

# Temporal Dynamics and Information Retrieval

*Susan Dumais*

Microsoft Research

<http://research.microsoft.com/~sdumais>

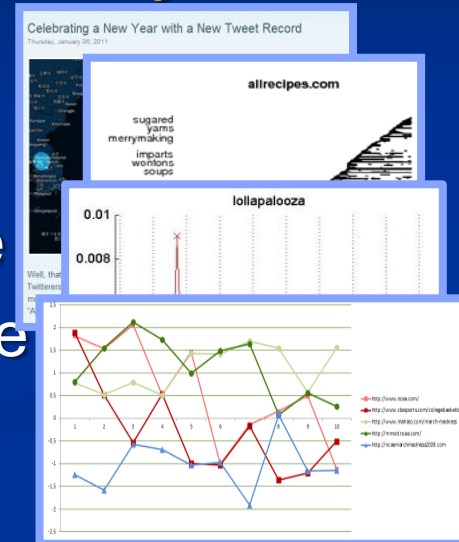
In collaboration with:

Eric Horvitz, Jaime Teevan, Eytan Adar, Jon Elsas, Dan Liebling,  
Richard Hughes, Krysta Svore, Kira Radinsky

# Change is Everywhere in IR

## ■ Change is everywhere in digital information systems

- New documents appear all the time
- Document content changes over time
- Queries and query volume change over time
- What's relevant to a query changes over time
  - E.g., *U.S. Open 2012* (in May vs. Sept)
- User interaction changes over time
  - E.g., anchor text, “likes”, query-click streams, social networks, etc.
- Relations between entities change over time
  - E.g., President of the US is  $\leftrightarrow$  [in 2008 vs. 2004 vs. 2000]

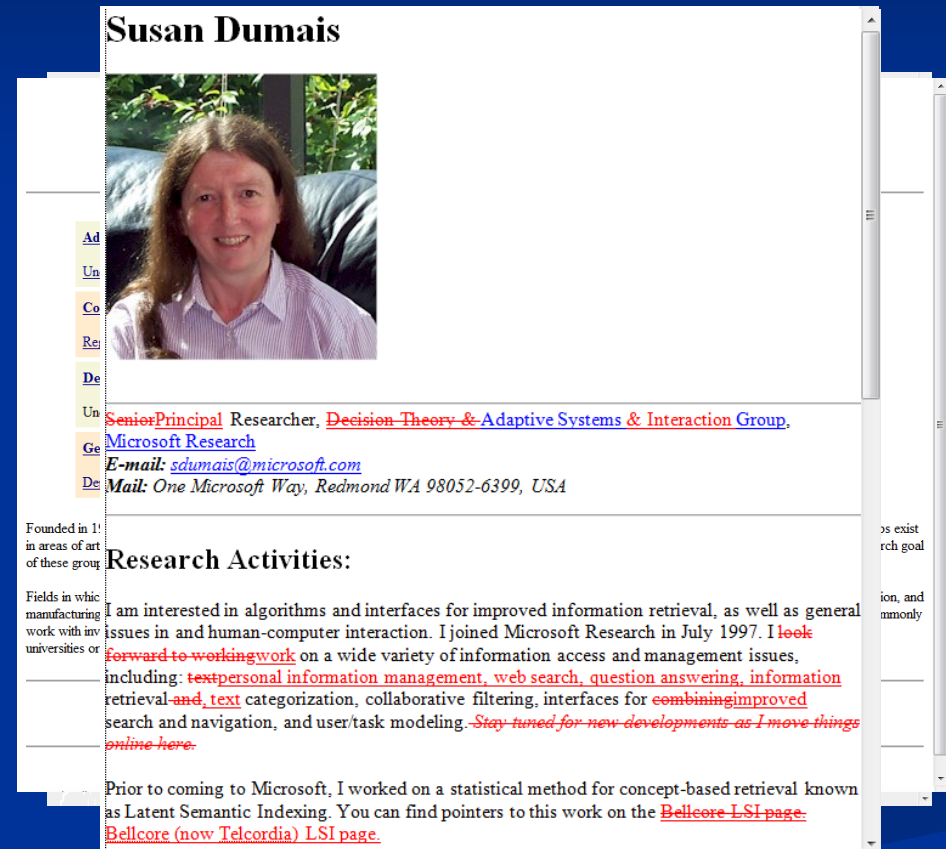


## ■ Change is pervasive in digital information systems ... yet, most retrieval systems ignore it !


# Digital Dynamics Easy to Capture

- Easy to capture

- But ... few tools support dynamics



**Susan Dumais**



**Ad**  
**Un**  
**Co**  
**Re**  
**De**  
**Un**  
**Ge**  
**De**

SeniorPrincipal Researcher, ~~Decision Theory & Adaptive Systems~~ & Interaction Group,  
[Microsoft Research](#)  
**E-mail:** [sdumais@microsoft.com](mailto:sdumais@microsoft.com)  
**Mail:** One Microsoft Way, Redmond WA 98052-6399, USA

Founded in 1' in areas of art of these group

Fields in whic manufacturing work with inv universities or

**Research Activities:**

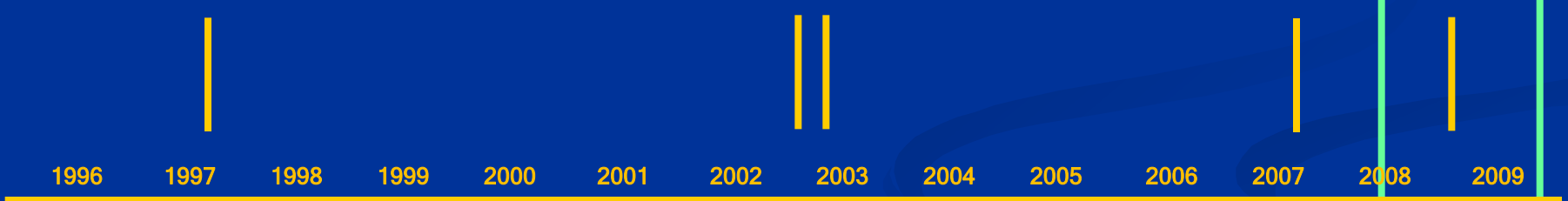
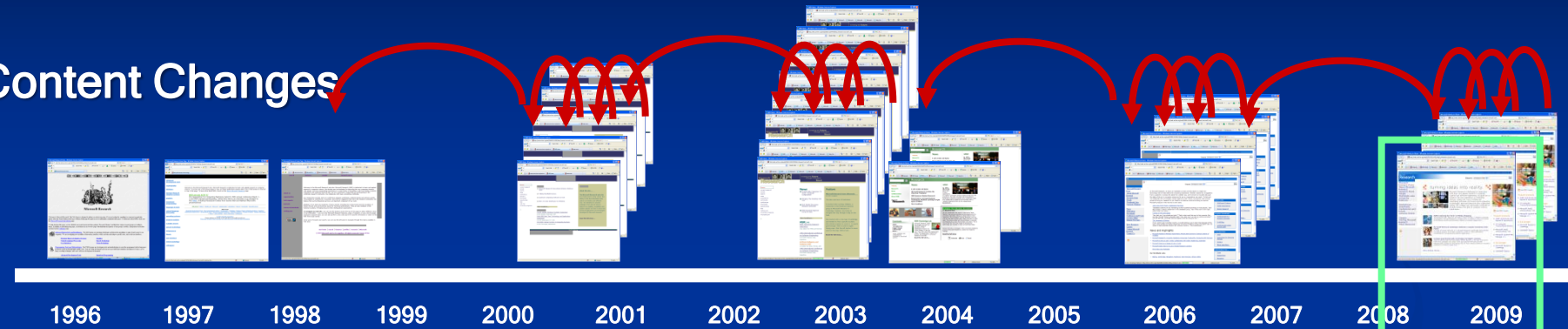
I am interested in algorithms and interfaces for improved information retrieval, as well as general issues in and human-computer interaction. I joined Microsoft Research in July 1997. I ~~look~~ **Forward to working work** on a wide variety of information access and management issues, including: ~~text~~ **personal information management, web search, question answering, information retrieval and, text** categorization, collaborative filtering, interfaces for ~~combining improved~~ search and navigation, and user/task modeling. ~~Stay tuned for new developments as I move things online here.~~

Prior to coming to Microsoft, I worked on a statistical method for concept-based retrieval known as Latent Semantic Indexing. You can find pointers to this work on the ~~Belleore LSI page.~~ **Bellcore (now Telcordia) LSI page.**

ys exist rch goal  
ion, and mmonly

# Web Dynamics

Content Changes



User Visitation/ReVisitation

Today's Browse and Search Experiences

But, ignores ...

# Overview

- Change on the Web
  - Content changes over time
  - User interaction varies over time (queries, re-visitation, anchor text, query-click stream, “likes”)
  - Tools for understanding Web change (e.g., Diff-IE)
- Improving Web retrieval using dynamics
  - Query trends over time
  - Retrieval models that leverage dynamics
  - Task evolution over time

# Overview

- **Change on the Web**
  - Content changes over time
  - User interaction varies over time (queries, re-visitation, anchor text, query-click stream, “likes”)
  - Tools for understanding Web change (e.g., Diff-IE)
- **Improving Web retrieval using dynamics**
  - Query trends over time
  - Retrieval models that leverage dynamics
  - Task evolution over time

# Characterizing Web Change

## Content Changes



- Large-scale Web crawls, over time
  - Revisited pages
    - 55,000 pages crawled hourly for 18+ months
    - Unique users, visits/user, time between visits
  - Pages returned by a search engine (for ~100k queries)
    - 6 million pages crawled every two days for 6 months

# Measuring Web Page Change

- Summary metrics
  - Number of changes
  - Amount of change
  - Time between changes
- Change curves
  - Fixed starting point
  - Measure similarity over different time intervals
- Within-page changes



# Measuring Web Page Change

## ■ Summary metrics

### ■ Number of changes

- 33% of Web pages change
- 66% of visited Web pages change
  - 63% of these change every hr.

