Temporal Dynamics and Information Retrieval

Susan Dumais

Microsoft Research

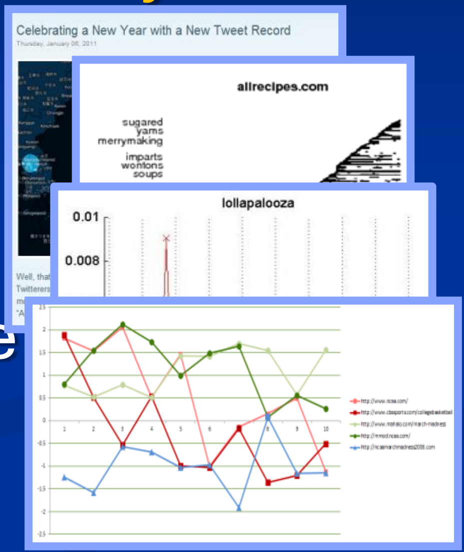
http://research.microsoft.com/~sdumais

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 <> [in 2008 vs. 2004 vs. 2000]



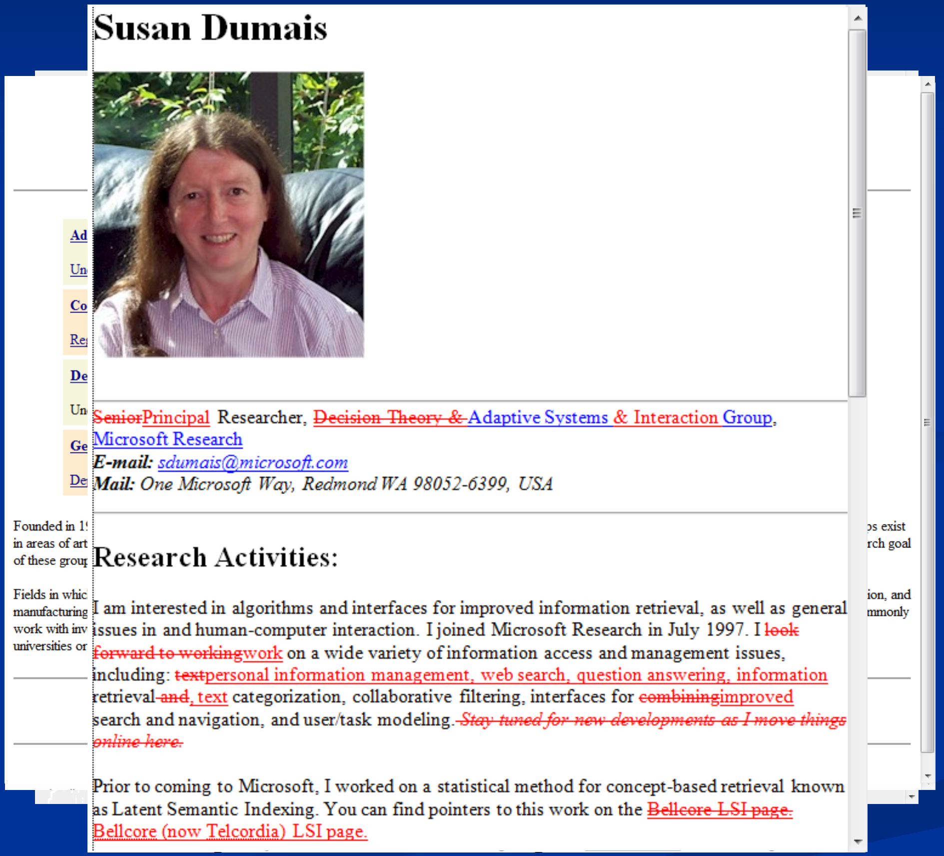
- Change is pervasive in digital information systems

  … yet, most retrieval systems ignore it !

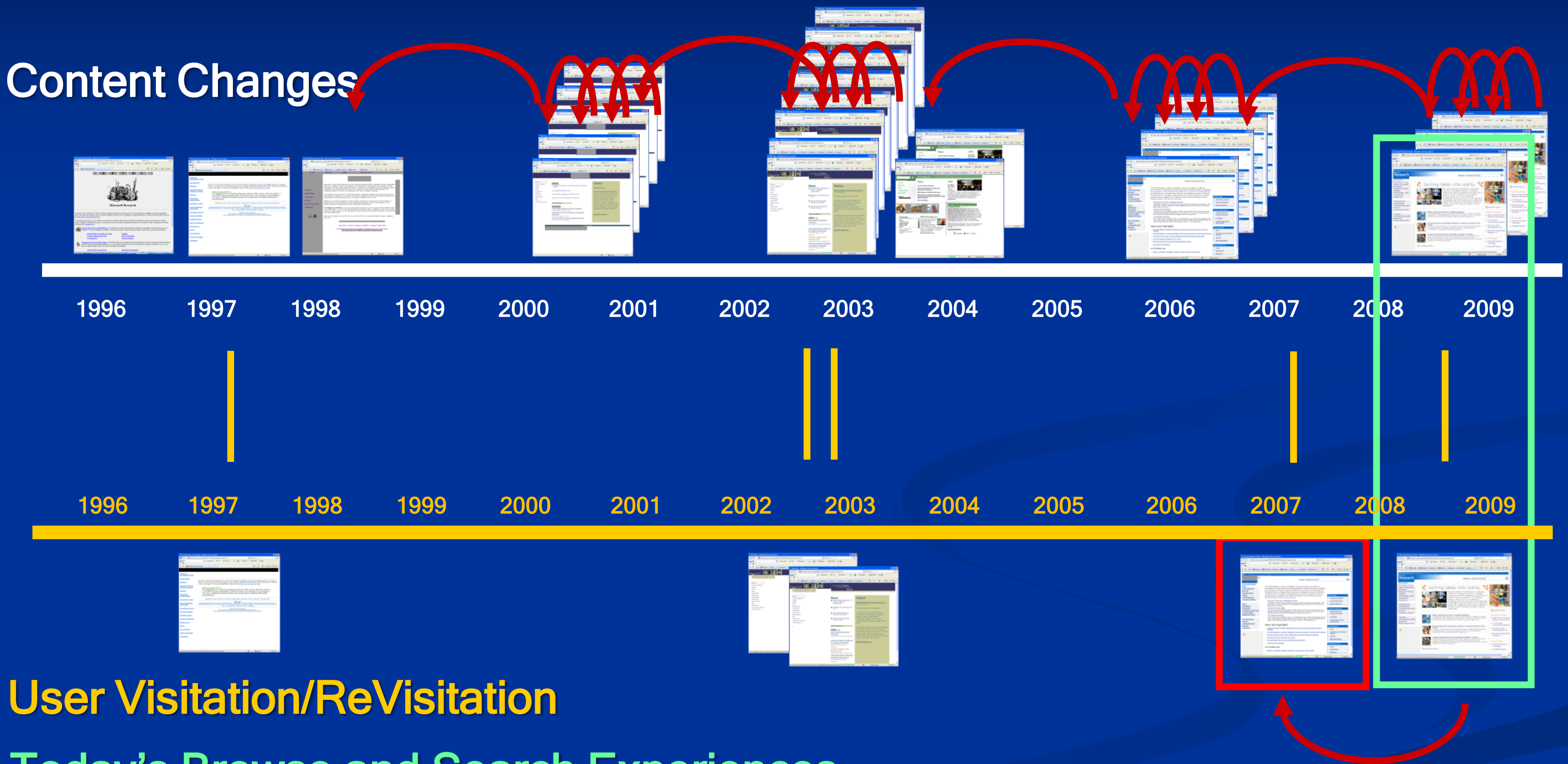
Digital Dynamics Easy to Capture

- Easy to capture

- But … few tools support dynamics



Web Dynamics



User Visitation/ReVisitation

Today's Browse and Search Experiences

But, ignores …

Overview

- Change on the Web
  - <u>Content</u> changes over time
  - <u>User interaction</u> 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

[Adar et al., WSDM 2009]

