Temporal Dynamics and Information Systems
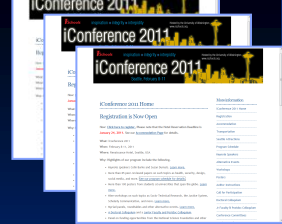
Susan Dumais

Microsoft Research

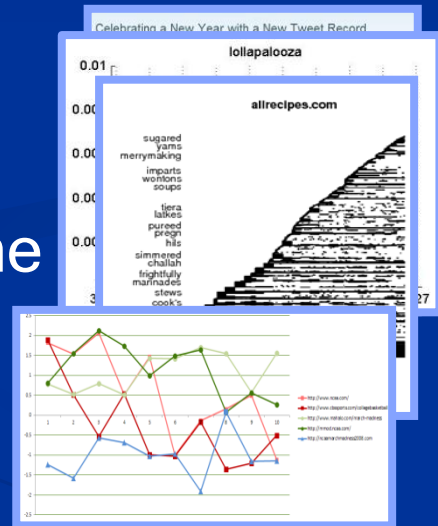http://research.microsoft.com/~sdumais

In collaboration with:

Eric Horvitz, Jaime Teevan, Eytan Adar, Jon Elsas, Ed Cutrell, Dan Liebling, Richard Hughes, Merrie Ringel Morris, Evgeniy Gabrilovich, Krysta Svore, Anagha Kulkani
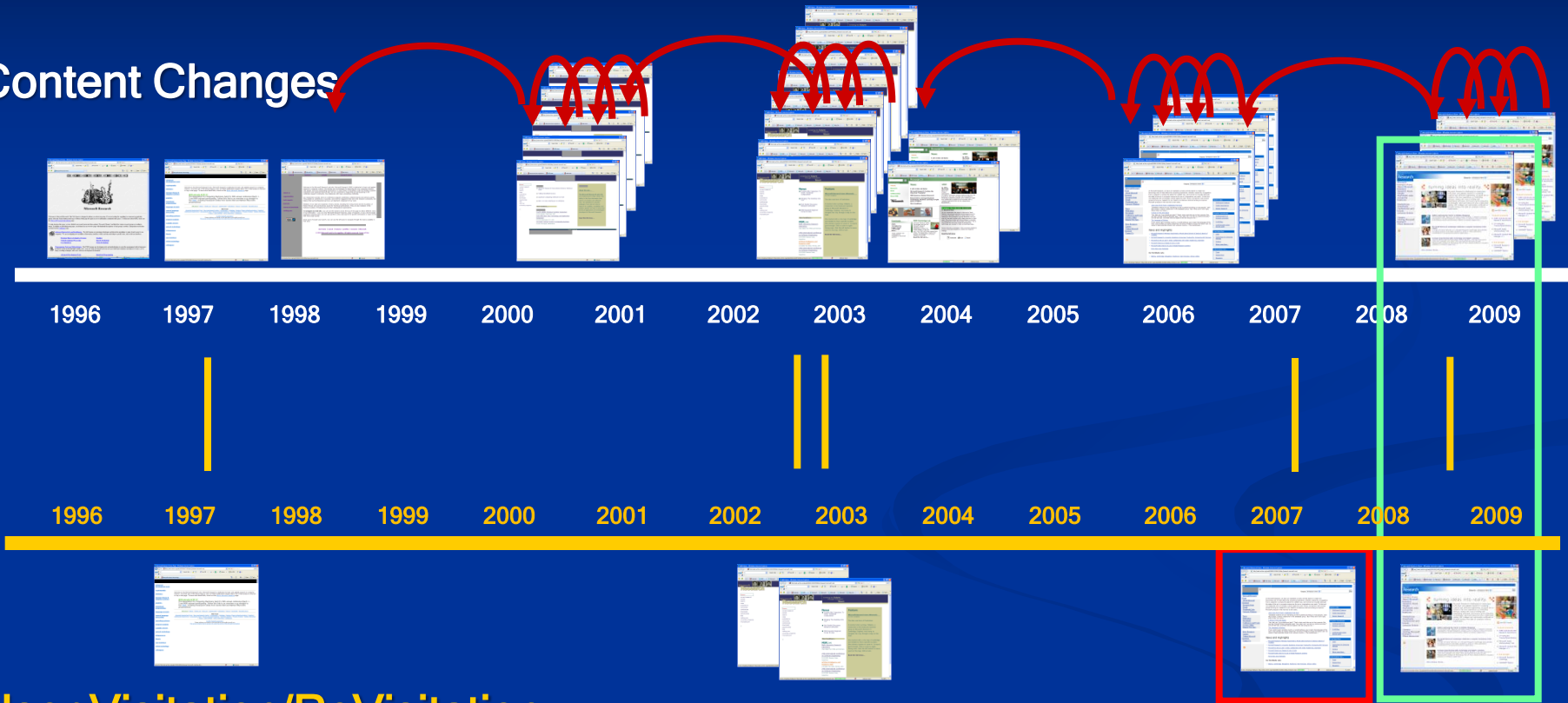
Information Dynamics

- **Many differences between physical & digital libraries**

- **Change is everywhere in digital information systems**
  - New documents (and queries) appear all the time
  - Query volume changes over time
  - Document content changes over time
  - What's relevant to a query changes over time
    - E.g., *U.S. Open 2010* (in May vs. Sept)
    - E.g., *Hurricane Earl* (in Sept 2010 vs. before/after)
  - User interaction changes over time
    - E.g., tags, anchor text, social networks, query-click streams, etc.

- **Change is pervasive in digital information systems
  … yet, we're not doing much about it !**

Information Dynamics

Content Changes

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

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

User Visitation/ReVisitation

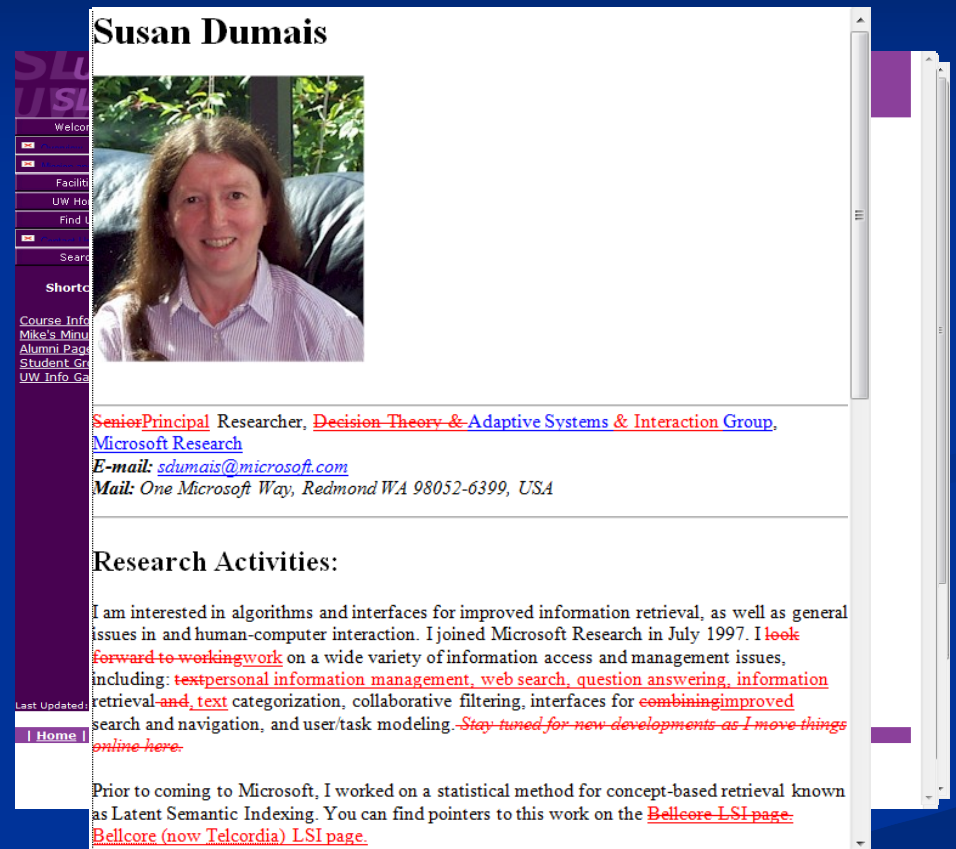**Today's Browse and Search Experiences**

But, ignores …

Digital Dynamics Easy to Capture

- Easy to capture

- But ... few tools support dynamics

Susan Dumais

~~Senior~~Principal Researcher, ~~Decision Theory &~~ Adaptive Systems & Interaction Group, Microsoft Research
E-mail: sdumais@microsoft.com
Mail: One Microsoft Way, Redmond WA 98052-6399, USA

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 ~~Bellcore LSI page.~~ Bellcore (now Telcordia) LSI page.