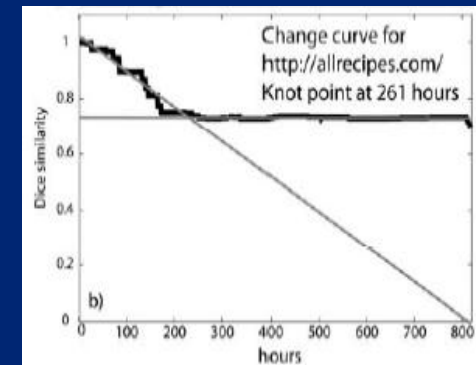
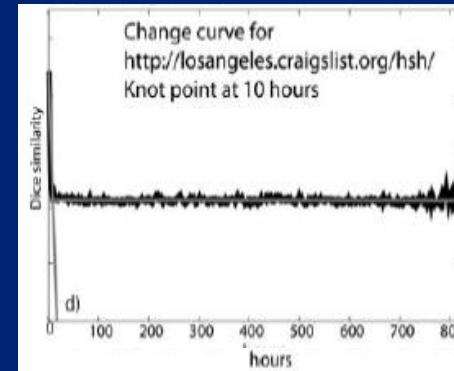
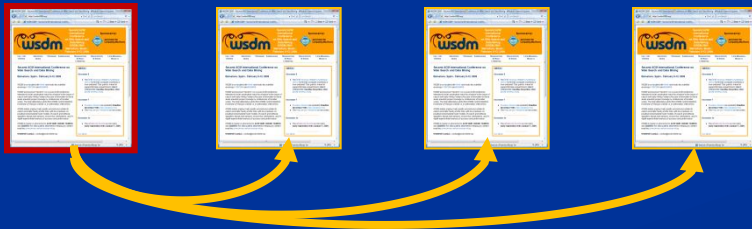
### ■ Amount of change

### ■ Time between changes

- Avg. Dice coeff. = 0.80
- Avg. time bet. change = 123 hrs.
- .edu and .gov pages change infrequently, and not by much
- .com pages change at an intermediate rate, but by a lot
- popular pages change more frequently, but not by much

# Measuring Web Page Change

- Summary metrics
  - Number of changes
  - Amount of change
  - Time between changes
- Change curves
  - Fixed starting point
  - Measure similarity over different time intervals



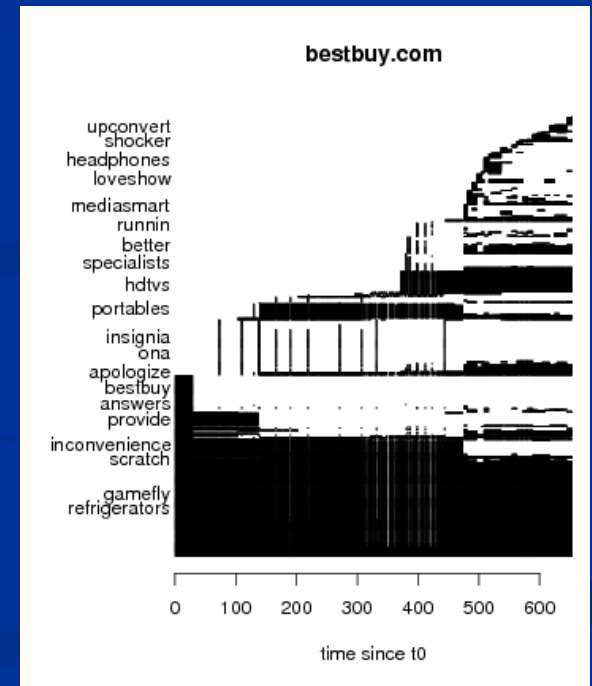
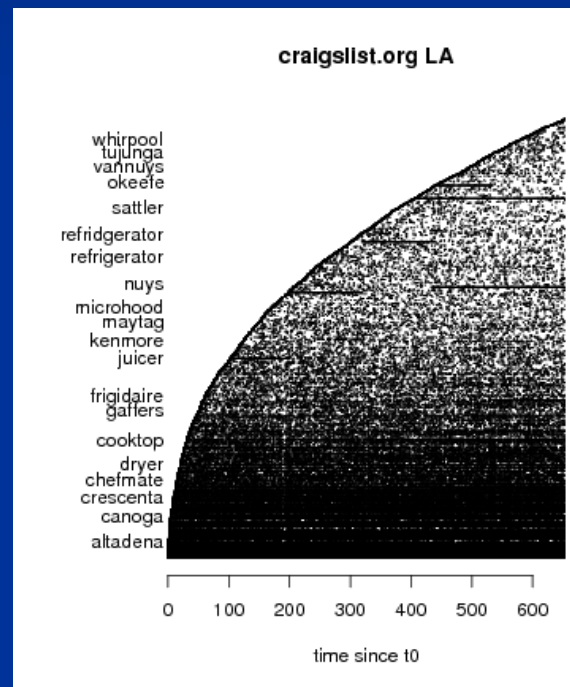
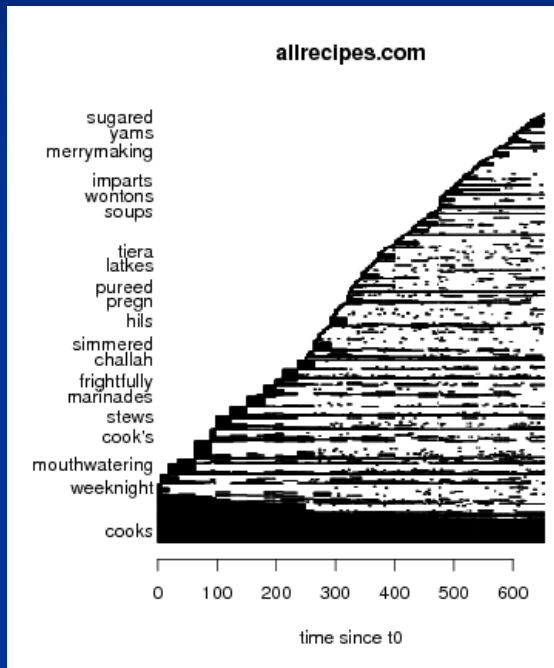
# Measuring Within-Page Change

- Term-level changes
  - Divergence from norm
    - cookbooks
    - salads
    - cheese
    - ingredient
    - bbq
    - ...
  - “Staying power” in page



Sep. Oct. Nov. Dec.  
Time

# Example Term Longevity Graphs



# Revisitation on the Web

- Revisitation patterns
  - Log analyses
    - Toolbar logs for *revisitation*
    - Query logs for *re-finding*
  - User survey to understand intent in revisitations

1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009



## User Visitation/ReVisitation

What was the last Web page you visited?

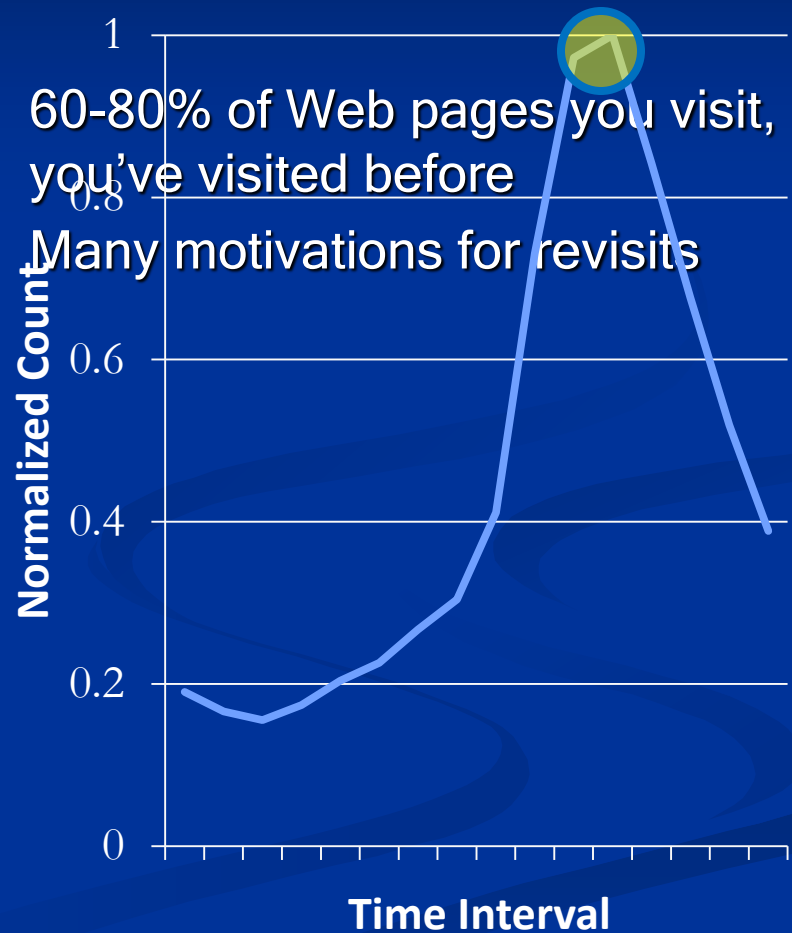
Why did you visit (re-visit) the page?