Characterizing Web Change

- 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

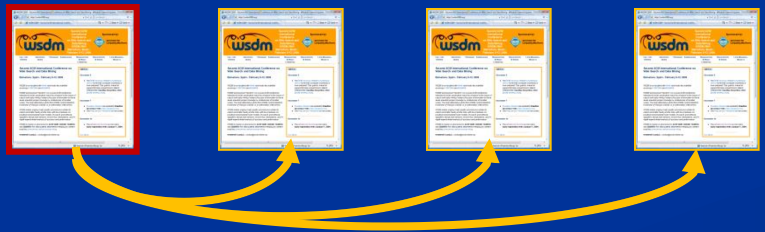
- Summary metrics
  - Number of changes
  - Amount of change
  - Time between changes
- Change curves
  - Fixed starting point
  - Measure similarity over different time intervals
- Within-page changes

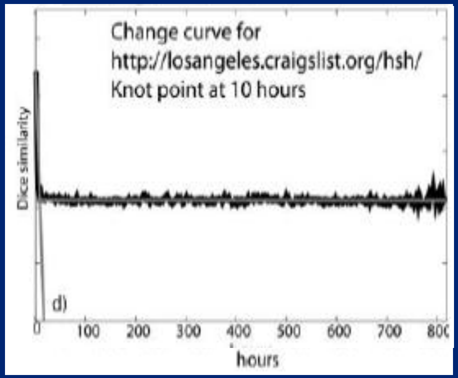
Measuring Web Page Change

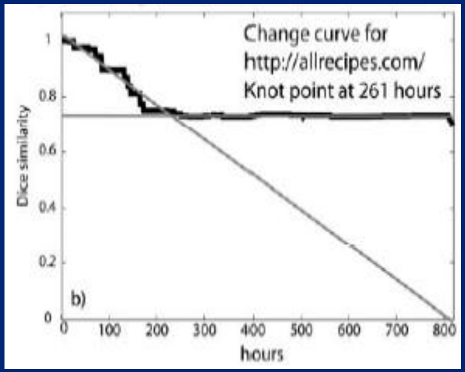
- Summary metrics
  - Number of changes
    - 33% of Web pages change
    - 66% of visited Web pages change
      - 63% of these change every hr.
  - Amount of change
  - Time between changes
    - Avg. Dice coeff. = 0.80
    - Avg. time bet. change = 123 hrs.
    - .edu and .gov pages change infrequently, 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



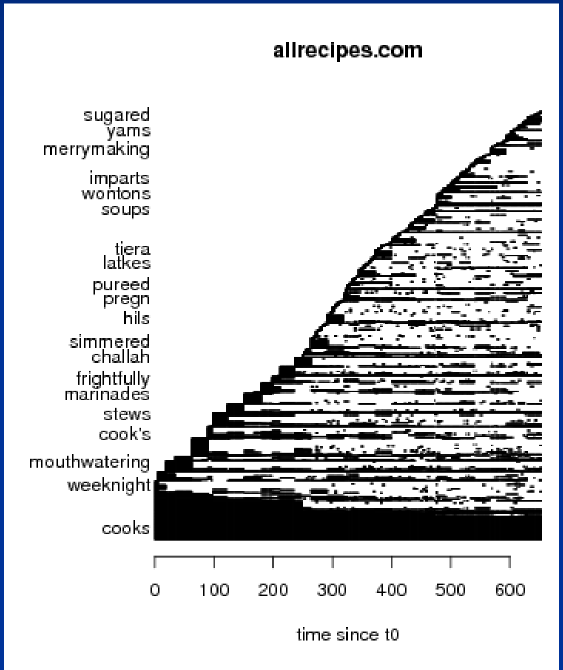


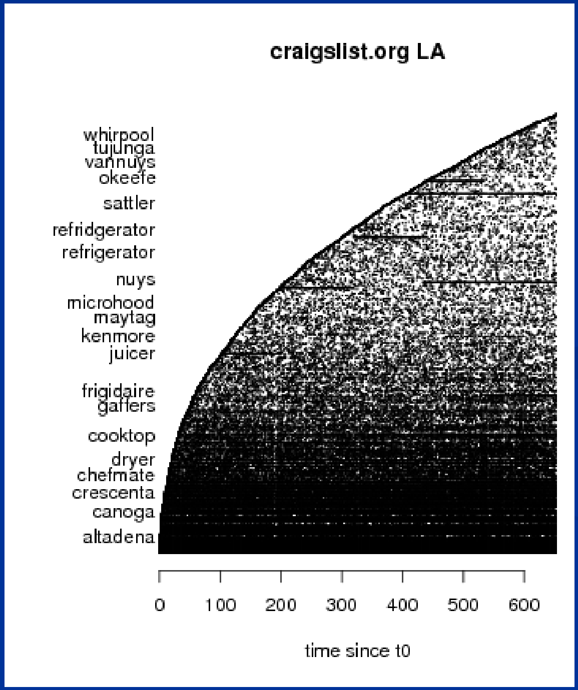


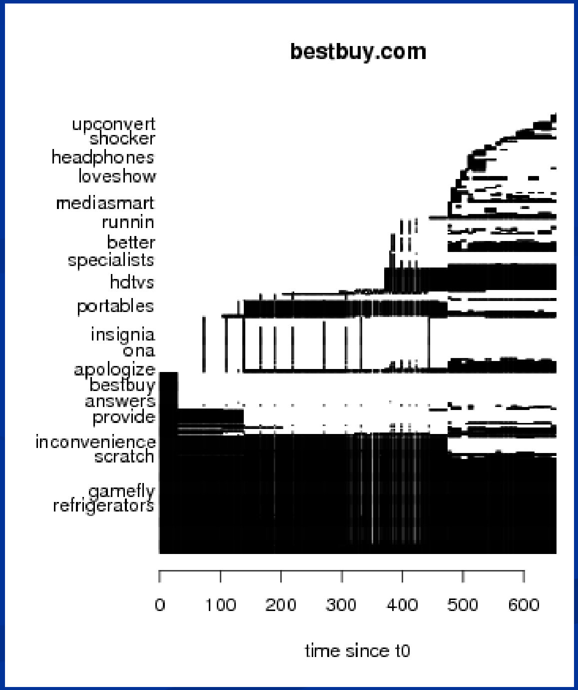
Measuring Within-Page Change

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

Example Term Longevity Graphs





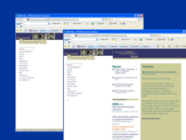


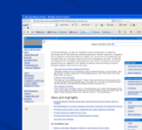
[Adar et al., CHI 2009]

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

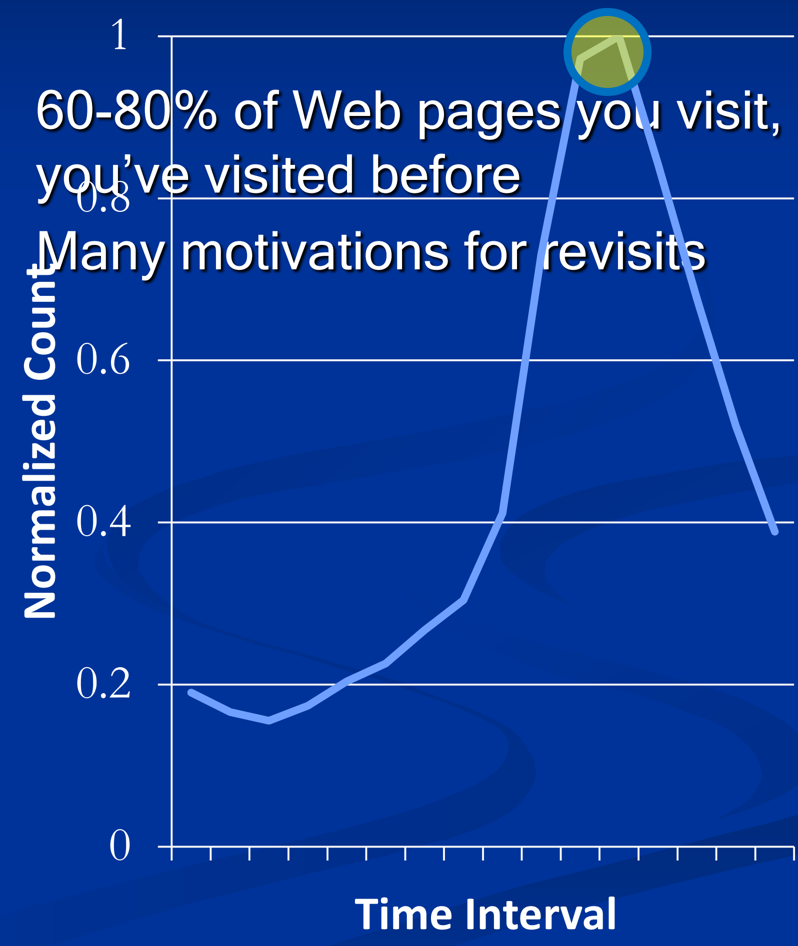
<u>What</u> was the last Web page you visited?