Overview

- Characterize change in digital content
  - Content changes over time
  - People re-visit and re-find over time
- Improve retrieval and understanding
  - Examples from our work on search and browser support … but more general
    - Desktop: Stuff I've Seen; Memory Landmarks; LifeBrowser
    - News: Analysis of novelty (e.g., NewsJunkie)
    - Web: Tools for understanding change (e.g., Diff-IE)
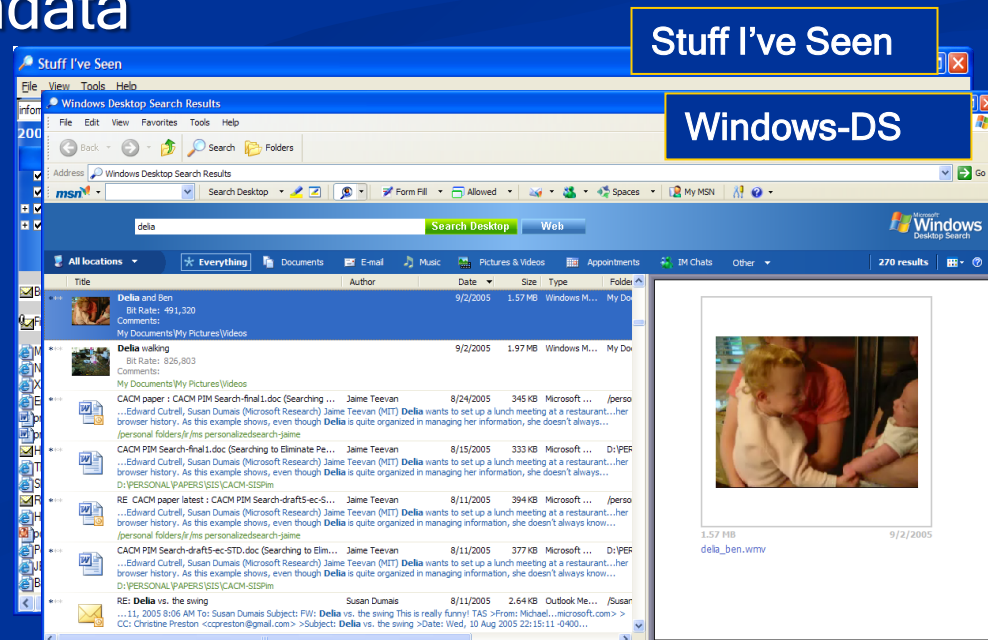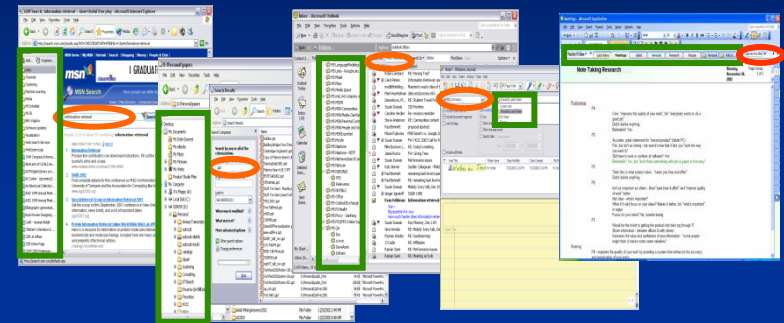   - Web: Retrieval models that leverage dynamics

Stuff I've Seen (SIS)

- ### Many silos of information

- ### SIS:
  - #### Unified access to distributed, heterogeneous content (mail, files, web, tablet notes, rss, etc.)
  - #### Index full content + metadata
  - #### Fast, flexible search
  - #### Information re-use

  - #### SIS ->
    #### Windows Desktop Search

Stuff I've Seen

Windows-DS

Example Desktop Searches

Lots of metadata ... especially time

Looking for: *recent email from Fedor that contained a link to his new demo*
Initiated from: Start menu
Query: from:Fedor

Looking for: *the pdf of a SIGIR paper on context and ranking (not sure it used those words) that someone (don't remember who) sent me about a month ago*
Initiated from: Outlook
Query: SIGIR

Looking for: *meeting invite for the last intern handoff*
Initiated from: Start menu
Query: intern handoff kind:appointment

Looking for: *C# program I wrote a long time ago*
Initiated from: Explorer pane
Query: QCluster*.*

Stuff I've Seen: Findings

- **Studied using:** free-form feedback, questionnaires, usage patterns from log data, in situ experiments, lab studies for richer data

- **Personal stores:** 5k-1500k items [SD: 100k items; 1k new items/wk]

- **Information needs:**
  - Desktop search != Web search
  - <u>People</u> are important – 29% queries involve names/aliases
  - <u>Date</u> is the most common sort order, even w/ "best-match" default
    - Few searches for "best" matching object
    - Many other criteria (e.g., time, people, type), depending on task
    - Need to support flexible access
  - Abstractions important – "useful" date, people, pictures
  - <u>Age of items retrieved</u>
    - Today (5%), Last week (21%), Last month (47%)
    - Need to support episodic access to memory

$Log(Freq) = -0.68 * log(DaysSinceSeen) + 2.02$
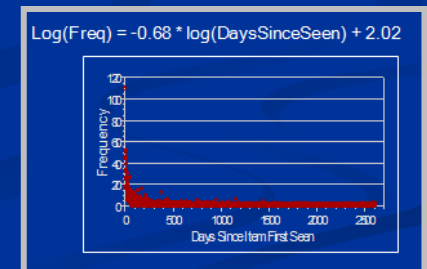
Memory Landmarks

- Importance of episodes in human memory
    - Memory organized into episodes (Tulving, 1983)
    - People-specific events as anchors (Smith et al., 1978)
    - Time of events often recalled relative to other events, historical or autobiographical (Huttenlocher & Prohaska, 1997)
- Identify and use landmarks facilitate search and information management
    - Timeline interface, augmented w/ landmarks
    - Learn Bayesian models to identify memorable events
- Extensions beyond search, e.g., Life Browser

Memory Landmarks

Distribution of Results Over Time

Stuff I've Seen

QUERY:
earthquake

submit   59 results found

⦿ AND   ○ fuzzy match
○ OR   ⦿ exact match

☐ show advanced controls

Search Results

9/12/2002
Talk at UUUU
Easter Day   4/15/2001
              4/7/2001

📊 history.xls

CHI   4/3/2001   Power Point Presentation
ACM1 SanJose 3/10/2001   from Michael B. Smith — Re: Greetings
1/1/2002   from John Smith — RE: Followup from Earthquake Meeting
                            from Tim Stevens— RE: followup from the "earthquake meeting"
                            from Janet S. Hazeltine — Re: Rockin and Rollin ...
                            from Bart Thomason— Followup from Earthquake Meeting
                            from Bart Thomason— Greetings
                            re jcdl 2002-3.txt
                            from Samuel Atkins French — Re: Shakin Sue ...
                            from Jennifer Rogerson — Re: lost letters
1/1/2001   from Bart Thomason— RE: quake?
                            from Edward Finerhold— A bit shaken?
                            Proceedings Template - WORD
                            from John Smith — RE: followup from the "earthquake meeting"
                            from Bart Thomason— Earthquake Report
                            from Mike Yin — RE: Relevance test review 1
                            from Jenny Henson— are you ok?
                            from Cargon, Angelina — FW: Earthquake
Seattle Earthquake measures 2/28/2001   from Trent van Erliche (MSR)— Earthquake coverage
6.8   from Phillip Lander— Earthquake
              2/24/2001
1/1/2000

1/1/20

Memory Landmarks
- General (world, calendar)
- Personal (appts, photos)