# Measuring Revisitation

- Summary metrics
  - Unique visitors
  - Visits/user
  - Time between visits
- Revisitation curves
  - Histogram of revisit intervals
  - Normalized

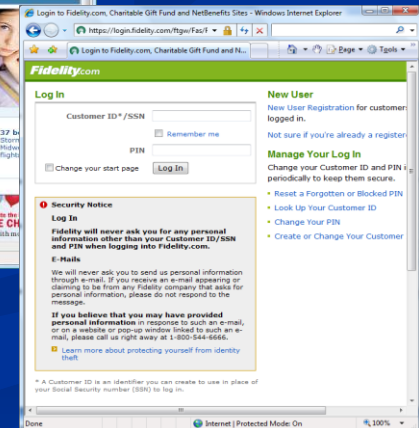


- 60-80% of Web pages you visit, you've visited before
- Many motivations for revisits



# Four Revisitation Patterns

- **Fast**
  - Hub-and-spoke
  - Navigation within site
- **Hybrid**
  - High quality *fast* pages
- **Medium**
  - Popular homepages
  - Mail and Web applications
- **Slow**
  - Entry pages, bank pages
  - Accessed via search engine



# Relationships Between Change and Revisitation

ACM SIGIR 2011  
July 24-28, 2011  
The 34<sup>th</sup> Annual International ACM SIGIR Conference

Home  
Important Dates  
News  
Program  
General Schedule  
Main Conference  
Papers  
Tutorials  
Workshops  
Demos  
Posters  
Doctoral Consortium  
Keynotes  
Industrial Track  
Elsevier App Challenge  
Other Events

The 34<sup>th</sup> Annual ACM SIGIR Conference

**Reminders for Attendees**  
**Reminders for Poster & Demo**

Registration

Important Dates

- 17 Jan 2011 : Abstracts for full research papers due
- 24 Jan 2011 : Full research paper submissions due
- 28 Jan 2011 : Workshop proposals due
- 11 Feb 2011 : Posters, demonstration, and tutorial proposals due
- 5 Mar 2011 : Notification of workshop acceptances
- 7 Mar 2011 : Doctoral consortium proposals due
- 8 Apr 2011 : All other acceptance notifications
- 24-28 Jul 2011 : Conference

News:  
June 20:

SIGIR  
Special Interest Group  
on Information Retrieval

- Interested in change
  - Monitor
- Effect change
  - Transact
- Change unimportant
  - Re-find old
  - Change can interfere with re-finding



# Revisitation and Search (Re-finding)

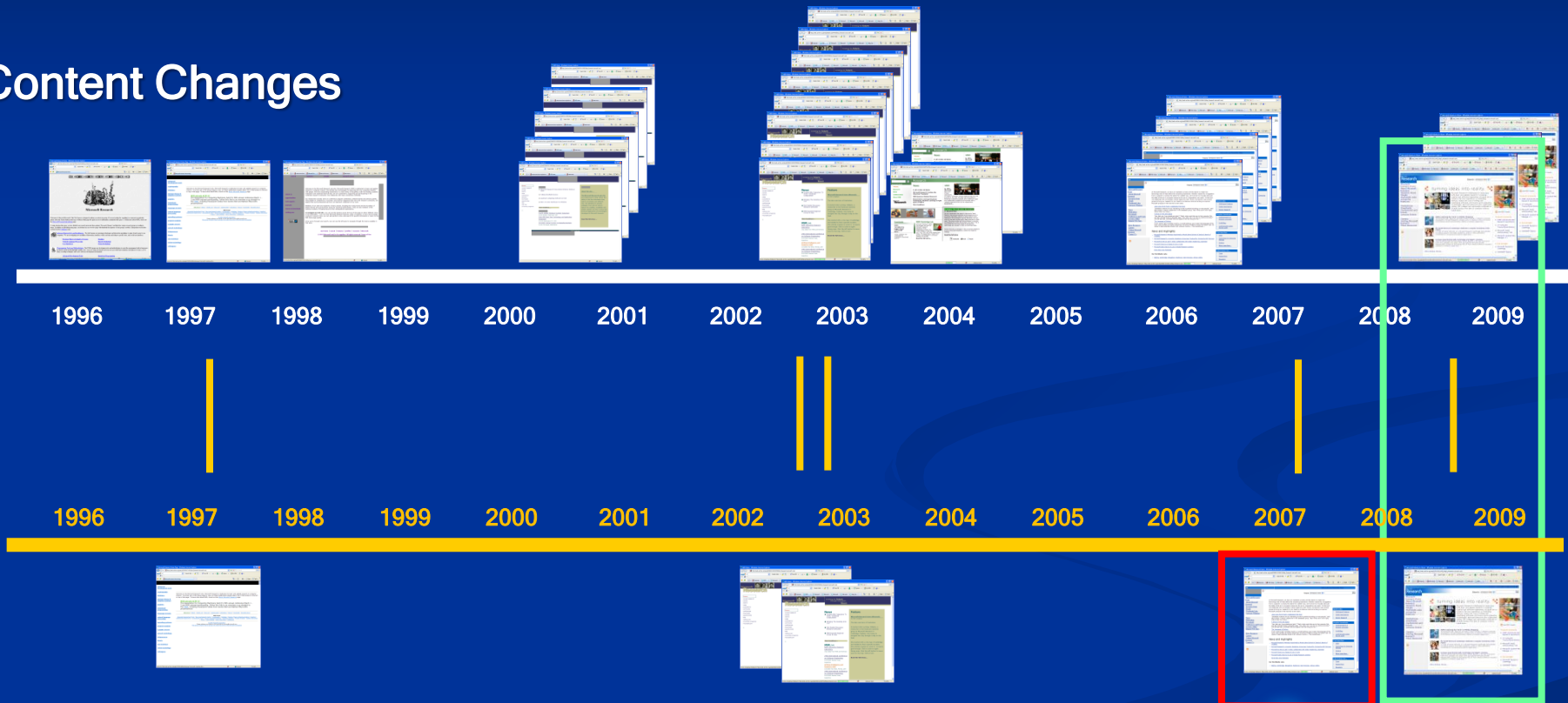
- 60-80% of the Web page visits are re-revisits
- 33%-43% of queries are re-finding

- Repeat query (33%)
  - Q: *microsoft research*
  - Click same or different URLs
- Repeat click (39%)
  - <http://research.microsoft.com/>
  - Q: *microsoft research; msr*
- Big opportunity (43%)
  - 24% “navigational revisits”