<u>Why</u> did you visit (re-visit) the page?

Measuring Revisitation

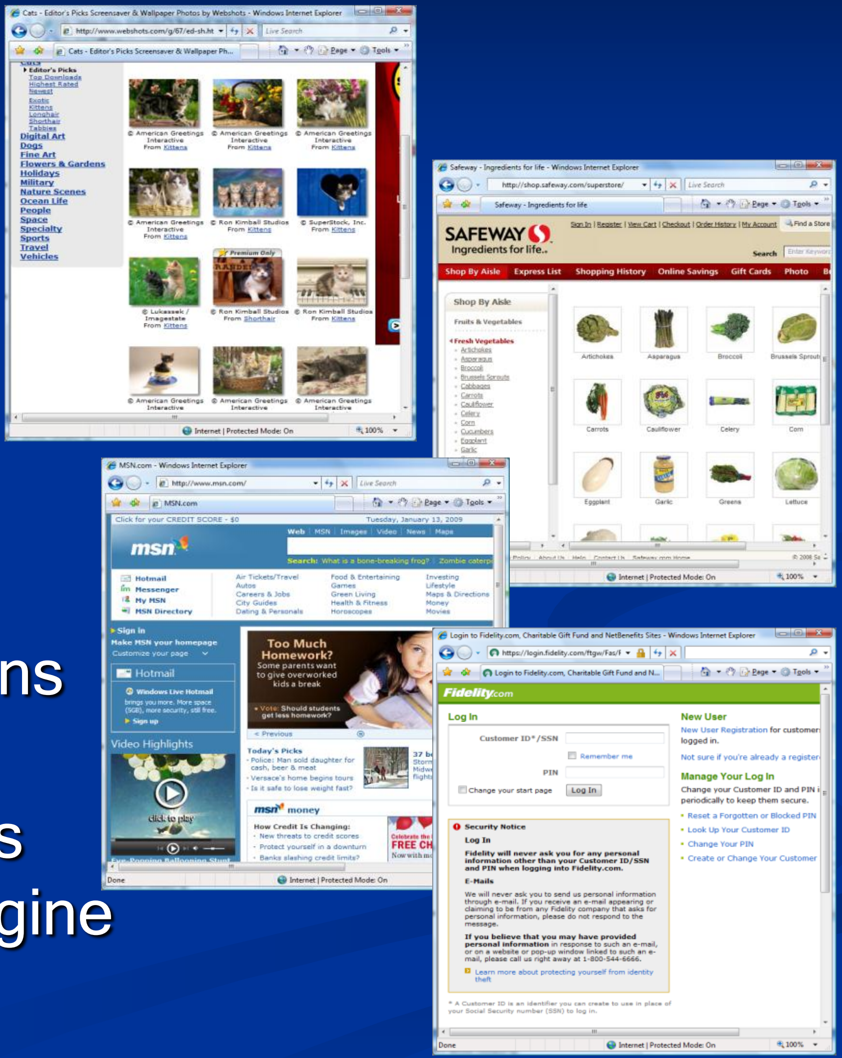
- Summary metrics
  - Unique visitors
  - Visits/user
  - Time between visits
- Revisitation curves
  - Histogram of revisit intervals
  - Normalized

- 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

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

[Teevan et al., SIGIR 2007]
[Tyler et al., WSDM 2010]
[Teevan et al., WSDM 2011]

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%		

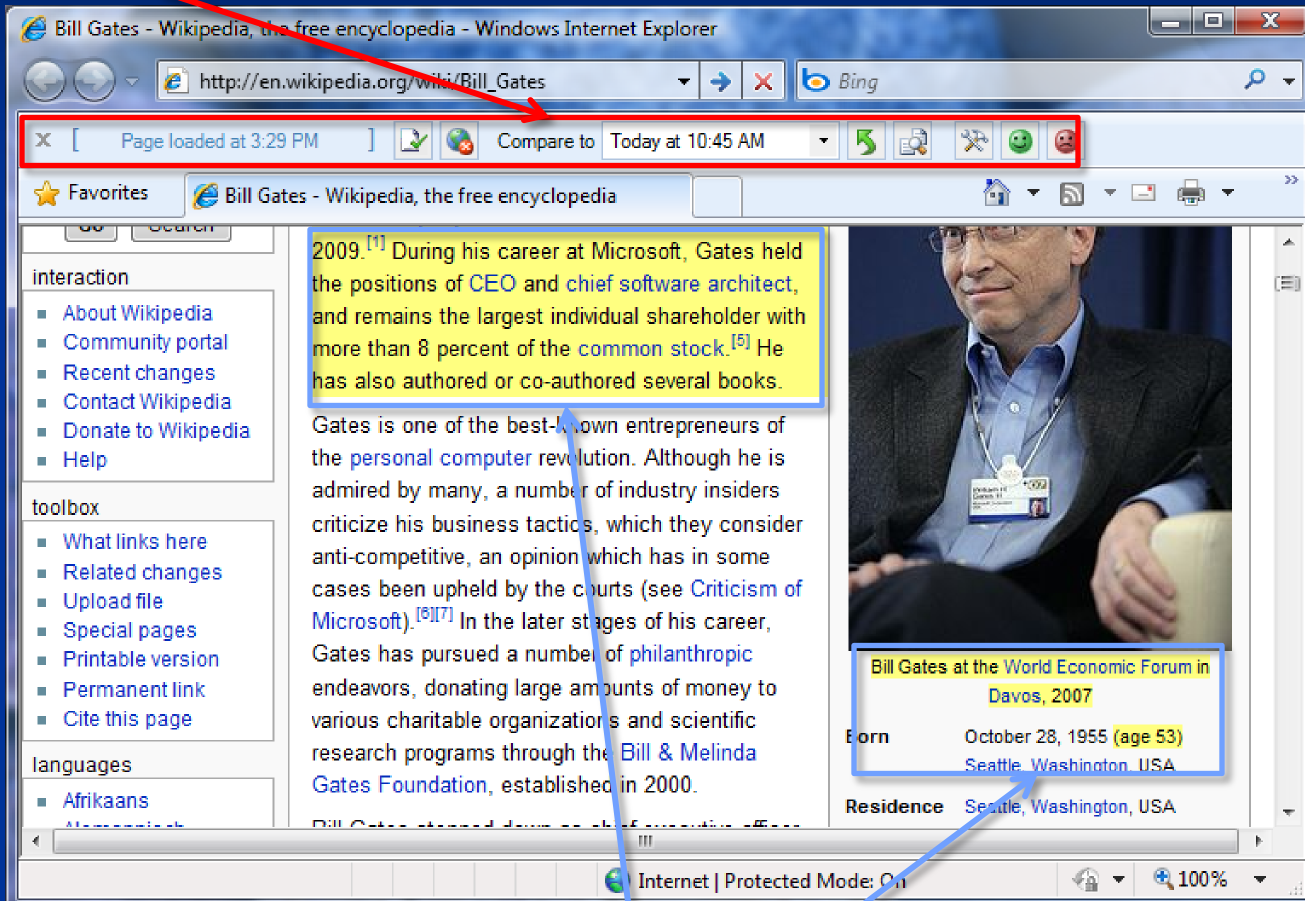
Building Support for Web Dynamics

Diff IE

[Teevan et al., UIST 2009]