Linked to results by time

Summer Olympics in Sydney, '9/15/2
Australia   8/1/20
1/1/1999
SIGIR and vacation   7/21/2
                            1/1/20
World Celebrates New   1/1/2000
Millennium
vacation   12/20/1999   from Christine Borgman — RE: Seattle plans
1/20/1998   sigir99_report.txt
                                          1999-earthquake-graph.gif

zoom ◀ | ▶ zoom
in   out

☐ study

iConfer

Memory Landmarks
Learned models of memorability
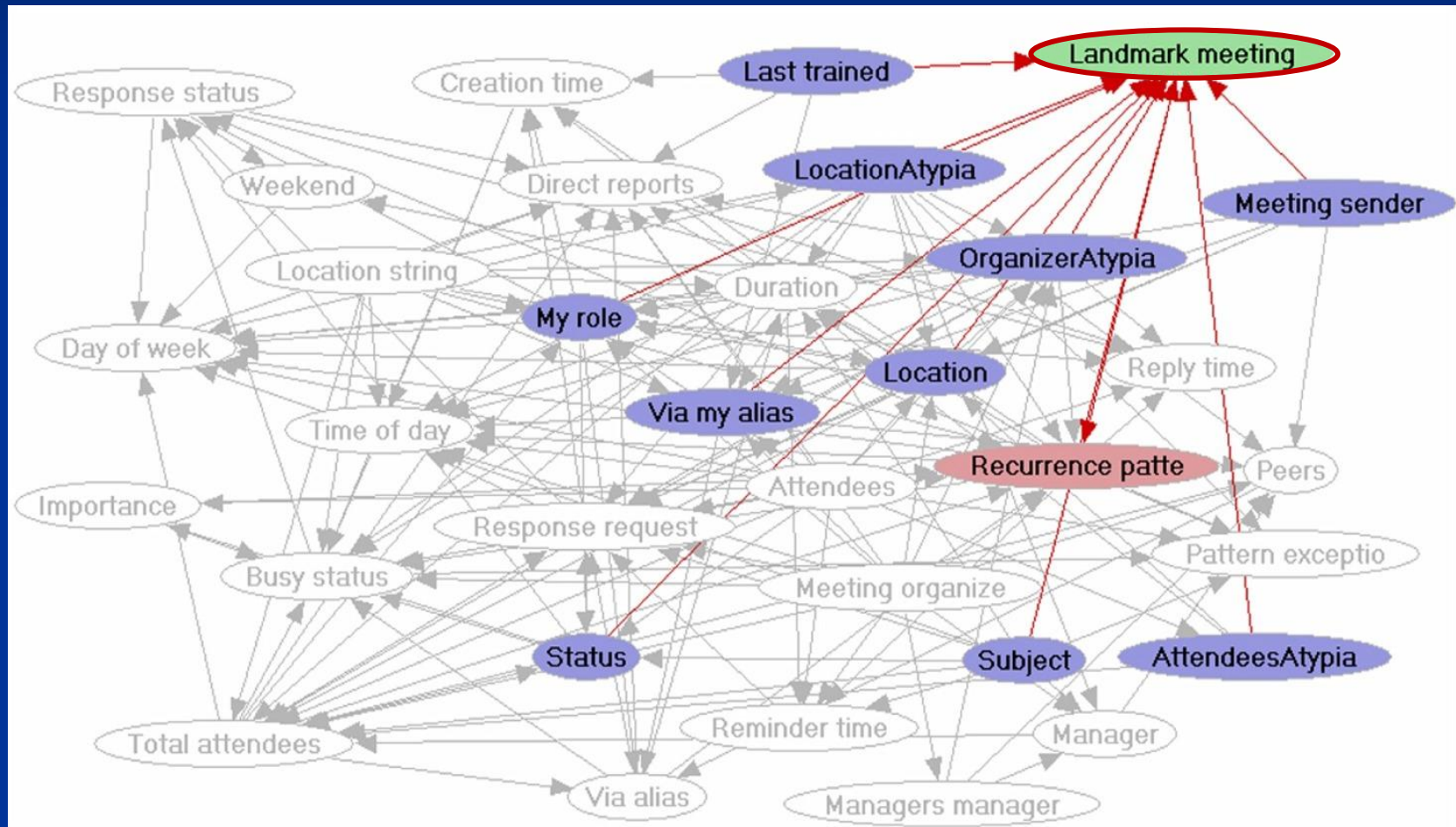
LifeBrowser

Images & videos

Desktop & search activity

Appts & events

Locations

Whiteboard capture

NewsJunkie
Evolution of Context over Time

- **News is a stream of information w/ evolving events**
  - **But, it's hard to consume it as such**
  - **Personalized news using information novelty**
- Identify clusters of related articles
- Characterize what a user knows about an event
- Compute the novelty of new articles, relative to this background (relevant & novel)
  - Novelty = KLDivergence (article || current_knowledge)
- Use novelty score and user preferences to guide what, when, and how to show new information

NewsJunkie in Action

NewsJunkie:
Pizza delivery man w/ bomb incident
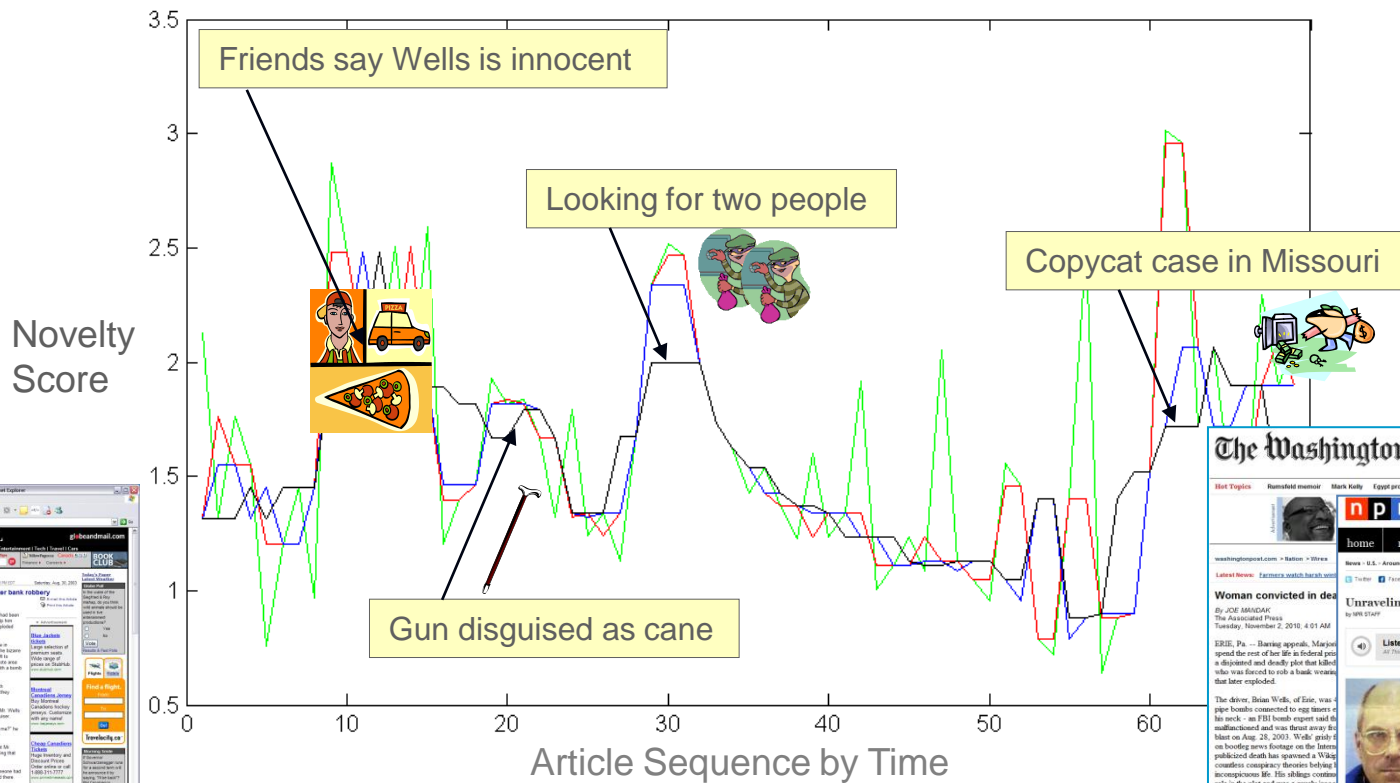
Characterizing Web Change

Content Changes

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

- 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 <u>visited</u> 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
    - 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