		Repeat Click	New Click
Repeat Query	33%	29%	4%
New Query	67%		

# Building Support for Web Dynamics

## Content Changes

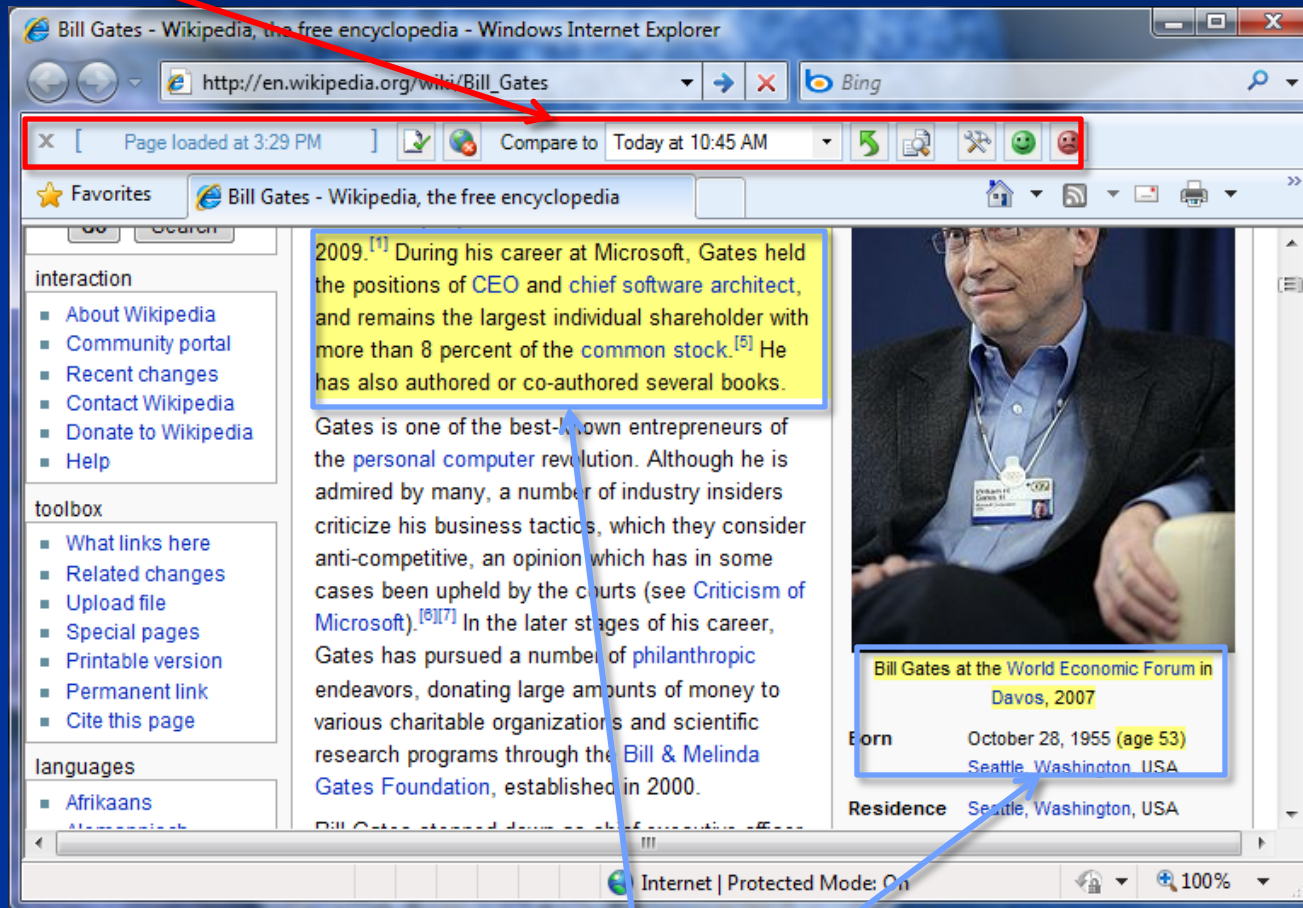


## User Visitation/ReVisitation



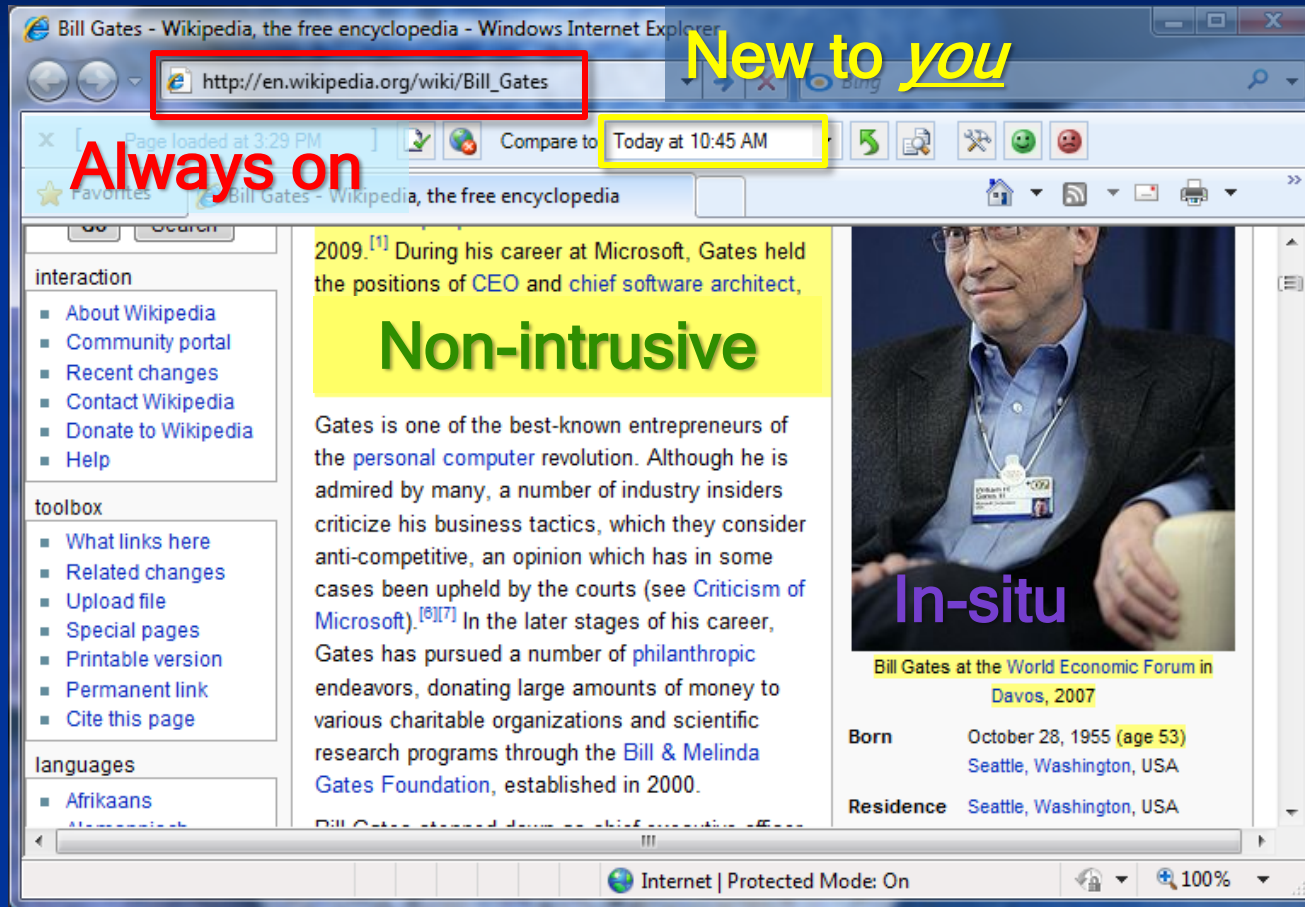
# Diff-IE

## Diff-IE toolbar



Changes to page since  
your last visit

# Interesting Features of Diff-IE



Try it: <http://research.microsoft.com/en-us/projects/diffie/default.aspx>

# Examples of Diff-IE in Action

# Expected New Content

HOME PAGE TODAY'S PAPER VIDEO MOST POPULAR Edition: U.S. / Global Subscribe: Digital / Home Delivery Log In Register Now

**Goldman Sachs 10000 WOMEN** **The New York Times** Thursday, March 8, 2012 Last Update: 1:38 PM ET

Search **ING DIRECT** Follow Us Subscribe to Home Delivery | Personalize Your Weather

**HICKEY FREEMAN**

**WORLD**  
U.S.  
POLITICS  
NEW YORK  
BUSINESS  
DEALBOOK  
TECHNOLOGY  
SPORTS  
SCIENCE  
HEALTH  
ARTS  
STYLE  
OPINION

Autos  
Blogs  
Books  
Cartoons  
Classifieds  
Crosswords  
Dining & Wine  
Education  
Event Guide  
Fashion & Style  
Home & Garden  
Jobs  
Movies  
Music  
Obituaries  
Real Estate  
Sunday Magazine  
T Magazine  
Television  
Theater  
Travel  
Weddings / Celebrations

**Syrian Minister Appears to Defect and Join Opposition**  
By RICK GLADSTONE and J. DAVID GOODMAN  
16 minutes ago  
A man identifying himself as the deputy oil minister, Abdo Hussameldin, appeared in a video posted to YouTube on Wednesday, and said he had left his post.  
[Post a Comment](#) | [Read \(23\)](#)

**Intractable Afghan Graft Hampering U.S. Strategy**  
By MATTHEW ROSENBERG and GRAHAM BOWLEY  
Despite years of urging and oversight by American advisers, President Hamid Karzai's government has yet to prosecute a high-level corruption case.  
[Post a Comment](#) | [Read \(321\)](#)

**A Quest for Truth About the Last Days of Bin Laden**  
By DECLAN WALSH  
An inquiry that began as a personal attempt to truth-check accounts of Osama bin Laden's final years in Pakistan is another addition to a long-gestating game.

**TimesCast** **HICKEY FREEMAN**  
  
Moises Saman for The New York Times  
**TimesCast: Syrians Seek Safety in Turkey**  
Syrians face difficult choices at the Turkish border | Daw Aung San Suu Kyi shifts from dissident to stump speaker on Myanmar's campaign trail.  
· [TimesCast Archive](#)

**OPINION** »  
ROOM FOR DEBATE  
**Hate and Hate Crimes**  
Do bias laws protect against or fuel intimidation and bigotry?  
· Kristof: Austerity's Pain  
· Collins: Dogging Romney  
· Blow: Mitt's Weakness  
· Editorial: Housing News  
· Townies: When the Drag Queens Go  
· The Loyal Opposition: The Two Mitt Romneys

**MARKETS** » At 1:40 PM ET  
S.&P. 500 1,364.91 +12.28 +0.91%  
Dow 12,907.22 +69.89 +0.54%  
Nasdaq 2,967.92 +32.23 +1.10%

GET QUOTES My Portfolios »  
Stock, ETFs, Funds

**CAMPAIGN 2012**  
**Democrats Warm to Obama as a Campaign Ally**  
By JENNIFER STEINHAUER 12:35 PM ET  
On Capitol Hill, Democrats have begun to mention Mr. Obama more often and to publicly back his proposals.  
· Obama Mines for Voters With High-Tech Tools  
[Post a Comment](#) | [Read \(2\)](#)

**Romney Fund-Raising Lags as Big Givers Hit Limits**  
By NICHOLAS CONFESSORE and ASHLEY PARKER  
With fewer small donors giving to Mitt Romney, there is a perception that he is favored more by his party's elite.  
· The Gingriches Take to the Dance Floor 12:47 PM ET  
[Post a Comment](#) | [Read \(52\)](#)