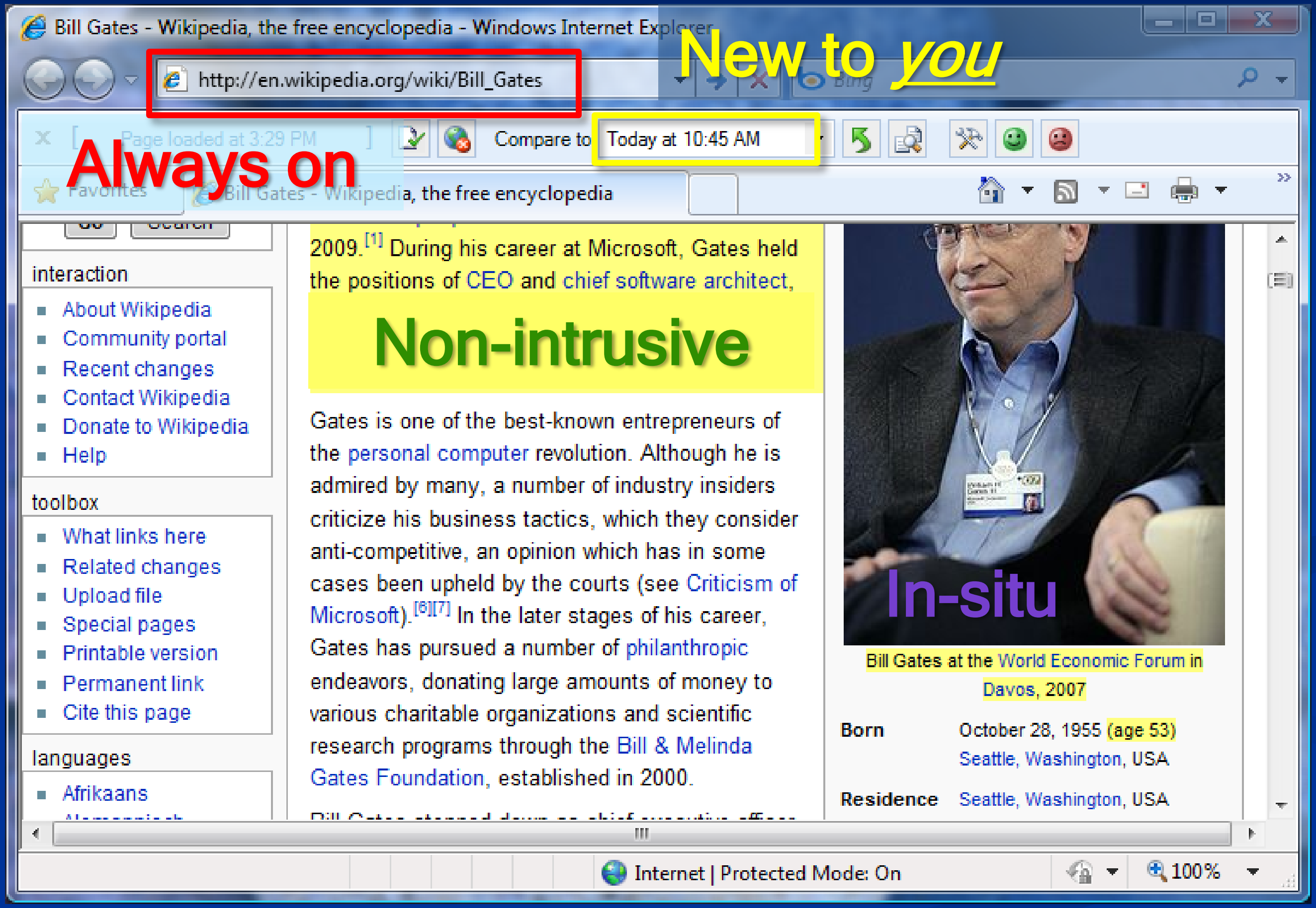
Diff-IE

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

The New York Times

Thursday, March 8, 2012 Last Update: 1:38 PM ET

Search

Follow Us | Subscribe to Home Delivery | Personalize Your Weather

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

TimesCast

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

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

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

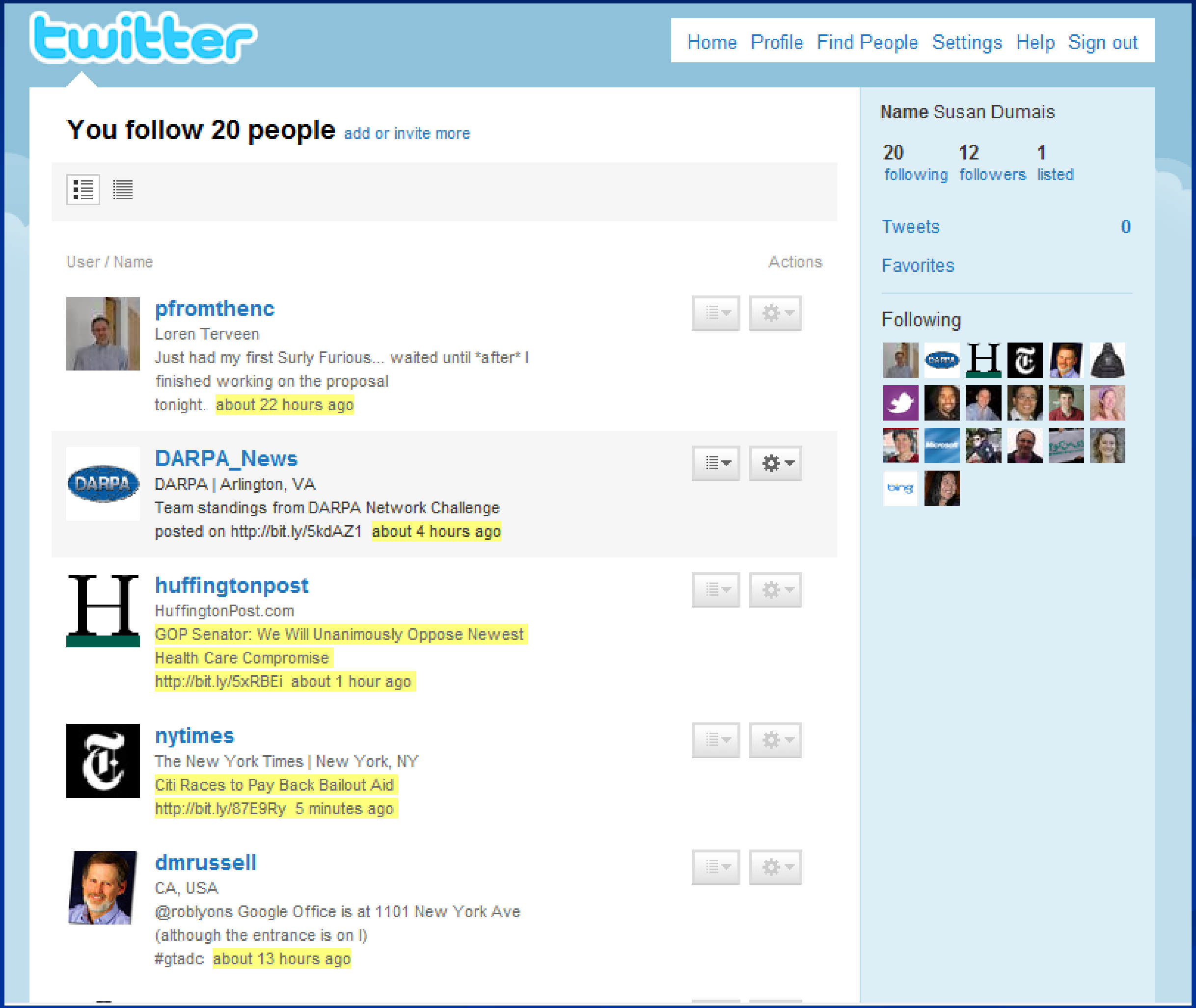
S.&P. 500	Dow	Nasdaq
1,364.91	12,907.22	2,967.92
+12.28	+69.89	+32.23
+0.91%	+0.54%	+1.10%