Change curve for
http://losangeles.craigslist.org/hsh/
Knot point at 10 hours

Change curve for
http://allrecipes.com/
Knot point at 261 hours
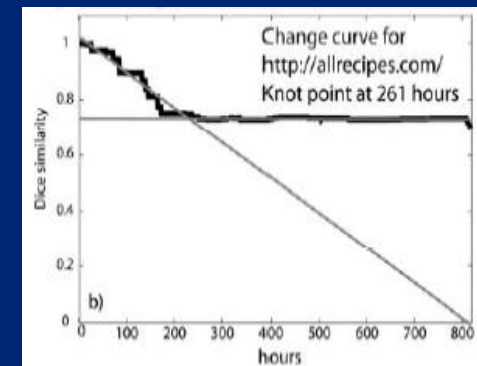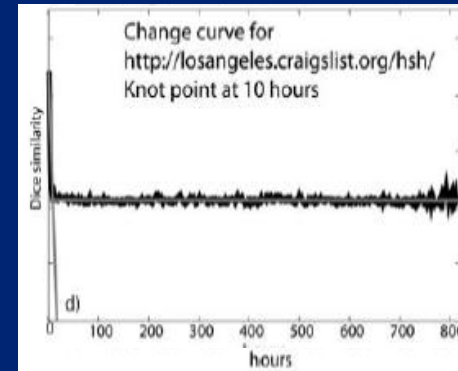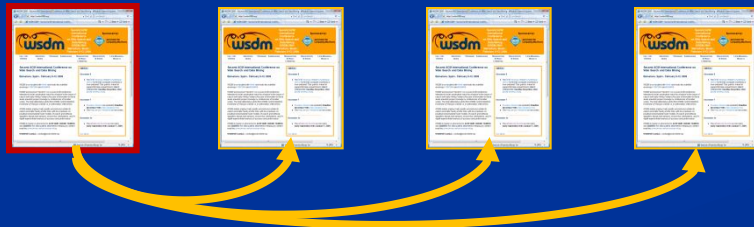
Measuring Within-Page Change

- **DOM-level changes**
- **Term-level changes**
  - **Divergence from norm**
    - cookbooks
    - salads
    - cheese
    - ingredient
    - bbq
    - …
  - **"Staying power" in page**

Sep.   Oct.   Nov.   Dec.

Time

Example Term Longevity Graphs
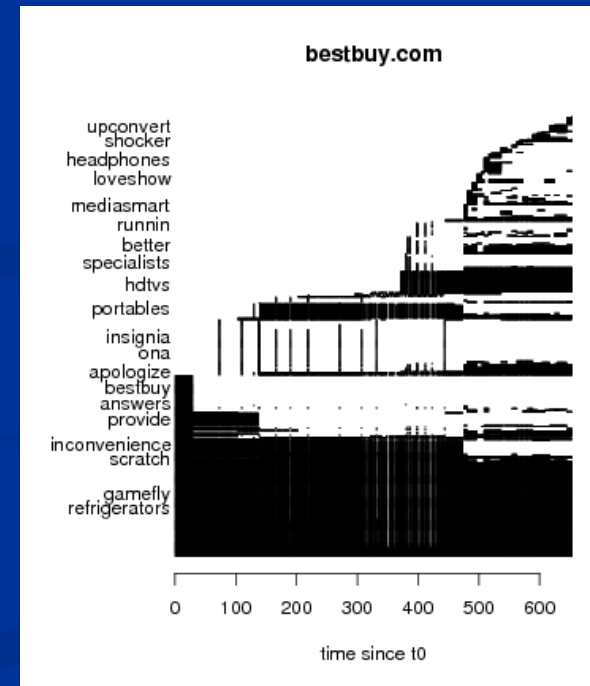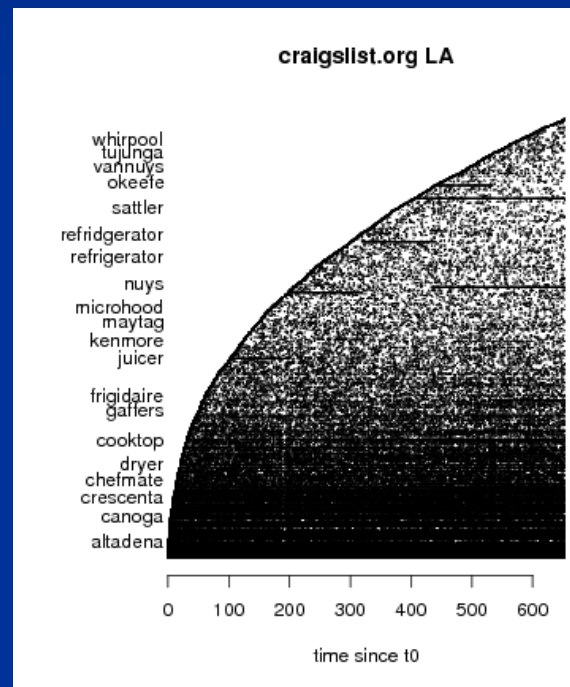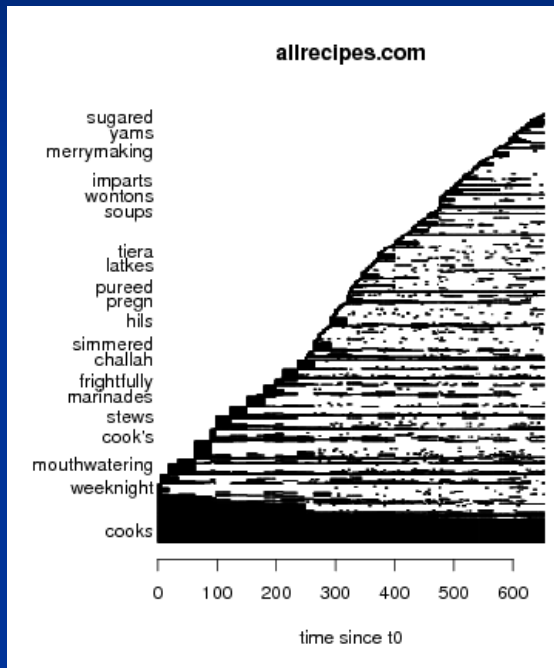
Revisitation on the Web

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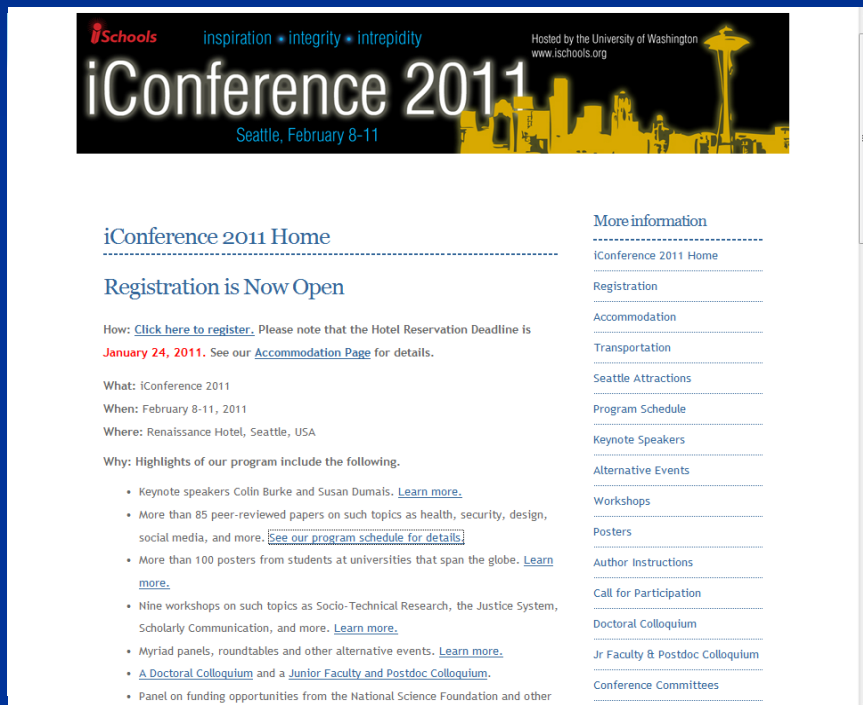
- **Summary metrics**
  - Unique visitors
  - Visits/user
  - Time between visits
- **Revisitation curves**
  - Histogram of revisit intervals
  - Normalized