**'Titanic' Director Takes On His Most Risky Project**  
By WILLIAM J. BROAD 2:01 AM ET

**THEATER** »

**The New York Times**  
SPECIAL OFFER  
4 WEEKS FOR 99¢  
[CLICK HERE](#)

**SPECIAL OFFER**  
4 WEEKS FOR 99¢  
[GET IT NOW](#)  
  
The New York Times  
NYTIMES.COM

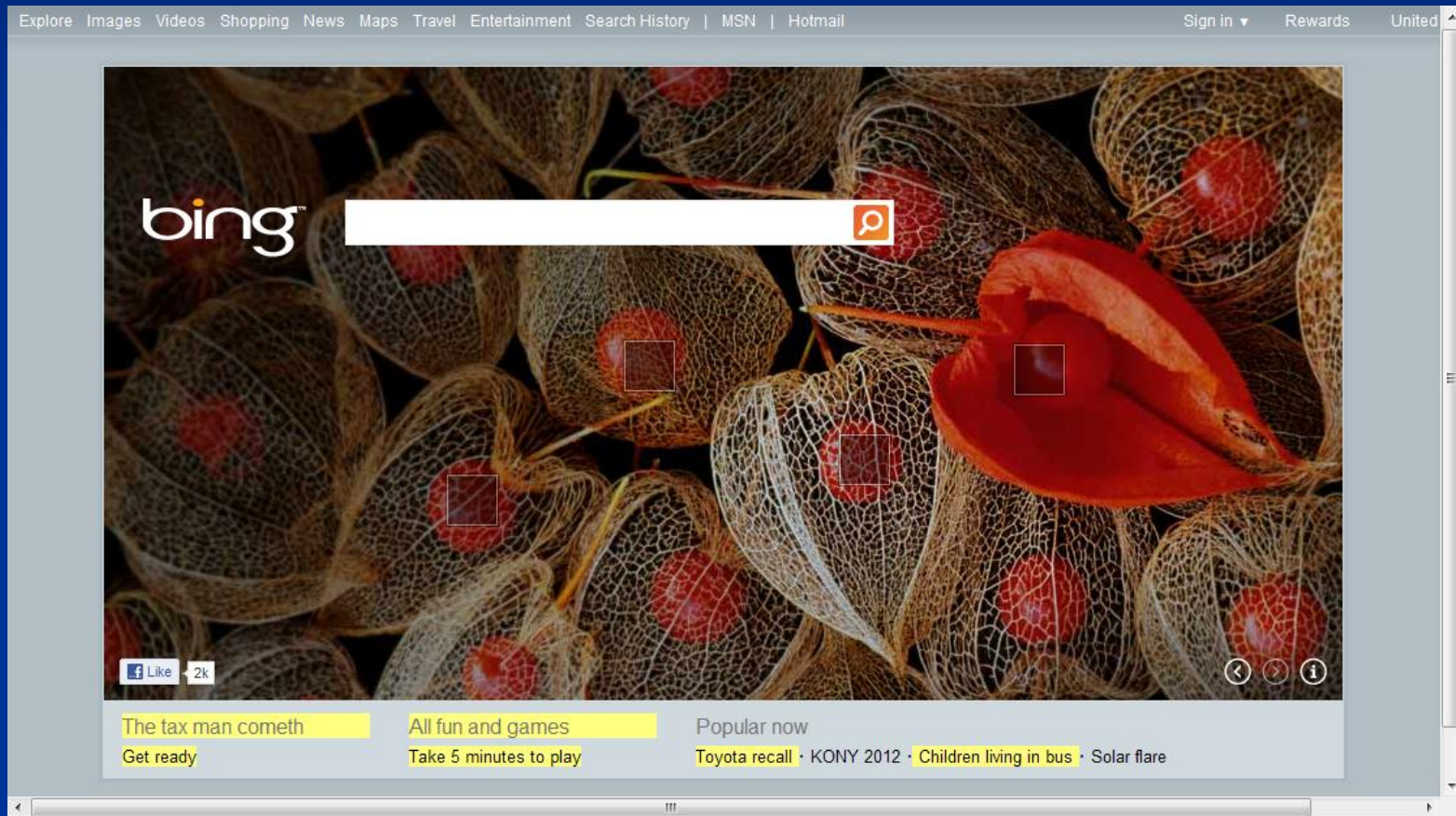
# Monitor

The screenshot shows a Twitter profile for Susan Dumais. The top navigation bar includes 'Home', 'Profile', 'Find People', 'Settings', 'Help', and 'Sign out'. The profile header shows 'Name Susan Dumais', '20 following', '12 followers', and '1 listed'. The main content area is titled 'You follow 20 people' and lists several followed accounts:

- pfromthenc** (Loren Terveen): Just had my first Surly Furious... waited until \*after\* I finished working on the proposal tonight. [about 22 hours ago](#)
- DARPA\_News** (DARPA | Arlington, VA): Team standings from DARPA Network Challenge posted on <http://bit.ly/5kdAZ1> [about 4 hours ago](#)
- huffingtonpost** (HuffingtonPost.com): [GOP Senator: We Will Unanimously Oppose Newest Health Care Compromise](#) <http://bit.ly/5xRBEI> [about 1 hour ago](#)
- nytimes** (The New York Times | New York, NY): [Citi Races to Pay Back Bailout Aid](#) <http://bit.ly/87E9Ry> [5 minutes ago](#)
- dmrussell** (CA, USA): @roblyons Google Office is at 1101 New York Ave (although the entrance is on I) #gtadc [about 13 hours ago](#)

On the right side, there are sections for 'Tweets' (0), 'Favorites', and 'Following' (a grid of profile pictures).

# Serendipitous Encounters





# Unexpected Important Content



The screenshot shows a Windows Internet Explorer browser window displaying the SIGIR 2009 website. The address bar shows the URL <http://www.sigir2009.org/>. The page title is "Welcome -- Please join us in Boston | SIGIR'09". The main content area features a banner for "The 32nd Annual ACM SIGIR Conference July 19-23 2009" and a large heading "Welcome -- Please join us in Boston". Below this, a paragraph of text describes the conference location and registration status. A navigation menu on the left lists various sections: Program, For attendees, For contributors, and About us. The "Recent news and upcoming deadlines" section contains several bullet points, with two highlighted in yellow. The browser's status bar at the bottom shows "Done" and "Internet" with a 100% zoom level.

Welcome -- Please join us in Boston | SIGIR'09 - Windows Internet Explorer

<http://www.sigir2009.org/>

File Edit View Favorites Tools Help

msn

Page loaded at 4:04 PM

Compare to Today at 9:04 AM

Welcome -- Please join us in Boston | SIGIR'09

## SIGIR 2009 Boston

The 32nd Annual ACM SIGIR Conference July 19-23 2009

### Welcome -- Please join us in Boston

The SIGIR 2009 conference opens in a few days in Boston, Massachusetts, at the Sheraton Boston Hotel and Northeastern University. The conference is chock full of exciting events and registrations are strong and still growing. We are looking forward to an exciting week.

[Conference registration site \(updates or late\)](#)

[NEU dorm registration \(payments only\)](#)

#### Program

- Schedule
- Tutorials
- Workshops
- Industry track
- Keynotes
- Papers
- Posters
- Demonstrations

#### For attendees

- Visas
- Boston
- Venue
- Registration
  - student support
  - accommodations
- Red Sox

#### For contributors

- Important dates
- Mentoring (closed)
- Call for... (closed)
- Submitting (closed)

#### About us

- History
- Organizers

#### Recent news and upcoming deadlines

- Please join your colleagues by starting the conference with a free continental breakfast in the Sheraton Hotel, Back Bay A&B, from 7:00am to 8:20am on Monday, July 20.
- The conference banquet is currently full. Effective July 15th, reservations will be wait-listed for the banquet.
- Standard conference registration closes the night of Sunday, July 12th (Boston time). See below for more information.
- Registration for Northeastern University housing has closed. If you need to change your reservation, contact [questions@sigir2009.org](mailto:questions@sigir2009.org)
- Three special tourism events have been added to the schedule: a panoramic view of the city on Sunday, a famous duck boat tour on Saturday, and a sunset harbor cruise on Wednesday. Look under "Boston" on the left for more information. Two of the events require sufficient interest to occur, so please fill out the questionnaire on that page.
- Industry Track registration is now available for students.

#### Arrival at the conference

- Tutorial and doctoral consortium registration is on the Northeastern campus, Sunday morning (the 19th). If you have registered for a tutorial or want to register for a tutorial, please get to Shillman Hall between 8:00 and 9:00 (the doctoral consortium starts at 8:30; morning tutorials start at 9:00). Volunteers and signs will provide direction from the hotel and the NEU dorms. [Here is a map.](#)
- Normal conference registration takes place at the Sheraton hotel, starting at 3:00pm on Sunday, July 19. Registration will be available throughout the conference if you are not arriving on Sunday.
- Workshop registrations may be made during normal registration periods or on the Northeastern campus on Thursday morning (the 23rd).
- If you have paid your fees in full in advance, your registration process should be very speedy.
- A free breakfast is available Monday morning (only) in the Sheraton.

For those who are driving, parking is available at the Sheraton and at Northeastern University. See [here](#)

Done Internet 100%

# Understand Page Dynamics

The image shows a screenshot of a Bing search results page for the query "jaime teevan". The page layout includes a search bar at the top with the query "jaime teevan" and a magnifying glass icon. Below the search bar are tabs for "Web" and "Images". The main content area displays search results under the heading "ALL RESULTS" and "1-10 of 11,700 results". The results are listed in a vertical column, each with a title, a brief description, and a URL. The results include:

- Jaime Teevan, Ph.D.**: Researcher studying information retrieval and human computer interaction at Microsoft Research. URL: [research.microsoft.com/en-us/um/people/teevan](http://research.microsoft.com/en-us/um/people/teevan)
- Jaime Teevan: Work**: Doctoral candidate at Massachusetts Institute of Technology. URL: [people.csail.mit.edu/teevan/work](http://people.csail.mit.edu/teevan/work)
- DBLP: Jaime Teevan**: 2010; 36 : Jaime Teevan, Susan T. Dumais, Daniel J. Liebling: A longitudinal study of how highlighting web content change affects people's web interactions. URL: [www.informatik.uni-trier.de/~ley/db/indices/a-tree/t/Teevan:Jaime.html](http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/t/Teevan:Jaime.html)
- Jaime Teevan - Pipl Profile**: Pipl profile of Jaime Teevan. Quick facts, personal profiles, publications, contact details and much more. URL: [pipl.com/directory/people/Jaime/Teevan](http://pipl.com/directory/people/Jaime/Teevan)
- Jaime Teevan: Work**: Doctoral candidate at Massachusetts Institute of Technology. Publications. URL: [people.csail.mit.edu/teevan/work/publications/subject.html](http://people.csail.mit.edu/teevan/work/publications/subject.html)
- Jaime Teevan - LinkedIn**: Research · 232 connections · Greater Seattle Area. View Jaime Teevan's professional profile on LinkedIn. URL: [www.linkedin.com/pub/jaime-teevan/0/542/7ab](http://www.linkedin.com/pub/jaime-teevan/0/542/7ab)
- TR35: Jaime Teevan, 32 - Technology Review**: From MIT. Information on Emerging Technologies & impact on business & society.

On the right side of the page, there is a "Sponsored sites" section with a yellow background, featuring a link to "We Found Jaime Teevan" from Intelius.com. The page also includes a "RELATED SEARCHES" section on the left with "Susan Dumais" listed. The bottom of the page shows a navigation bar with a magnifying glass icon and a "111" indicator.

Expected



Expected New Content



Monitor



Unexpected Important Content

33 - Estimate Mothers & More Expo... Use Chapter Names and Events for more information. Yeah, baby!