GET QUOTES My Portfolios »

Stock, ETFs, Funds Go

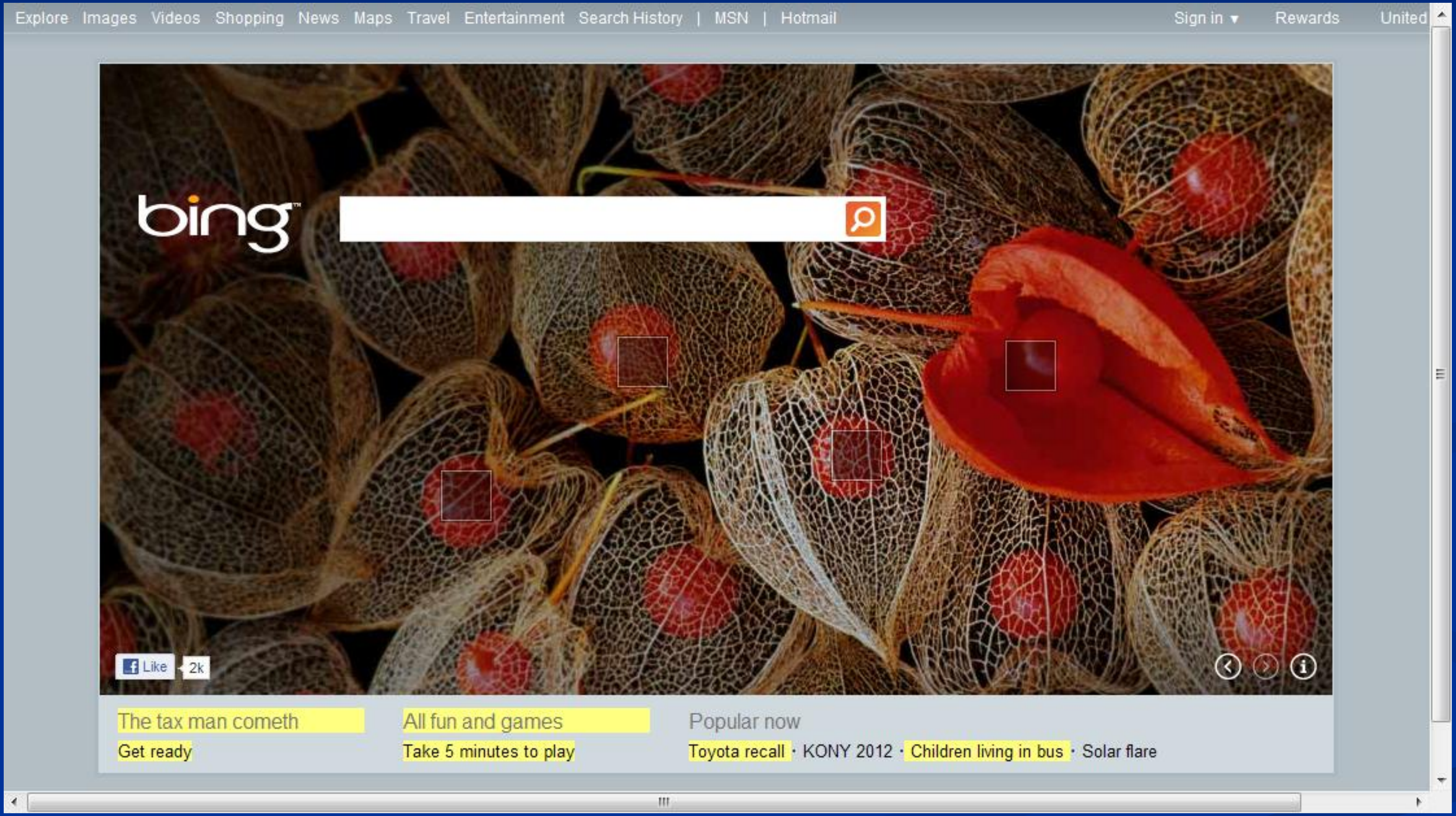
THEATER »

Monitor



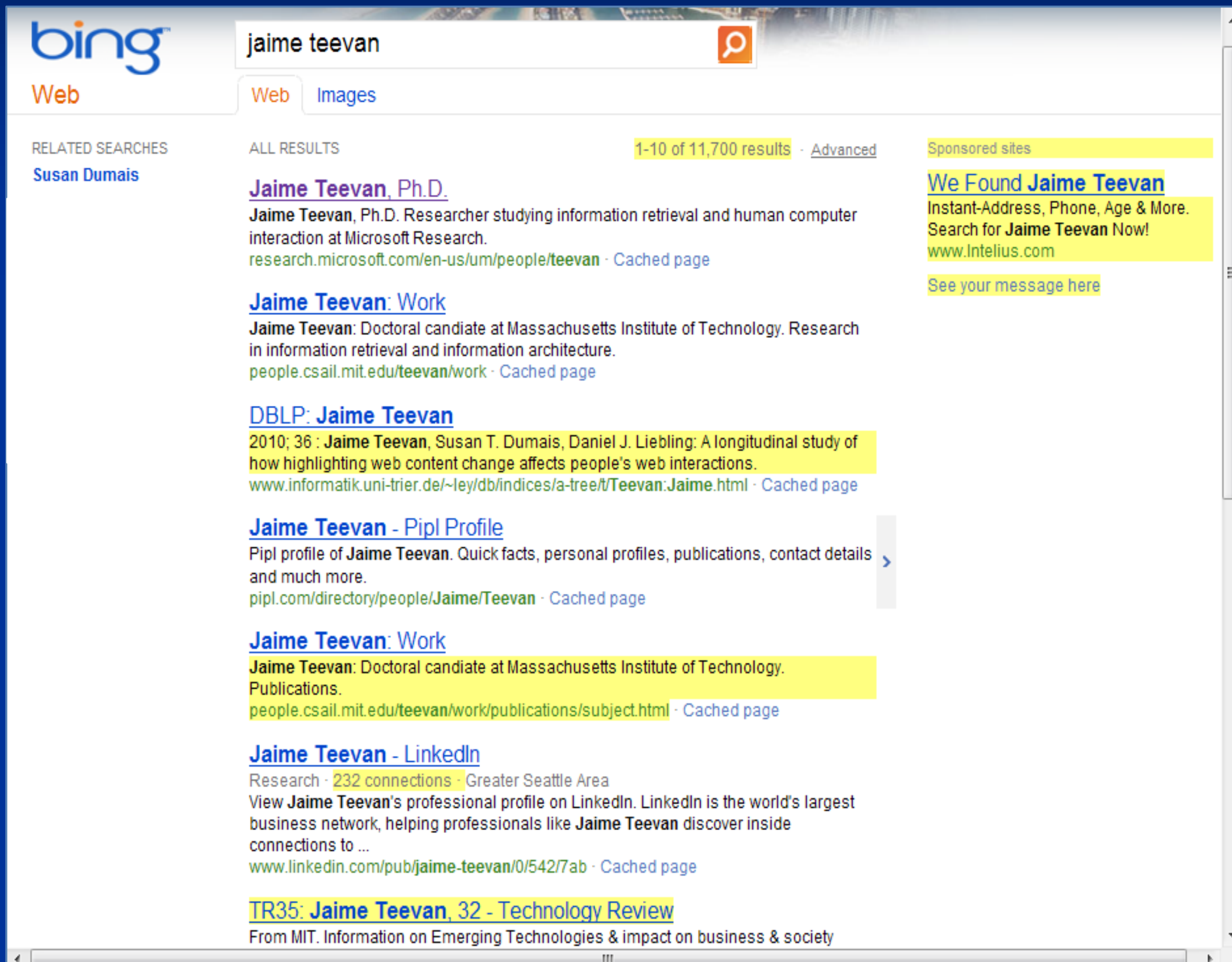
Home Profile Find People Settings Help Sign out
You follow 20 people add or invite more
User / Name
Actions
pfromthenc
Loren Terveen
Just had my first Surly Furious... waited until *after* I finished working on the proposal tonight. about 22 hours ago
DARPA_News
DARPA | Arlington, VA
Team standings from 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
Name Susan Dumais
20 following
12 followers
1 listed
Tweets 0
Favorites
Following