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

Normalized Count

Time Interval

1
0.8
0.6
0.4
0.2
0
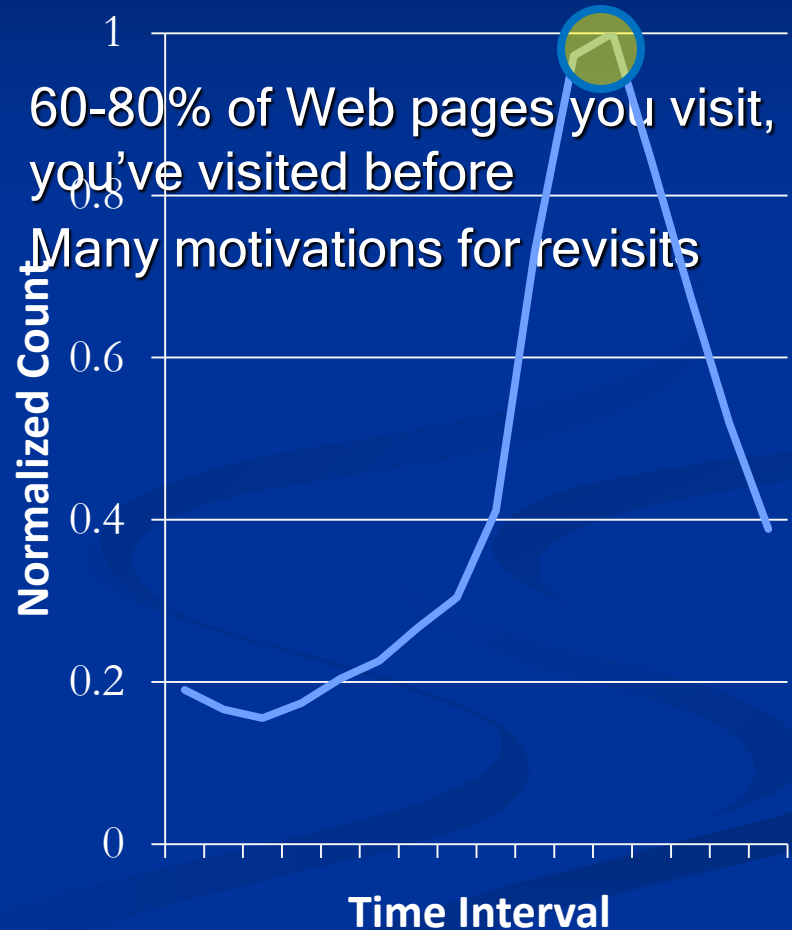
Possible Relationships Between Change and Revisitation

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

Revisitation and Search (ReFinding)

- # Repeat query (33%)
  - Q: *iconference 2011*
- # Repeat click (39%)
  - http://www.ischools.org/iConference11
  - Q: *iconference 2011; iconference*
- # Big opportunity (43%)
  - 24% "navigational revisits"

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

Building Support for Web Dynamics

Temporal IR

Content Changes

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

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

User Visitation/ReVisitation
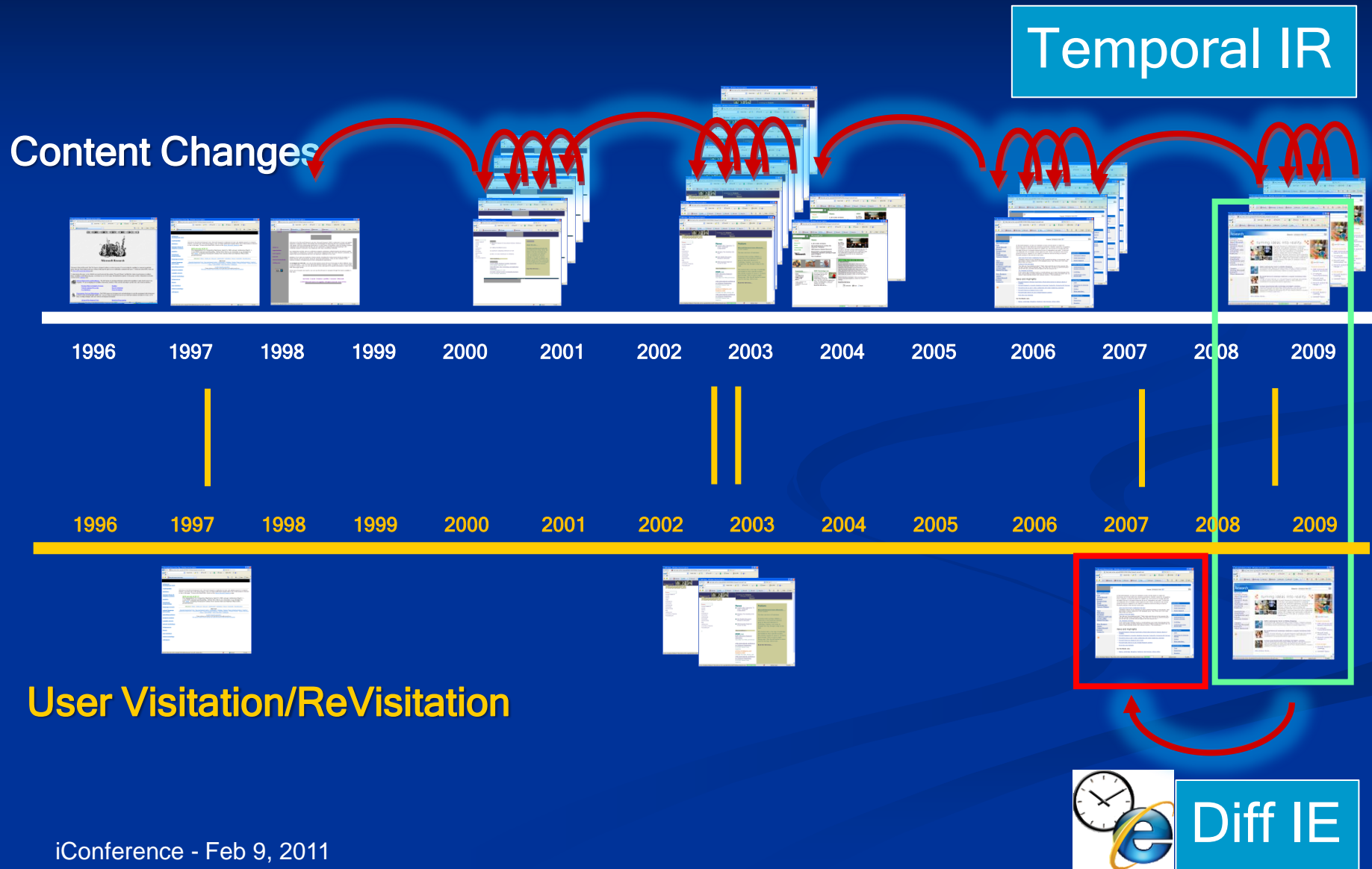
Diff IE

Diff-IE

Diff-IE toolbar
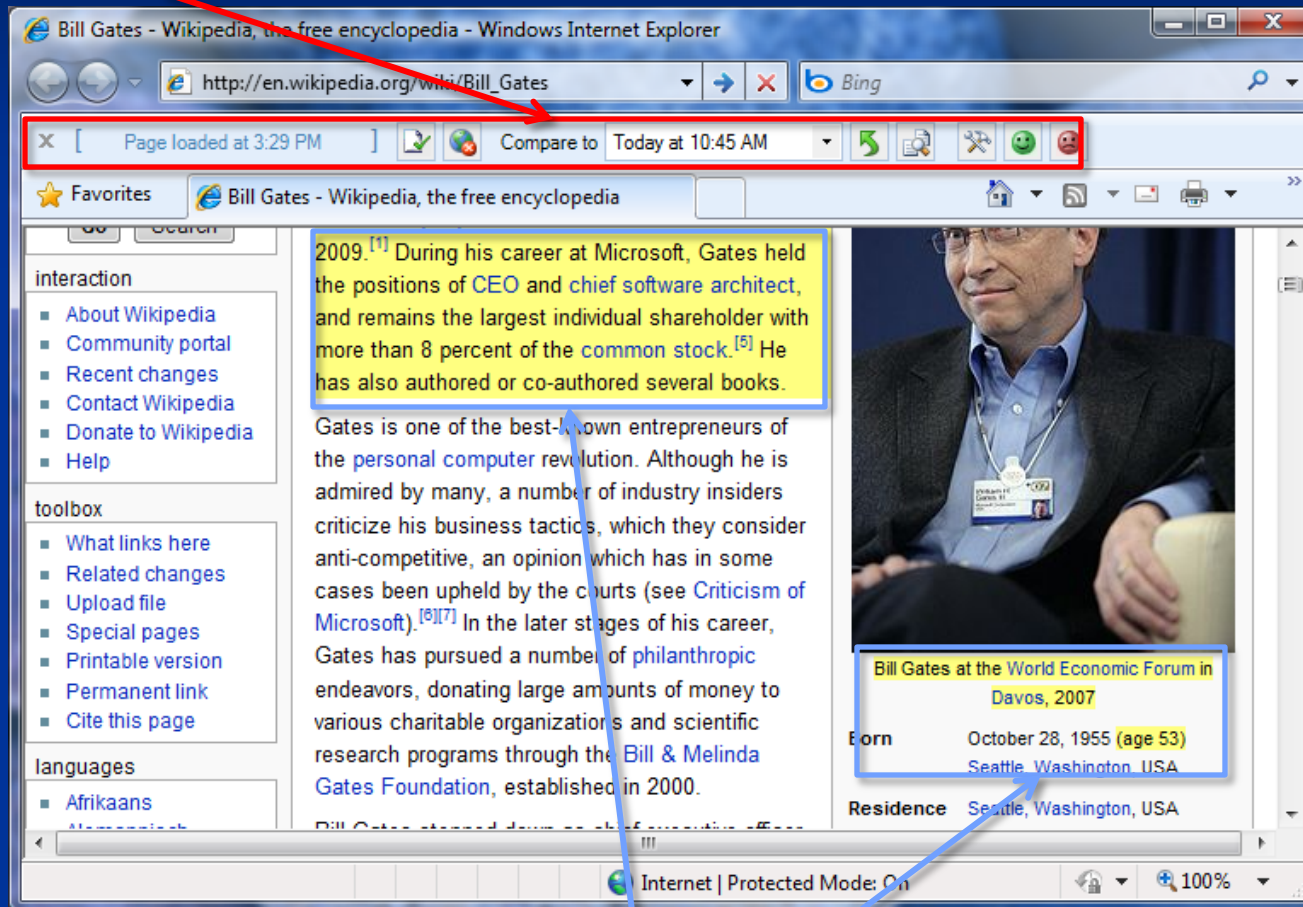
Changes to page since your last visit