Subject	Replies	Views	Last post
Small Business + ...	21	206	Monday 16, 2008, 9:00 AM
Confusing Invoicing	10	83	Monday 16, 2008, 9:00 AM
Back to work Part II	8	43	Monday 16, 2008, 9:00 AM
Aldergate Christian Preschool - HIKING	0	17	Monday 16, 2008, 9:00 AM
do you know anyone in Italy or Sweden?	3	20	Monday 16, 2008, 9:00 AM
Call an Interview and I'm Recalling ...	94	562	Monday 16, 2008, 9:00 AM
Autism and Industry? Finding the Right Fit	7	105	Monday 16, 2008, 9:00 AM
Any members work at Amazon.com?	10	118	Monday 16, 2008, 9:00 AM
Time Magazine on SAHRA - getting back into work!	1	75	Monday 16, 2008, 9:00 AM
Freemason Resurgence	12	41	Monday 16, 2008, 9:00 AM
Congrat designer or Admin?	0	45	Monday 16, 2008, 9:00 AM

Attend to Activity



Serendipitous Encounter

Unexpected



Edit



Understand Page Dynamics



Unexpected Unimportant Content

# Studying Diff-IE

- Internal study of Diff-IE
- Logging
  - URLs visited
  - Amount of change when revisited
- Feedback buttons
- Survey
  - Prior to installation
  - After a month of use
- Experience interview



6. How often do you find the following types of pages change?

Required	Always	Often	Sometimes	Rarely	Never
News pages	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Message boards, forums, newsgroups	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Company homepages	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Personal homepages of people you know	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Pages with product information	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reference pages (dictionaries, yellow pages, maps)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Wikipedia pages	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Blogs you read	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Search engine results	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>






*In situ*

*Representative*




*Experience*

*Longitudinal*

# People Revisit More

- Perception of revisitation remains constant
  - How often do you revisit? 
  - How often are revisits to view new content? 
- Actual revisitation increases
  - First week: 39.4% of visits are revisits
  - Last week: 45.0% of visits are revisits 
- Why are people revisiting more with DIFF-IE?

# Revisited Pages Change More

- Perception of change increases
  - What proportion of pages change regularly?  8%
  - How often do you notice unexpected change?  17%
- Amount of change seen increases
  - First week: 21.5% revisits changed, by 6.2%  51+%
  - Last week: 32.4% revisits changed, by 9.5%
- Diff-IE is driving visits to changed pages
  - It supports people in understanding change

# Other Examples of Dynamics and User Experience

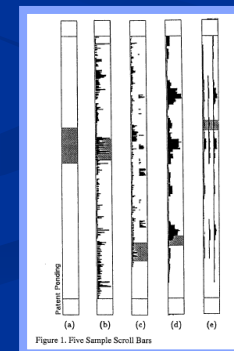
## ■ Content changes

- Diff-IE (Teevan et al., 2008)
- Zoetrope (Adar et al., 2008)
- Diffamation (Chevalier et al., 2010)
- Temporal summaries and snippets ...



## ■ Interaction changes

- Explicit annotations, ratings, "likes", etc.
- Implicit interest via interaction patterns
  - Edit wear and read wear (Hill et al., 1992)



# Overview

## ■ Change on the Web

- Content changes over time
- User interaction varies over time (queries, re-visitation, anchor text, query-click stream, “likes”)
- Tools for understanding Web change (e.g., Diff-IE)

## ■ Improving Web retrieval using dynamics

- Query trends over time
- Retrieval models that leverage dynamics
- Task evolution over time

## Questions?



# Overview

- **Change on the Web**
  - Content changes over time
  - User interaction varies over time (queries, re-visitation, anchor text, query-click stream, “likes”)
  - Tools for understanding Web change (e.g., Diff-IE)
- **Improving Web retrieval using dynamics**
  - Query trends over time
  - **Retrieval models that leverage dynamics**
  - Task evolution over time

# Temporal Retrieval Models 1

## (content-based)

- Current retrieval algorithms look only at a single snapshot of a page
- But, Web page content changes over time
- Can we can leverage this to improved retrieval?
  - Pages have different *rates of change*
    - Different priors (using change rate vs. link structure)
  - Terms have *different longevity (staying power)*
    - Some are always on the page; some transient
  - Language modeling approach to ranking

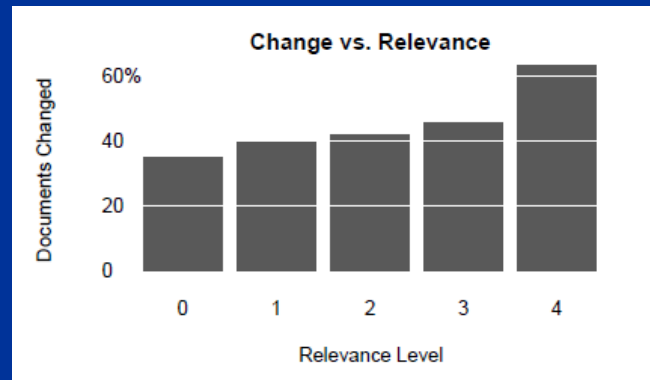
$$P(D|Q) = P(D) \cdot P(Q|D)$$

Change prior

Term longevity

# Relevance and Page Change

- Page change is related to relevance
  - Human relevance judgments
    - 5-point scale - Perfect/Excellent/Good/Fair/Bad
  - Rate of Change -- 60% Perfect pages; 30% Bad pages



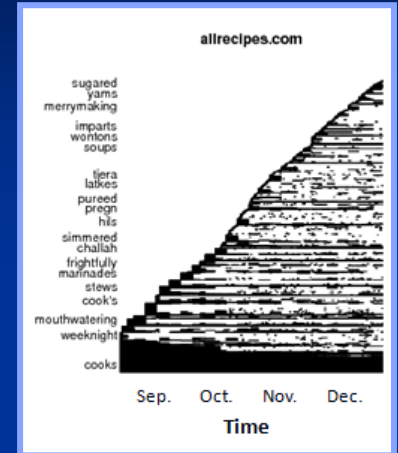
- Use change rate as a document prior (vs. priors based on link structure like Page Rank)
  - Shingle prints to measure change

$$P(D | Q) = P(D) \cdot P(Q | D)$$

Change prior

# Relevance and Term Change

- Terms patterns vary over time
- Represent a document as a mixture of terms with different “staying power”
  - Long, Medium, Short



$$P(Q | D) = \lambda_L P(Q | D_L) + \lambda_M P(Q | D_M) + \lambda_S P(Q | D_S)$$

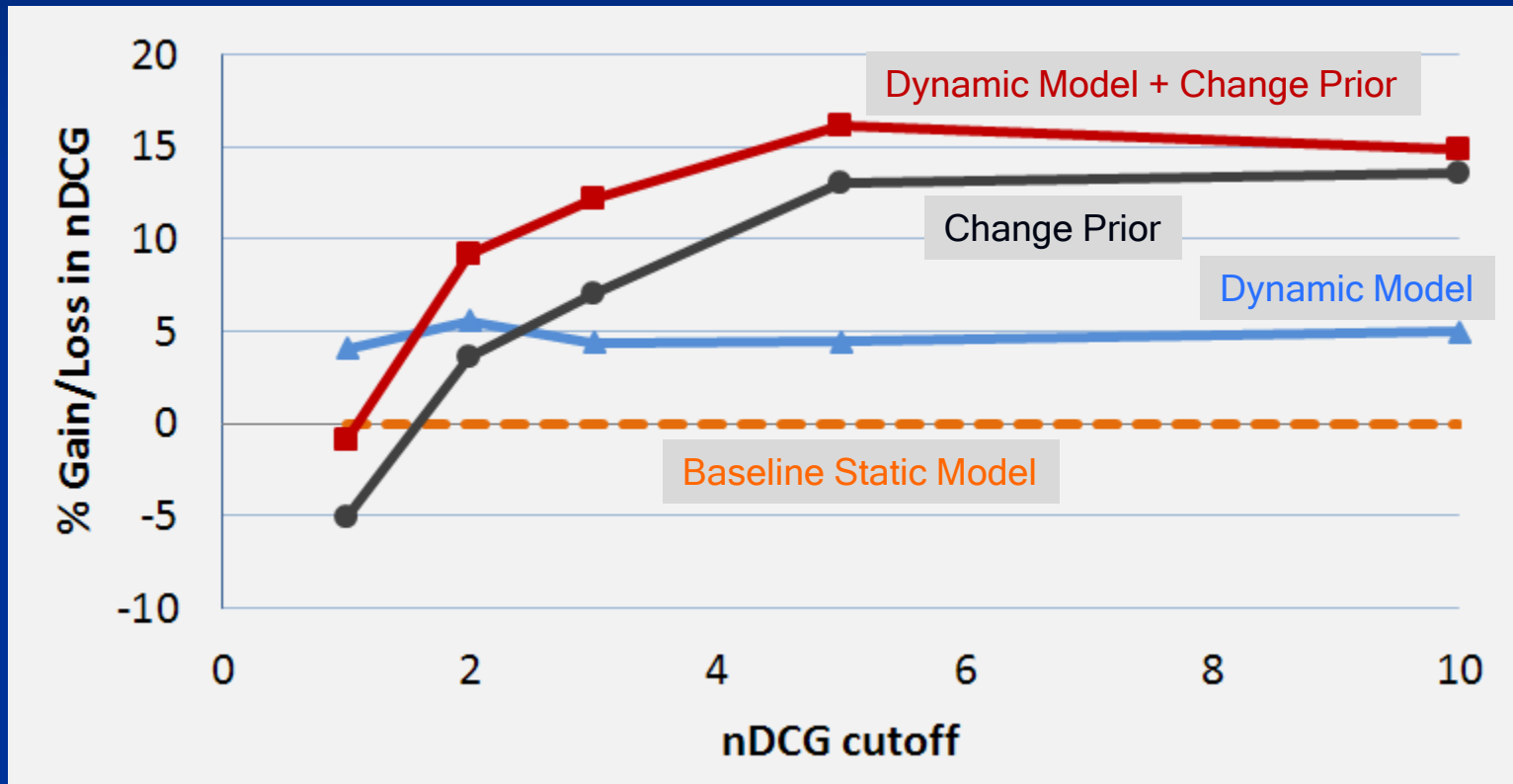
$$P(D | Q) = P(D) \cdot P(Q | D)$$

↑  
Term longevity

# Evaluation: Queries & Documents

- 18K Queries, 2.5M Judged Documents
  - 5-level relevance judgment (Perfect ... Bad)
- 2.5M Documents crawled weekly for 10 wks
- Navigational queries
  - 2k queries identified with a “Perfect” judgment
  - Assume these relevance judgments are consistent over time
- Measure changes in nDCG

# Experimental Results



# Temporal Retrieval Models 2

## (behavior-based)

### ■ Initial evaluation

- Navigational queries; assume relevance is “static” over time

### ■ But, relevance often changes over time

- E.g., *Super Bowl* -- in 2012 vs. in 2011
- E.g., *US Open 2012* -- in May (golf) vs. in Sept (tennis)
- E.g., *March madness 2012* -- before/during/after event
  - Before event: Schedule and tickets, e.g., stubhub
  - During event: Real-time scores, e.g., espn, cbssports
  - After event: General sites, e.g., wikipedia, ncaa

### ■ Current evaluation

- Collect explicit and implicit relevance judgments, query frequency, interaction data, and page content over time

# Relevance over Time

- Query: *football*  
[season Sep - Jan]
- Need to model  
time of query,  
pages and  
events

The screenshot shows a Bing search results page for the query "football". The search bar at the top contains the word "football" and the Bing logo. Below the search bar, there are navigation links for "Web", "Visual Search", "Shopping", "News", "Images", "Videos", and "More". The page displays "ALL RESULTS" for "1-10 of 176,000,000 results".