Serendipitous Encounters



Unexpected Important Content

Understand Page Dynamics



Expected

Expected
New Content

Monitor

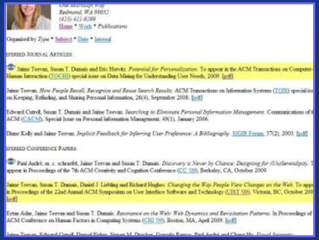
Unexpected

Unexpected
Important Content

Edit

Attend to Activity

Understand
Page Dynamics



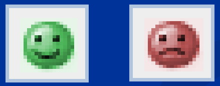
Serendipitous
Encounter

Unexpected
Unimportant Content

Studying Diff-IE

- Internal study of Diff-IE
- Logging
  - URLs visited
  - Amount of change when revsited

- Feedback buttons



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

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

	Always	Often	Sometimes	Rarely	Never
News pages					
Message boards, forums, newsgroups					
Company homepages					
Personal homepages of people you know					
Pages with product information					
Reference pages (dictionaries, yellow pages, maps)					
Wikipedia pages					
Blogs you read					
Search engine results					

- 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
  - First week: 39.4% of visits are revisits
  - Last week: 45.0% of visits are revisits

14%

- Why are people revisiting more with DIFF-IE?

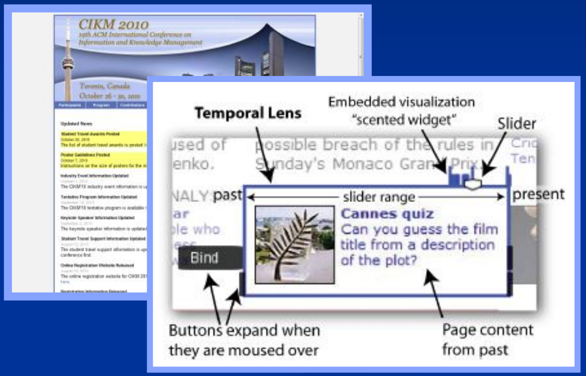
Revisited Pages Change More

- Perception of change increases
  - What proportion of pages change regularly? 8%
  - How often do you notice unexpected change? 17%
- Amount of change seen increases
  - First week: 21.5% revisits changed, by 6.2%
  - Last week: 32.4% revisits changed, by 9.5%

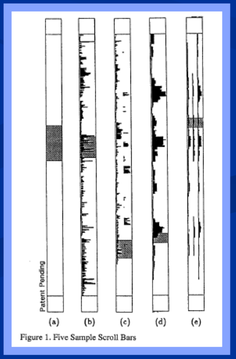
  51+%
- Diff-IE is driving visits to changed pages
  - It supports people in understanding change

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, “likes”, etc.
  - Implicit interest via interaction patterns
    - Edit wear and read wear (Hill et al., 1992)



Overview

- Change on the Web
  - <u>Content</u> changes over time
  - <u>User interaction</u> 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

[Elsas et al., WSDM 2010]

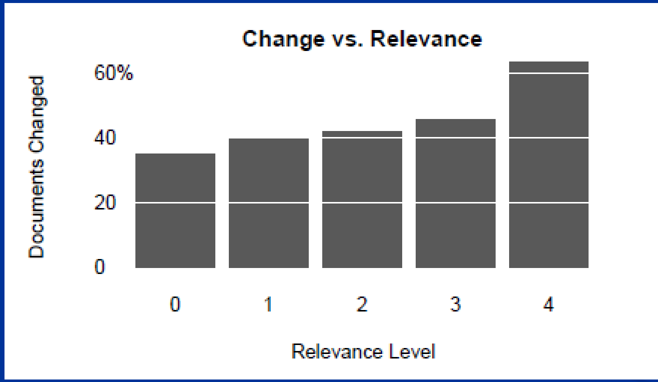
Temporal Retrieval Models 1 (content-based)

- 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

$$P(D|Q) = \underbrace{\boxed{P(D)}}_{\text{Change prior}} \cdot \underbrace{\boxed{P(Q|D)}}_{\text{Term longevity}}$$

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

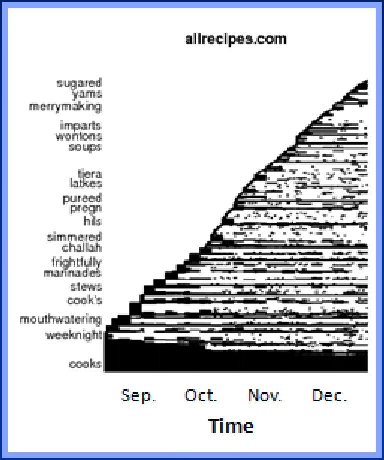


- Use change rate as a document prior (vs. priors based on link structure like Page Rank)
  - Shingle prints to measure change

