(\theta)\, X_{t-1} + G(\theta)\, \epsilon_{t-1} \qquad \text{(state transition eqn.)}$$

- Model state with Holt-Winters decomposition
  - Smoothing
  - Trend
  - Periodic/Seasonal



Temporal Retrieval Model

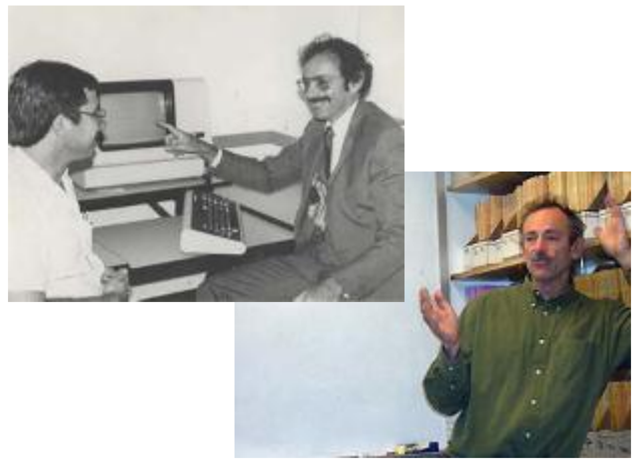
- Learn: Time series model of user behavior
  - Model can be query or URL dependent
- Predict: Future query and click behavior
- Results:
  - For predicting behavior
  - As features for improved ranking of results
    - 110% improvements across all queries
    - Best performance for smoothing + trend
    - Important to detect surprises quickly



Summary

- Practical challenges
  - Understanding how people retrieve information from computer systems
  - Improving access using simple statistical models, operating over large representative data
- Data-driven approach
  - Leads to improvements in information systems
    - LSI, Personalization, Temporal Retrieval models
  - Provides a unique perspective for understanding of the diversity and dynamics of information needs
  - Beyond the search lens

- Thanks Rich!



- Additional details:
  - http://research.microsoft.com/~sdumais