On the left side, there are sections for "RELATED SEARCHES" (Walter Football, Arena Football, Play Football Games, NFL Football Players, NFL Football Standings, English Football Results, Football Teams, Football Info) and "SEARCH HISTORY" (football, nfl, jeopardy, sigir). There are also options to "See all" and "Clear all · Turn off".

The main results area shows a "Shop for football" section with a table of products:

Category	Price	Brand
Uniforms & Clothing	below \$10	Wilson Sporting Goods
Balls	\$10-\$60	Team Effort
Protective Equipment	above \$60	Under Armour
		Electronic Arts

Below the shopping section, there are several search results:

- NFL.com - Official Site of the National Football League**: The official source for NFL news, video highlights, fantasy football, game-day coverage, schedules, stats, scores and more. [www.nfl.com](http://www.nfl.com) · Mark as spam
- Football.com | Scores, Video, Live Sportscasts, Headlines, News ...**: Get the latest Football News, Scores, Video, Live Sportscasts, Headlines, Stats, Fantasy Football, Hotties, Sports Rumors & More. [football.com](http://football.com) · Mark as spam
- Football - Wikipedia, the free encyclopedia**: Common elements · Etymology · Early history · Establishment of ... The game of **football** is any of several similar team sports, of similar origins which involve advancing a ball into a goal area in an attempt to score. [en.wikipedia.org/wiki/Football](http://en.wikipedia.org/wiki/Football) · Mark as spam
- American football - Wikipedia, the free encyclopedia**: History · Rules · Players · Basic strategy American **football**, known in the United States as **football**, is a sport played between two teams of eleven. The objective of the game is to score points by advancing the ... [en.wikipedia.org/wiki/American\\_football](http://en.wikipedia.org/wiki/American_football) · Mark as spam
- FIFA.com - Fédération Internationale de Football Association (FIFA)**: The official site of the international governing body of **football** with news, national associations, competitions, results, fixtures, development, organisation, world ... [www.fifa.com](http://www.fifa.com) · Mark as spam

At the top of the results, there are four player portraits with captions: Tim Tebow, Kyle Boller, Reggie Bush, and Tom Brady. A link "Compare NFL players: passing, rushing, receiving, sacks" is also visible.

Feb - Aug

Sep - Jan

Feb - Aug



# Relevance over Time

- Query: *sigir*
- Why is old content ranked higher?
  - User interaction data more prevalent for older documents
  - E.g., query-clicks, anchor text, etc.
- Need to weight user behavior signals appropriately

Web Images Videos Shopping News Maps More | MSN Home

bing MS Beta 0 sigir

Web More ▾

RELATED SEARCHES

- SIGIR 2010
- SIGIR Iraq
- SIGIR Reports
- Special Inspector General for Iraq
- Special Interest Group on Information Retrieval
- Sigar
- KDD 2010
- ACM SIGIR

SEARCH HISTORY

Search more to see your history

See all

Clear all · Turn off

ALL RESULTS 1-10 of 167,000 results · All

**Welcome to SIGIR Home**  
An Iraqi fisherman pushes his boat off-shore to depart on his daily fishing trip. View the Report.  
[www.sigir.mil](#) · Mark as spam

**ACM SIGIR Special Interest Group on Information Retrieval Home Page**  
Welcome to the ACM SIGIR Web site. ACM SIGIR addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization ...  
[www.sigir.org](#) · Mark as spam

**home [ACM SIGIR 2010]**  
ACM-SIGIR 2010 was held at UniMail, Geneva, Switzerland between 19th  
Thanks to all the participants!!! The story continues with ACM-SIGIR 2011.  
[www.sigir2010.org](#) · Mark as spam

**SIGIR 2009 Archive | SIGIR'09**  
The SIGIR 2009 conference ran July 19-23, 2009, in Boston, Massachusetts  
Hotel and Northeastern University. The conference was chock full of ...  
[sigir2009.org](#) · Mark as spam

**Welcome to The 34th Annual ACM SIGIR Conference**  
The 34 th Annual ACM SIGIR Conference Important Dates. 17 Jan 2011 :  
papers due; 24 Jan 2011 : Full research paper submissions due  
[sigir2011.org](#) · Mark as spam

**SIGIR'08 - Singapore**  
SIGIR is the major international forum for the presentation of new research results and for the demonstration of new systems and techniques in the broad field of ...  
[www.sigir2008.org](#) · Mark as spam

**About SIGIR**  
About SIGIR The Office of the Special Inspector General for Iraq Reconstruction (SIGIR) is the successor to the Coalition Provisional Authority Office of ...  
[www.sigir.mil/about/index.html](#) · Mark as spam

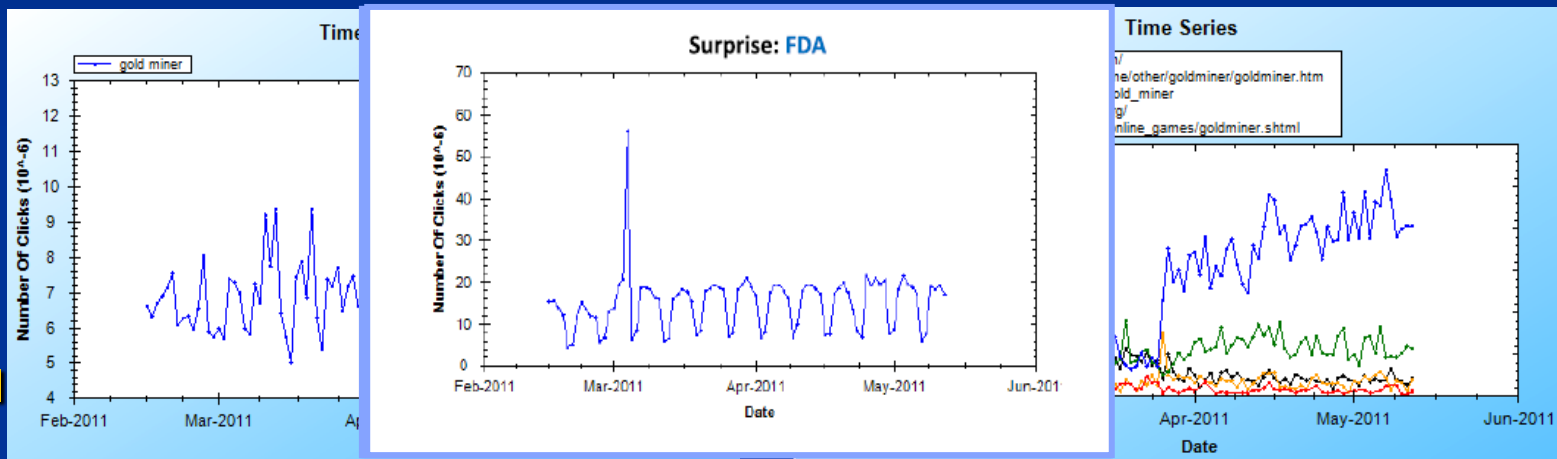
**ACM SIGIR - News**  
SIGIR News 2011. Get ready for SIGIR 2011 in Beijing, China! SIGIR invites applications for Donald B. Crouch Travel Grants to help cover the cost of travel, living ...  
[www.sigir.org/news.html](#) · Mark as spam

**SIGIR - Wikipedia, the free encyclopedia**  
SIGIR may refer to: Special Inspector General for Iraq Reconstruction; Special Interest Group on Information Retrieval, a Special Interest Group (SIG) of the Association for ...  
[en.wikipedia.org/wiki/SIGIR](#) · Mark as spam

# Experimental Setup

## ■ Data

- Queries and clicked URLs, over 4 months



- Actual user search behavior over time (implicit measure)

- Model temporal dynamics of behavior
- Use model to improved ranking

# Time Series Modeling

- Model search behavior as time series
  - Assume that the series of behavioral observations  $Y_1 \dots Y_n$  is generated sequentially based on some underlying structure (e.g., a sequence of *state vectors*)

- Linear State Space Model (SSM)

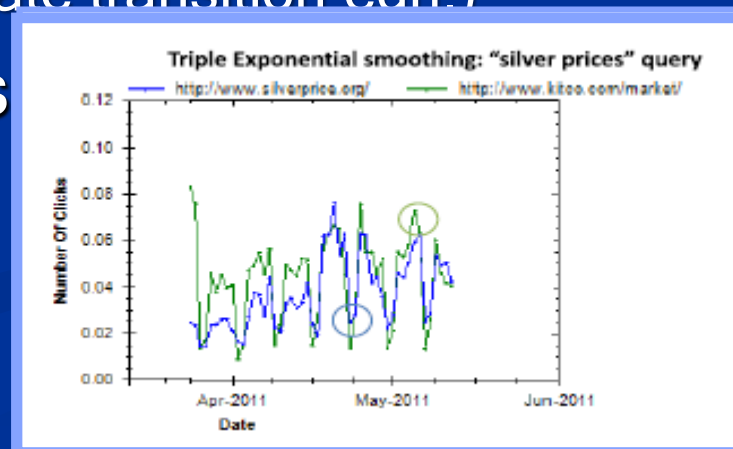
- Let  $X_t$  be a state vector at moment of time t, then a semi-linear state space model is defined by:

$$Y_t = w(\theta)X_t + \epsilon_t \quad (\text{observation eqn.})$$

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

- Model state with Holt-Winters

- Smoothing
- Trend (+Level)
- Periodic/Seasonal



# Experimental Details

- Train: Learn time series models
- Predict: Future query and click behavior
- Ranking models
  - Predicted clicks as the only feature for ranking
  - Temporal features (+other features) as input to learned ranker
- Three types of features
  - No user behavior (i.e., just content)
  - Historical average of user behavior
    - Uniform, Linear, Power
  - Temporal models of user behavior
    - Smoothing, +Trend, +Trend+Periodicity
- Measure: Correlation (predicted vs. actual) rankings

# Experimental Results

## ■ Predicted clicks as the only feature

Query Type	Baselines		
	Average	Linear weight	Power weight
General	0.91	0.92	0.93
Tail	<b>0.18</b>	<b>0.21</b>	<b>0.22</b>
Periodic	0.91	0.92	0.93
Dynamic	<b>0.28</b>	<b>0.35</b>	<b>0.38</b>
Alternating	0.80	0.82	0.84
Temp Reform	0.95	0.95	0.95

Table 2: Pearson correlation on ordering of our temporal models compared to baseline models. Statistically significant differences based on a paired t-test ( $p < .05$ ) are shown in bold.