Interesting Features of Diff-IE

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

iConference - Feb 9, 2011
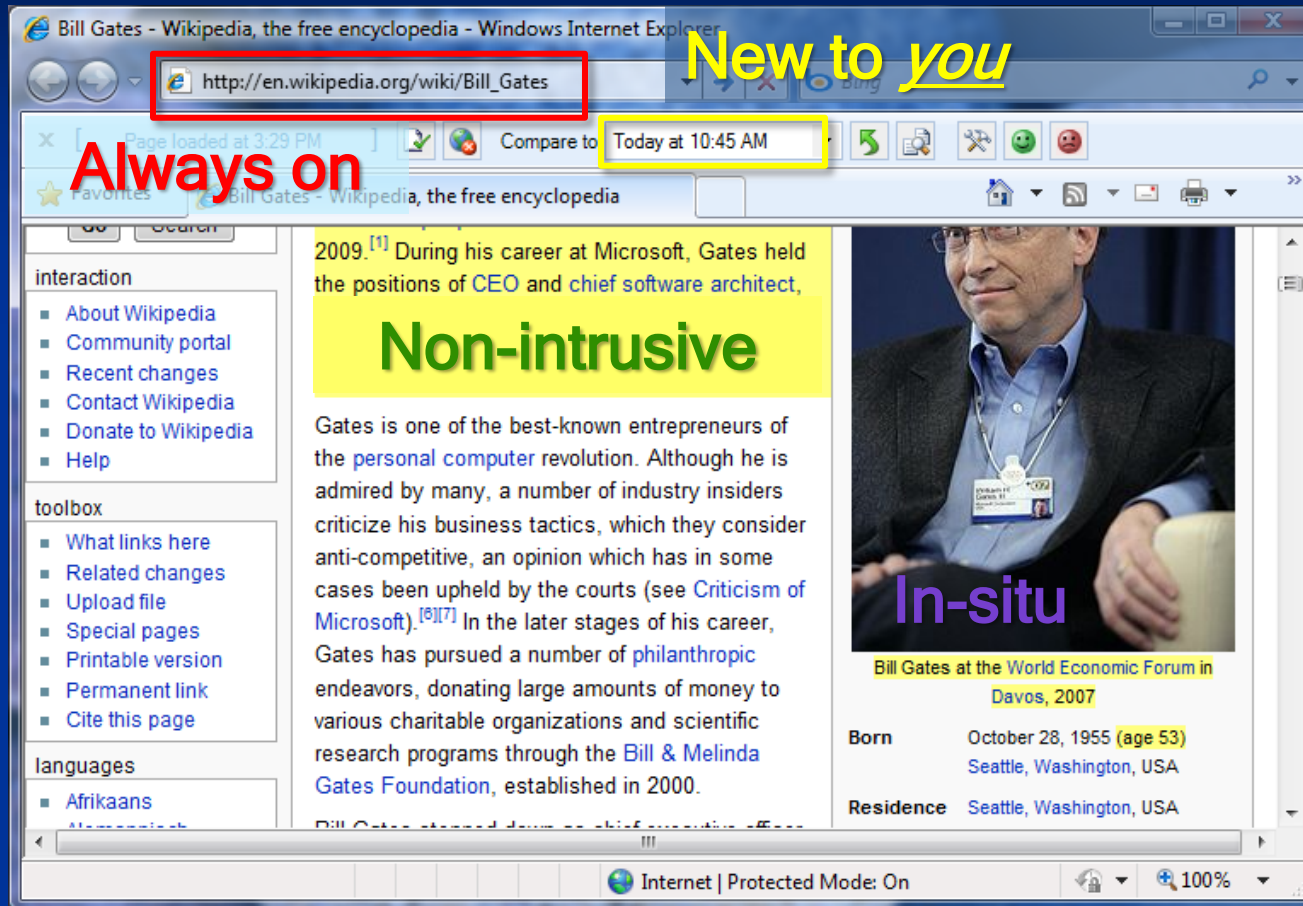
Examples of Diff-IE in Action

Expected New Content

Monitor

Unexpected Important Content

Designated as an architectural masterpiece and home to a collection of one million books and other materials, you'll want to tour this Seattle treasure. Tours will begin at 8 p.m., an hour before the reception, to help ensure that everyone wanting a tour can be accommodated. Arrive at the 5th Avenue entrance of the library and wear your conference badge.

9:00 PM - 11:30 PM: Opening Reception / Seattle Public Library Norcliff Room

Opening Reception at the Norcliffe Room of the Central Library. Catered by Ray's Boathouse, celebrated Seattle restaurant and pioneer of Pacific Northwest Cuisine.

Wednesday, February 9

7:00 AM - 5:00 PM: Registration

8:45 AM - 10:30 AM: Native American Blessing / Welcoming led by Julian Argel (Washington) followed by Keynote Address by Susan Dumais (Microsoft Research)

Location: B Level, Courtyard Ballroom

Please note that the doors will close for a short time at the commencement of our ceremony out of deference to the Native American Blessing. Stragglers will be asked to wait until the blessing has concluded before being permitted to enter.

Session Chair: Jonathan Grudin (Microsoft Research)

10:30 AM - 4:00 PM: Special Event: Microsoft Visit

This half-day visit to the Microsoft campus in Redmond will involve presentations and demos by Microsoft Research staff, as well as tours at the Microsoft Corporate headquarters. Please note that it conflicts with iConference sessions 1 through 12. *UPDATE: this tour is officially full, and space is limited to attendees who already signed up at registration.*

10:30 AM - 11:00 AM: Morning Break

11:00 AM - 12:30 PM:

Session 1: Alternative Event

UNIVERSITY LIBRARIES
UNIVERSITY *of* WASHINGTON

MORGAN & CLAYPOOL PUBLISHERS

texifter

iConference 2011 is hosted by the University of Washington Information School

Information School
UNIVERSITY *of* WASHINGTON

iConference - Feb 9, 2011

Understand Page Dynamics

Expected

Unexpected

Expected
New Content

Unexpected
Important Content

Edit

Attend to Activity

Monitor

Understand
Page Dynamics

Serendipitous
Encounter

Unexpected
Unimportant Content

Studying Diff-IE

- **Feedback buttons**

- **Survey**
  - Prior to installation
  - After a month of use

- **Logging**
  - URLs visited
  - Amount of change when revisited

- **Experience interview**

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
  - Last week: 45.0% of visits are revisits
  - First week: 39.4% of visits are revisits   14%
- 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
  - Last week: 32.4% revisits changed, by 9.5%
  - First week: 21.5% revisits changed, by 6.2%　↑ 51+%
- 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, wikis, etc.
  - Implicit interest via interaction patterns
    - Edit wear and read wear (Hill et al., 1992)
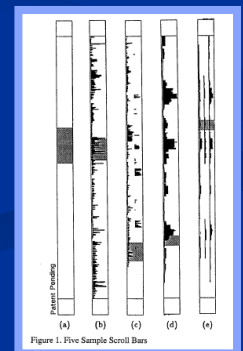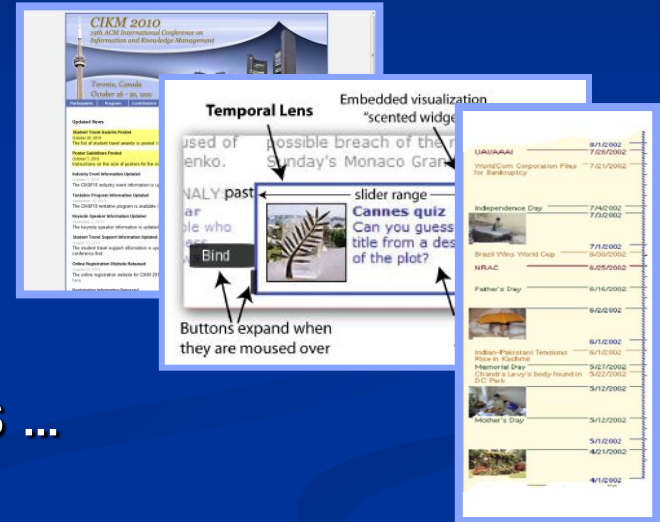
Leveraging Dynamics for Retrieval