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

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

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

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

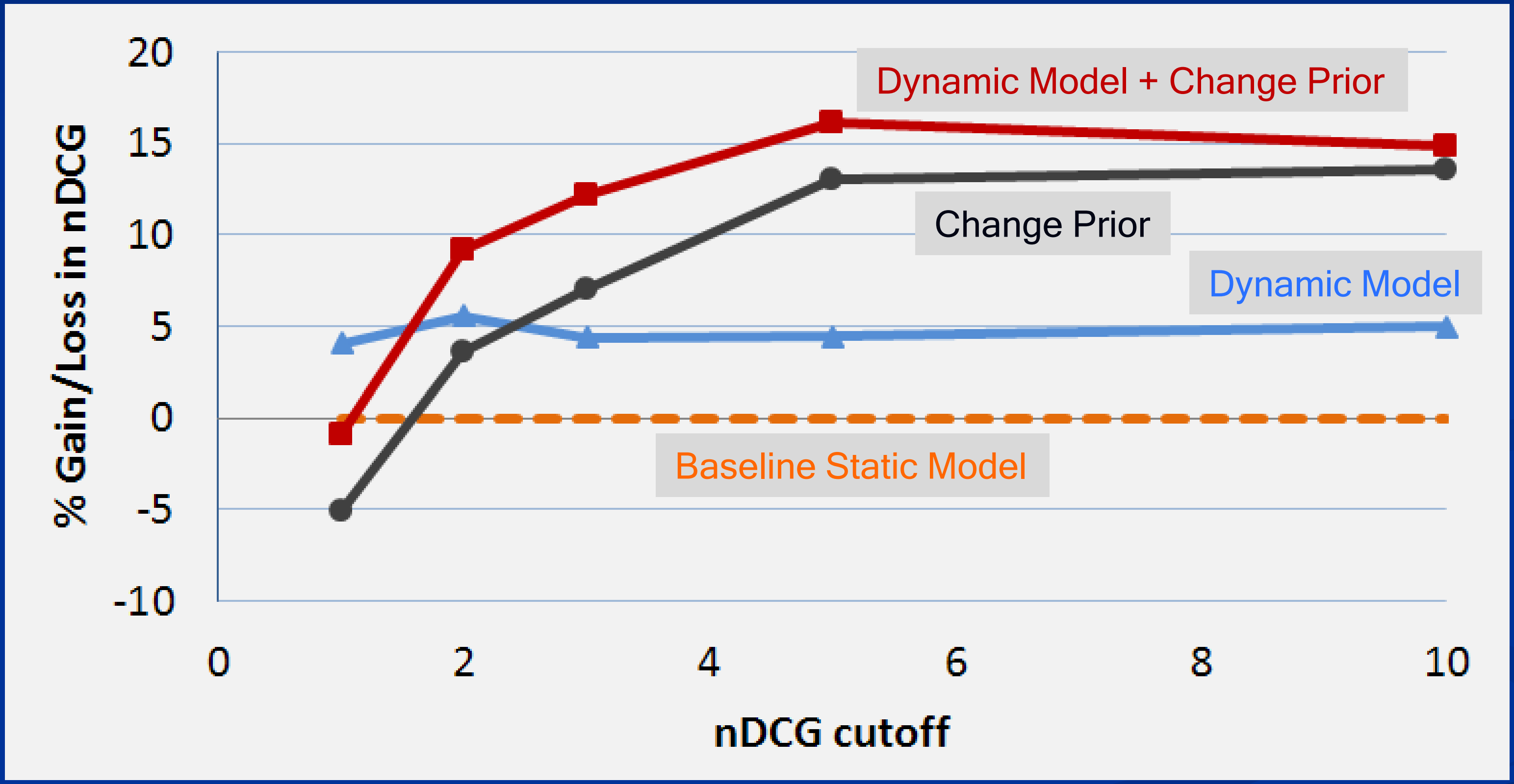
Term longevity

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



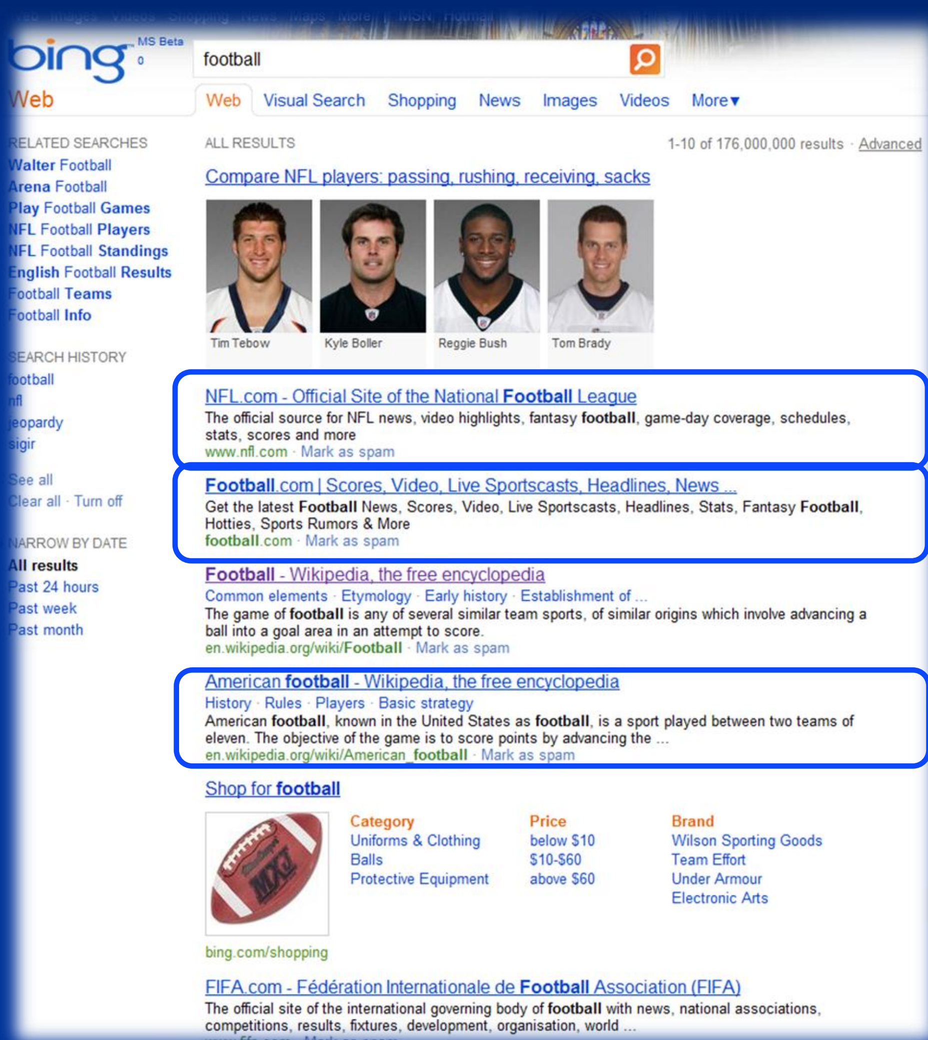
[Radinsky et al., submitted]

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 <u>over time</u>

Relevance over Time

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



Feb - Aug

Sep - Jan

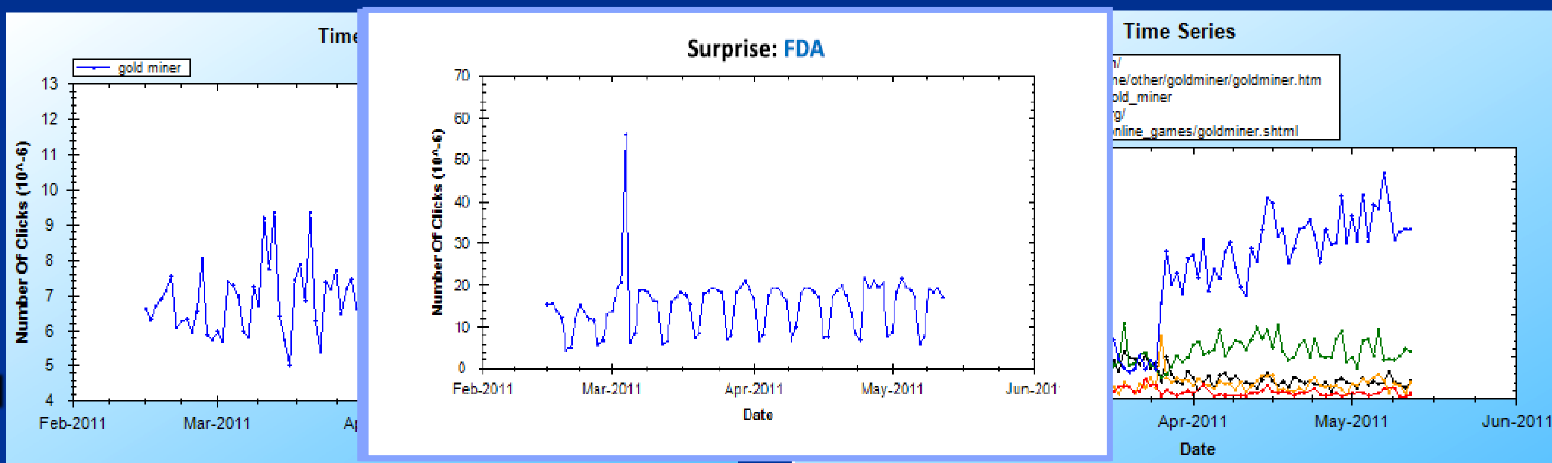
Feb - Aug

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

Experimental Setup

- Data
  - Queries and clicked URLs, over 4 months



ation)

-
  - Actual user search behavior over time (implicit measure)
- Model temporal dynamics of behavior
- Use model to improved ranking

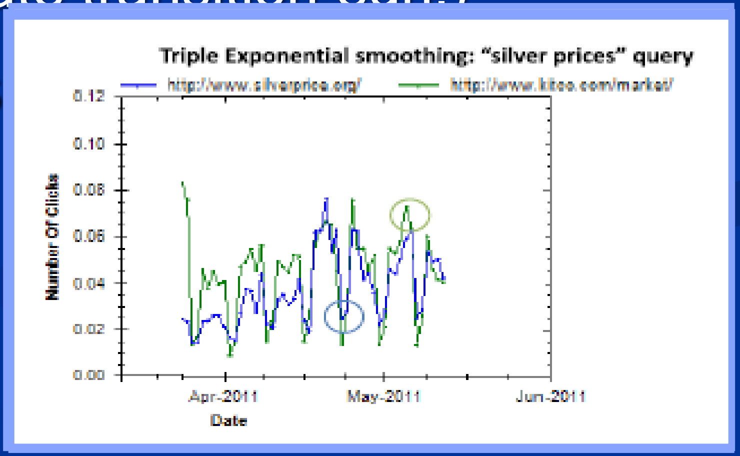
Time Series Modeling

- 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 <u>only</u> feature for ranking
  - Temporal features (+other features) as <u>input to learned ranker</u>
- 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	0.18	0.21	0.22
Periodic	0.91	0.92	0.93
Dynamic	0.28	0.35	0.38
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	**0.31**	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	**0.90**	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 stat
using our temporal models. Statistically significant differences based on a paired t-test when c
performing algorithm ($p < .05$) are shown in bold.