## ■ Ranker trained with temporal features

Query Type	No User Behavior	Baseline Models		
	Base Features	Base Features +Average	Base Features +Linear weight	Base Features +Power weight
General	0.47	0.97	0.98	0.98
Tail	<b>0.31</b>	0.20	0.07	0.02
Periodic	0.78	0.87	0.91	0.91
Dynamic	-0.08	0.30	0.30	0.39
Alternating	0.23	0.64	<b>0.90</b>	0.74
Temp Reform	0.19	0.73	0.97	0.96

Table 4: Pearson Correlation on ranking using Base features without user behavior, with statistical significance using our temporal models. Statistically significant differences based on a paired t-test when comparing our performing algorithm ( $p < .05$ ) are shown in bold.

## ■ Best-performing queries

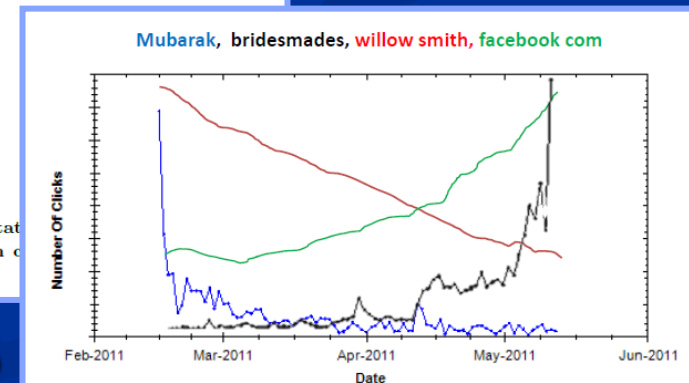
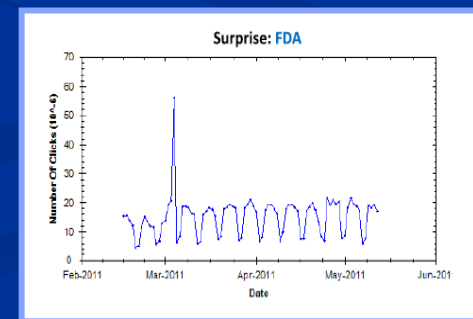


Figure 6: Dominant query shapes for queries where temporal model yielded better rankings than baseline rankers.

# Temporal IR Summary

- Goal: Improve Web retrieval by modeling temporal dynamics
- Content-based models
  - Rate of page change
  - Detailed term-level changes
- Behavior-based models
  - Query frequency over time
  - Click patterns over time
- Ongoing work
  - Combine content and behavior features
  - Surprise detection
  - Snippet generation
  - ...



# Other Examples of Dynamics and Information Systems

## ■ Temporal retrieval models

- Radinski et al. (submitted); Elsas & Dumais (2010); Liu & Croft (2004); Efron (2010); Aji et al. (2010)

## ■ Document dynamics, for crawling and indexing

- Adar et al. (2009); Cho & Garcia-Molina (2000); Fetterly et al. (2003)

## ■ Query dynamics

- Kulkarni et al. (2011); Jones & Diaz (2004); Diaz (2009); Kotov et al. (2010)

## ■ Extraction of temporal entities within documents

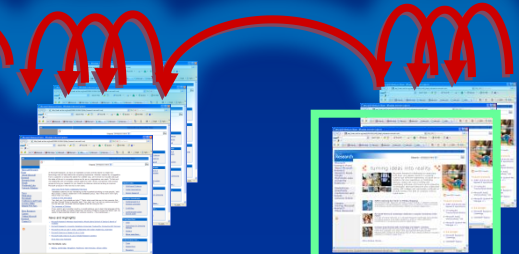
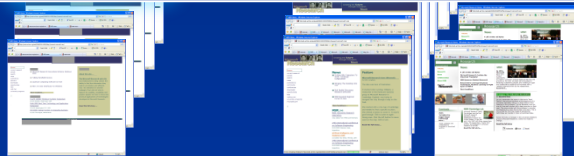
## ■ Protocol extension for retrieving versions over time

- E.g., Memento (Van de Sompel et al., 2010)

# Summary

**Temporal IR:**  
Leverages change  
for improved IR

Web content changes: page-level, term-level

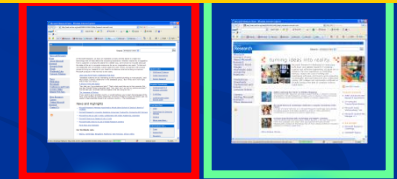


Relating revisitation and change allows us to

- Identify pages for which change is important
- Identify interesting components within a page



People revisit and re-find Web content



**Diff-IE:** Supports (and influences) interaction and understanding

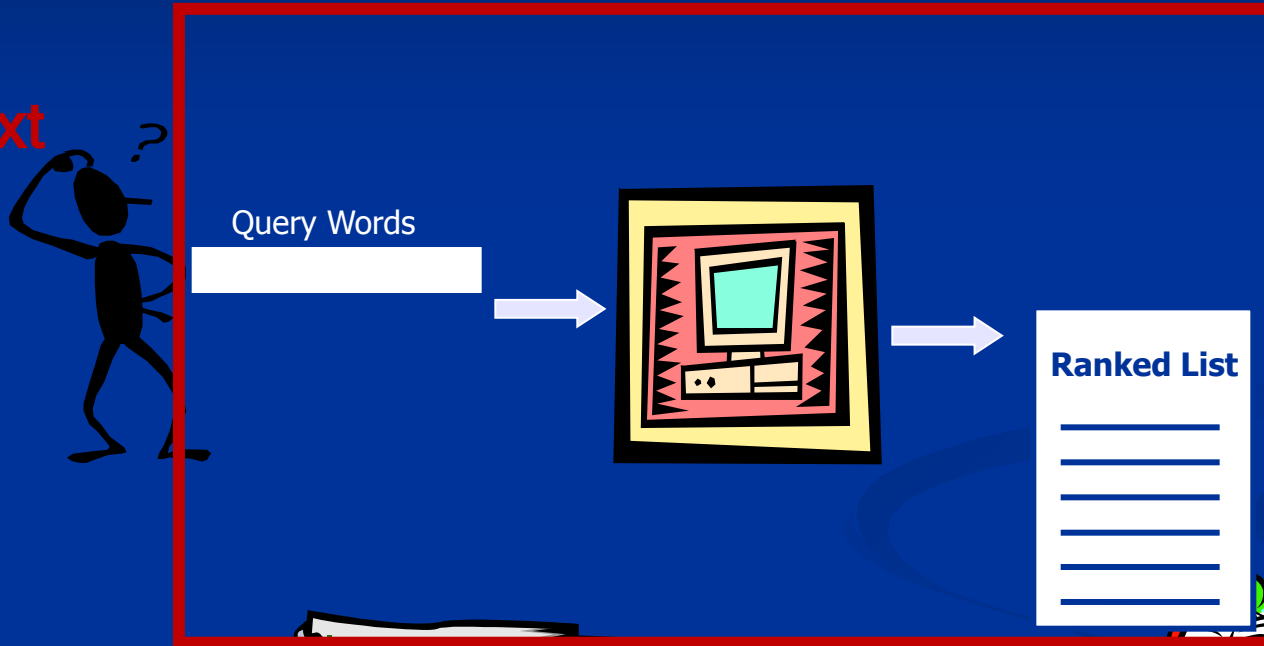


# Challenges and Opportunities

- Temporal dynamics are pervasive in information systems
- Influence many aspect of information systems
  - Systems: protocols, crawling, indexing, caching
  - Document representations: meta-data generation, information extraction, sufficient statistics at page and term-level
  - Retrieval models: term weights, document priors, etc.
  - User experience and evaluation
- Better supporting temporal dynamics of information
  - Requires digital preservation and temporal metadata extraction
  - Enables richer understanding of the evolution (and prediction) of key ideas, relations, and trends over time
- Time is one important example of context for IR
  - Others include: location, individuals, tasks ...

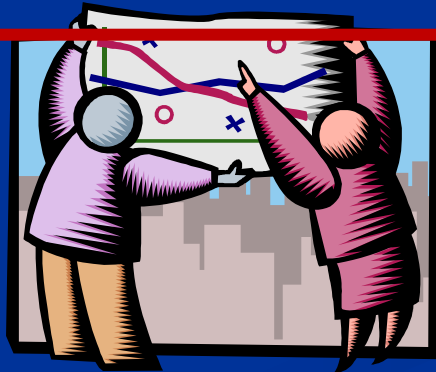
# Think Outside (Search Research) Boxes

User  
Context



Document  
Context

Task/Use  
Context



# Thank You !

- Questions/Comments ...

- More info,

<http://research.microsoft.com/~sdumais>

Diff-IE ... try it!

<http://research.microsoft.com/en-us/projects/diffie/default.aspx>