Temporal IR

Content Changes

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

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

User Visitation/ReVisitation
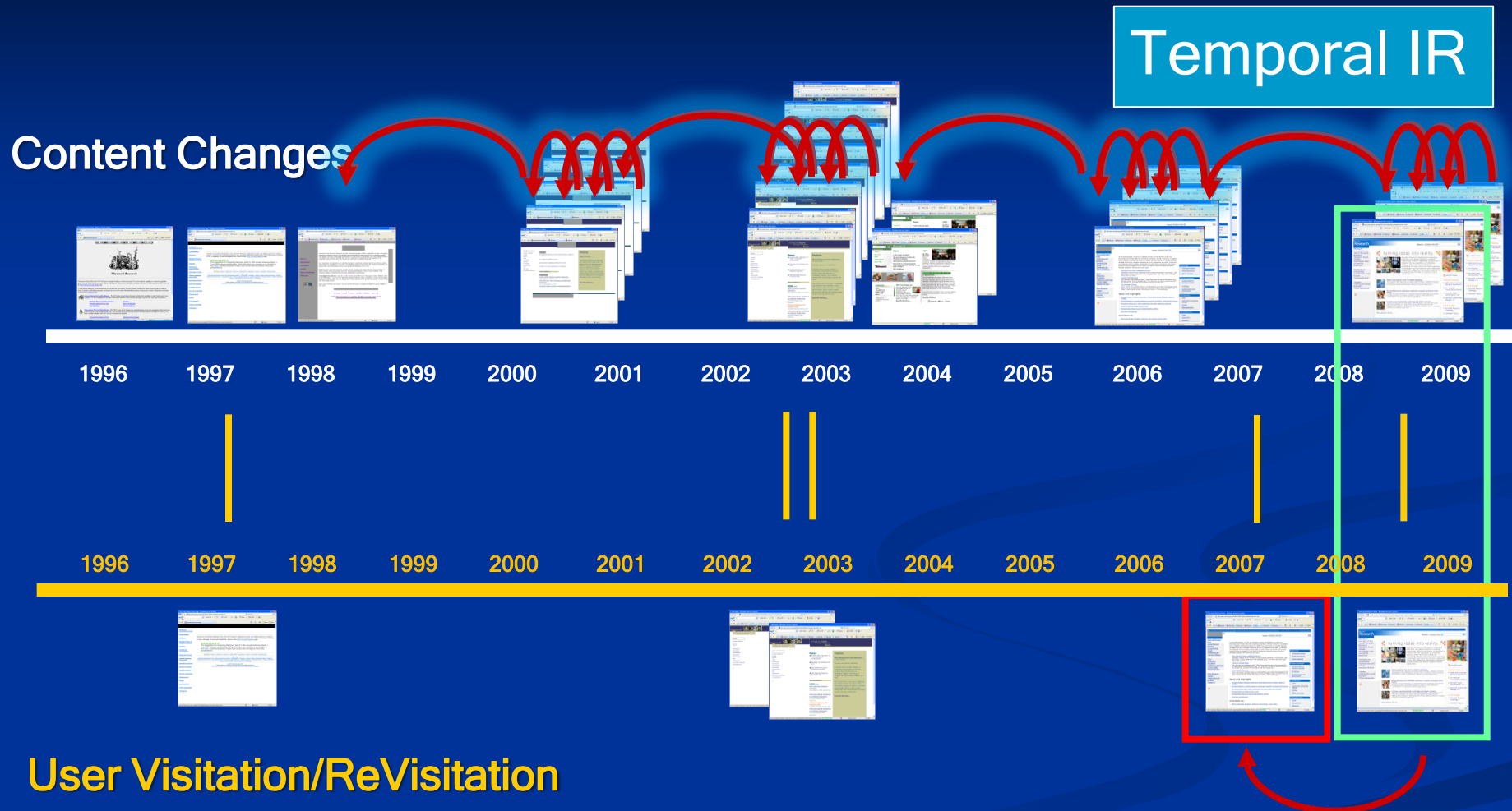
Temporal Retrieval Models

- Current retrieval algorithms look only at a single snapshot of a page

- But, Web pages change over time

- Can we can leverage this to improved retrieval?
  - Pages have different *rates of change*
    - Different priors (using change vs. link structure)
  - Terms have *different longevity (staying power)*
    - Some are always on the page; some transient
  - Language modeling approach to ranking

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

Change prior   Term longevity

Relevance and Page Change

- **<u>Page change</u>** is related to relevance judgments
  - Human relevance judgments
    - 5 point scale – Perfect/Excellent/Good/Fair/Bad
  - Rate of Change -- 60% Perfect pages; 30% Bad pages
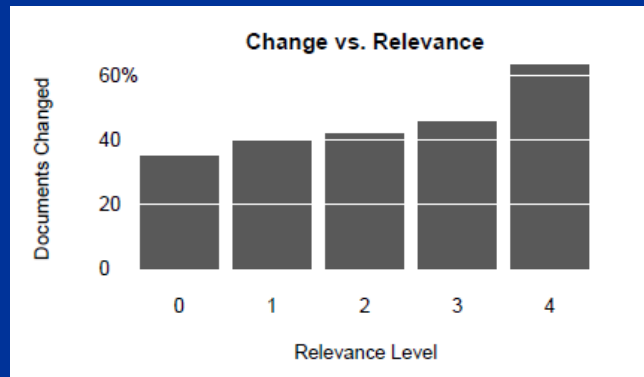
- Use change rate as a document prior (vs. priors based on link structure like Page Rank)
  - Shingle prints to measure change

$$P(D \mid Q) = \boxed{P(D)} \cdot P(Q \mid D)$$

Change prior

Relevance and Term Change

- <u>Terms patterns</u> vary over time

- Represent a document as a mixture of terms with different "staying power"
  - Long, Medium, Short
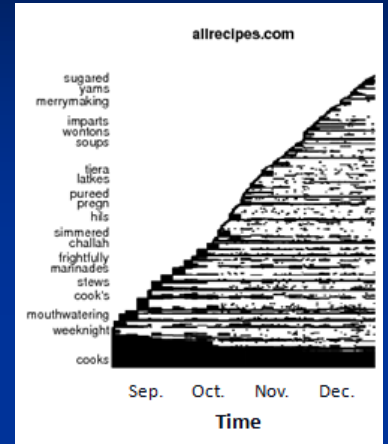
$$P(Q \mid D) = \lambda_L P(Q \mid D_L) + \lambda_M P(Q \mid D_M) + \lambda_S P(Q \mid D_S)$$