- Best-performing queries

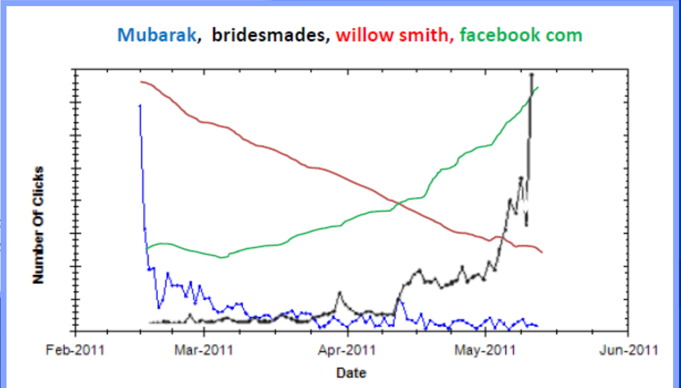
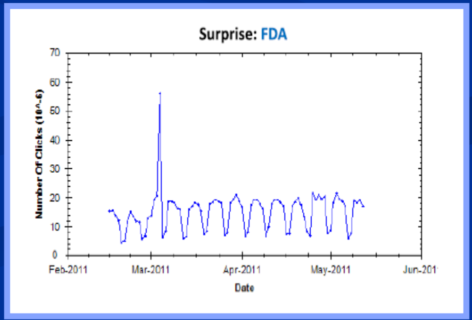


Figure 6: Dominant query shapes for queries where temporal model yielded better rankings than baseline rankers.

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
  - …



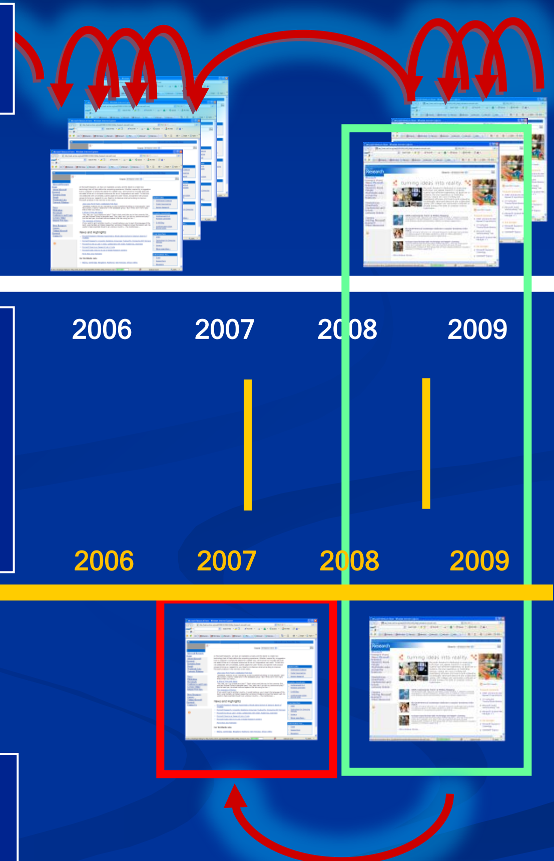
Other Examples of Dynamics and Information Systems

- 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

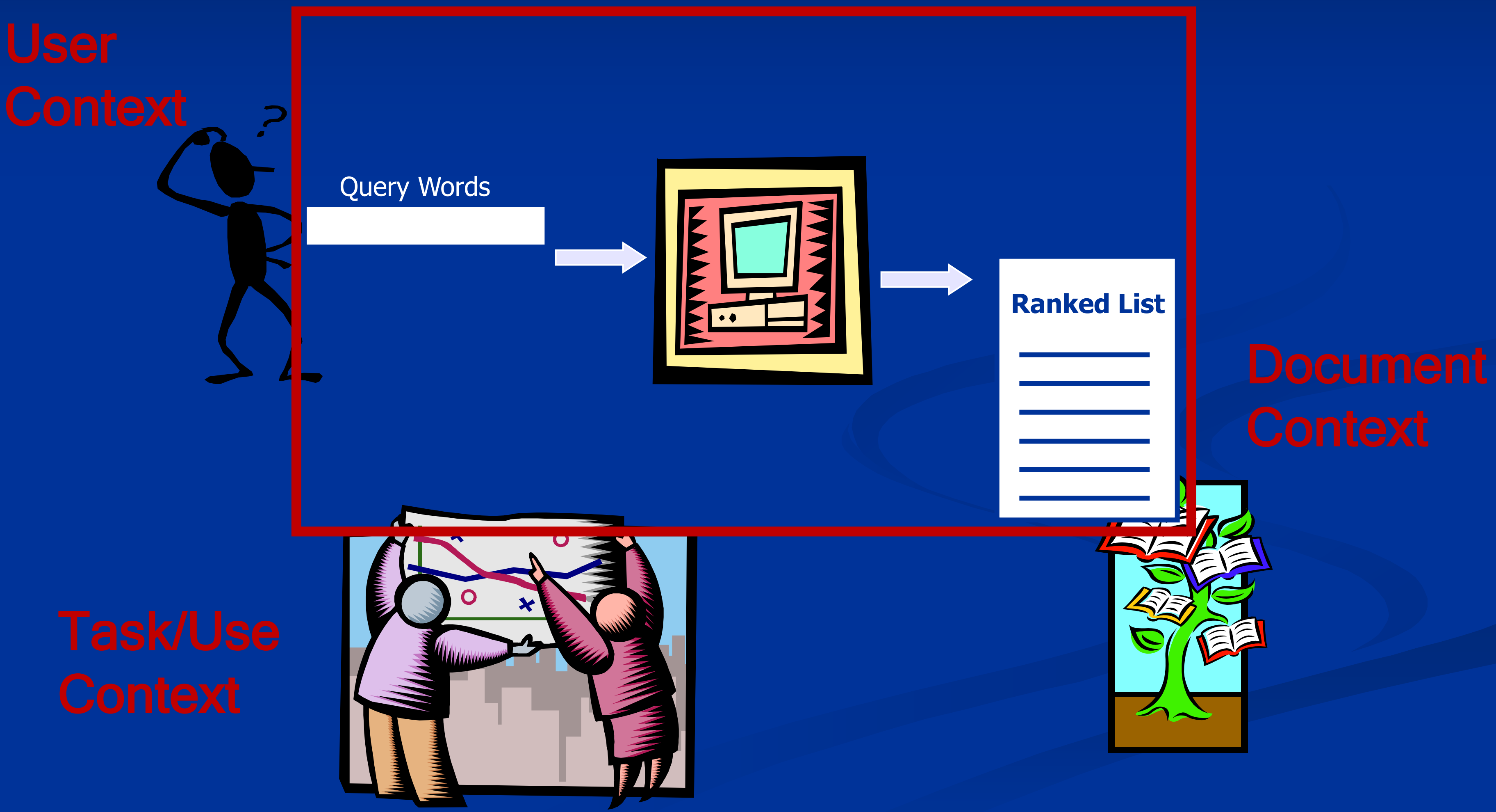
Diff-IE: Supports (and influences) interaction and understanding

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 the (Search) Boxes

Search Research



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