$$P(D \mid Q) = P(D) \cdot \boxed{P(Q \mid D)}$$

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

Experimental Results
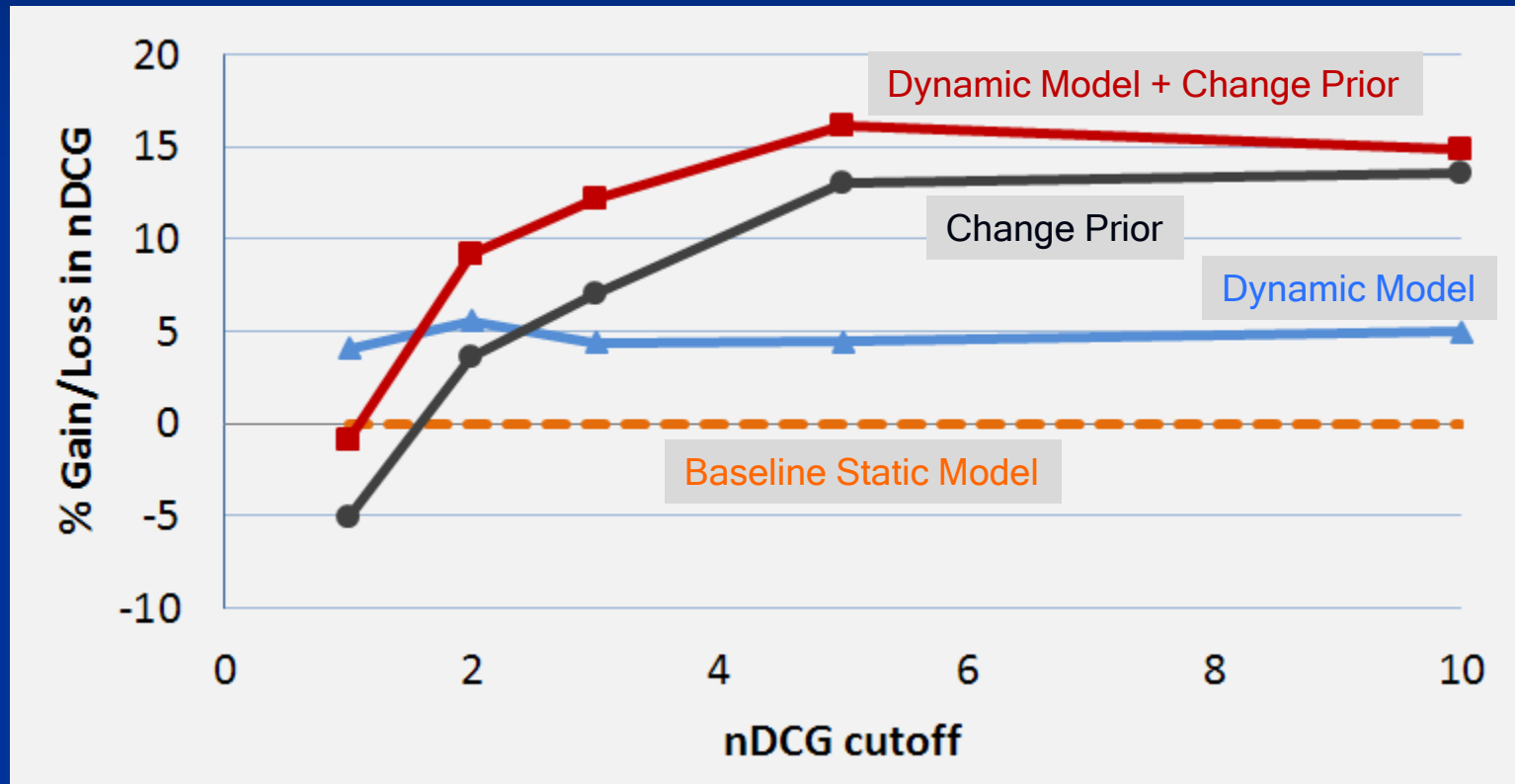
Temporal Retrieval, Ongoing Work

- Initial evaluation
  - Focused on navigational queries
  - Assumed their relevance is "static" over time

- But, there are many other cases …
  - E.g., *US Open 2010* (in June vs. Sept)
  - E.g., *World Cup Results* (in 2010 vs. 2006)
- Ongoing evaluation
  - Collecting explicit relevance judgments, query frequency, interaction data and page content over time
  - Developing temporal IR models, temporal snippets

Relevance over Time

- Query: *march madness* [Mar 15 – Apr 4, 2010]
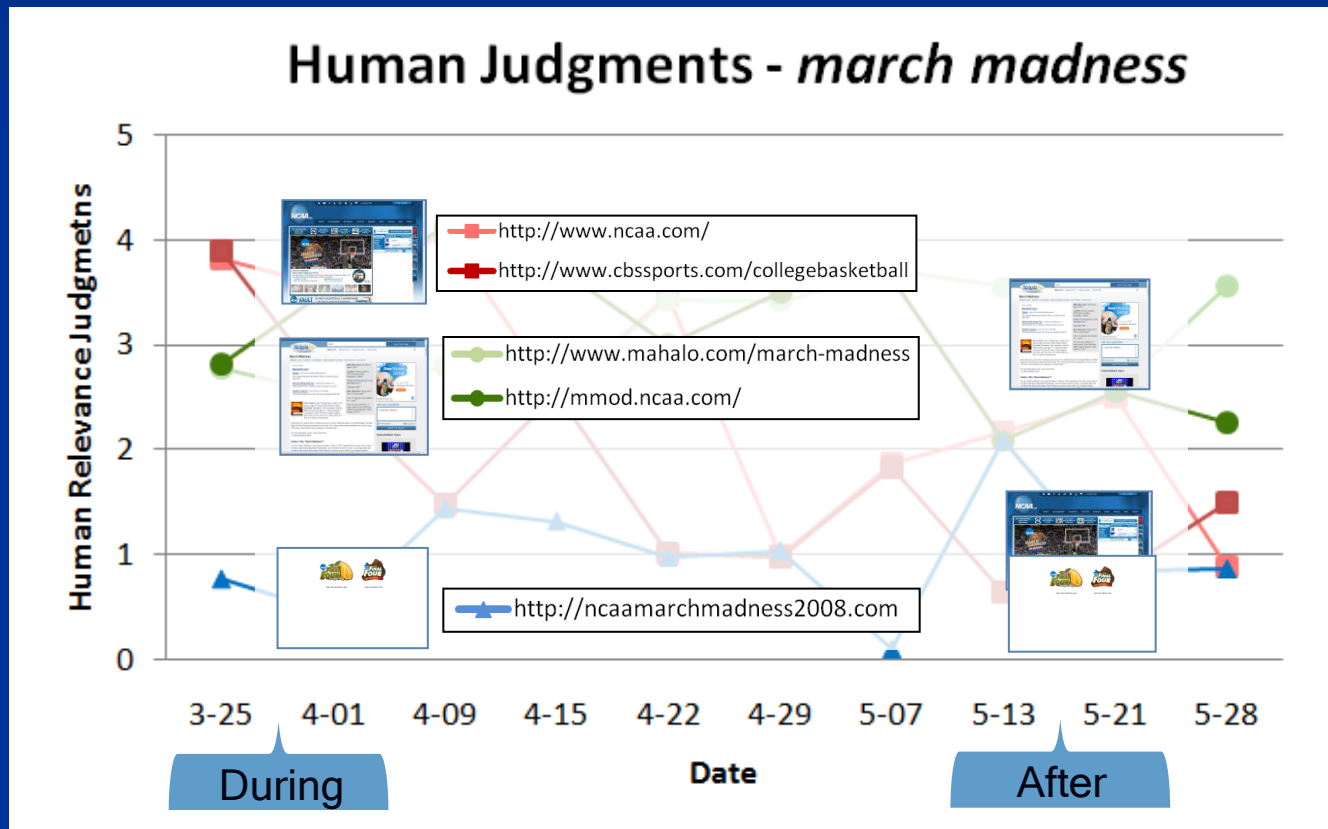
Other Examples of Dynamics and Information Systems

- Query dynamics
  - Kulkarni et al. (2011); Jones & Diaz (2004); Diaz (2009); Kotov et al. (2010)
- Document dynamics, for crawling and indexing
  - Adar et al. (2009); Cho & Garcia-Molina (2000); Fetterly et al. (2003)
- Temporal retrieval models
  - Elsas & Dumais (2010); Liu & Croft (2004); Efron (2010); Aji 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

Web content changes: page-level, term-level

Temporal IR: Leverages change for improved IR
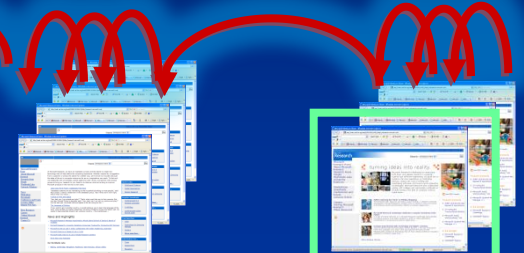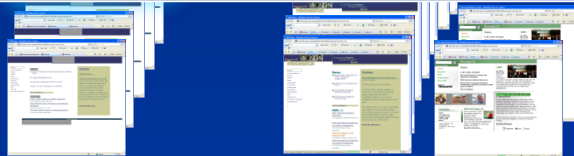
Relating revisitation and change allows us to
– Identify pages for which change is important
– Identify interesting components within a page

2006　2007　2008　2009

2006　2007　2008　2009

People revisit and re-find Web content
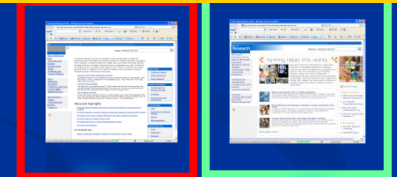
Diff-IE: Supports (and influences) interaction and understanding

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 and IR
  - Others include: location, individuals, tasks …

Think Outside the (Search) Boxes
Search (Re)Search

User Context

Task/Use Context

Document Context

Query Words

Ranked List

iConference - Feb 9, 2011

Feb 2011   Feb 2005   Feb 2000

Diff-IE … try it!
http://research.microsoft.com/en-us/projects/diffie/default.aspx