Temporal Dynamics and Information Retrieval

Susan T. Dumais

Microsoft Research

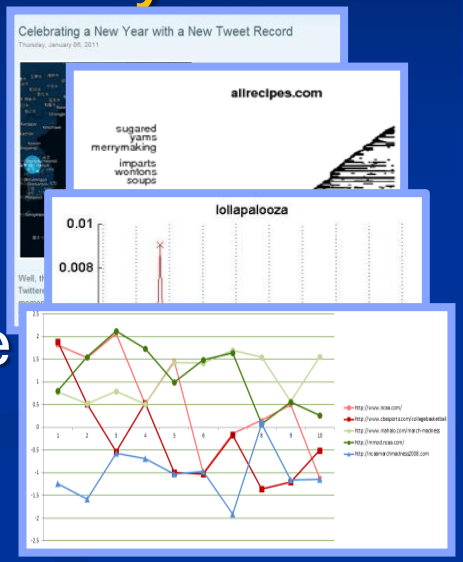
http://research.microsoft.com/~sdumais

In collaboration with:

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

Change is Everywhere in IR

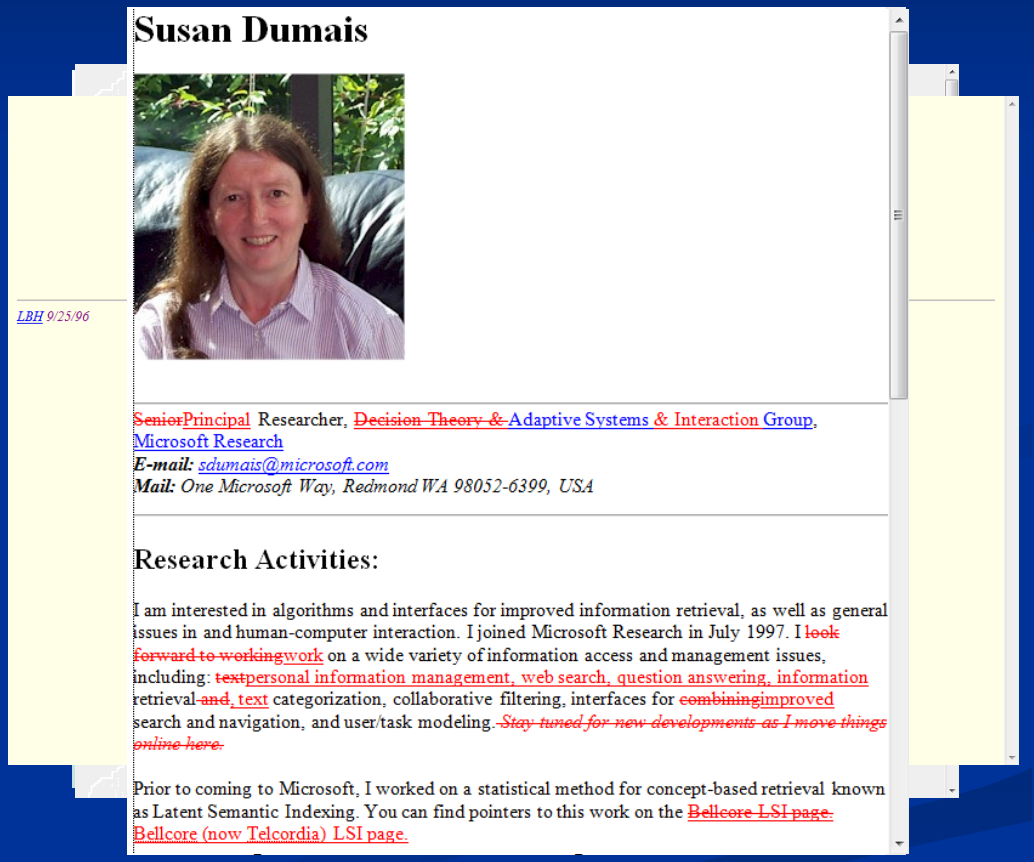
- Change is everywhere in digital information systems
  - New documents appear all the time
  - Document content changes over time
  - Queries and query volume change over time
  - What's relevant to a query changes over time
    - E.g., *U.S. Open 2012* (in June vs. Sept)
  - User interaction changes 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 !

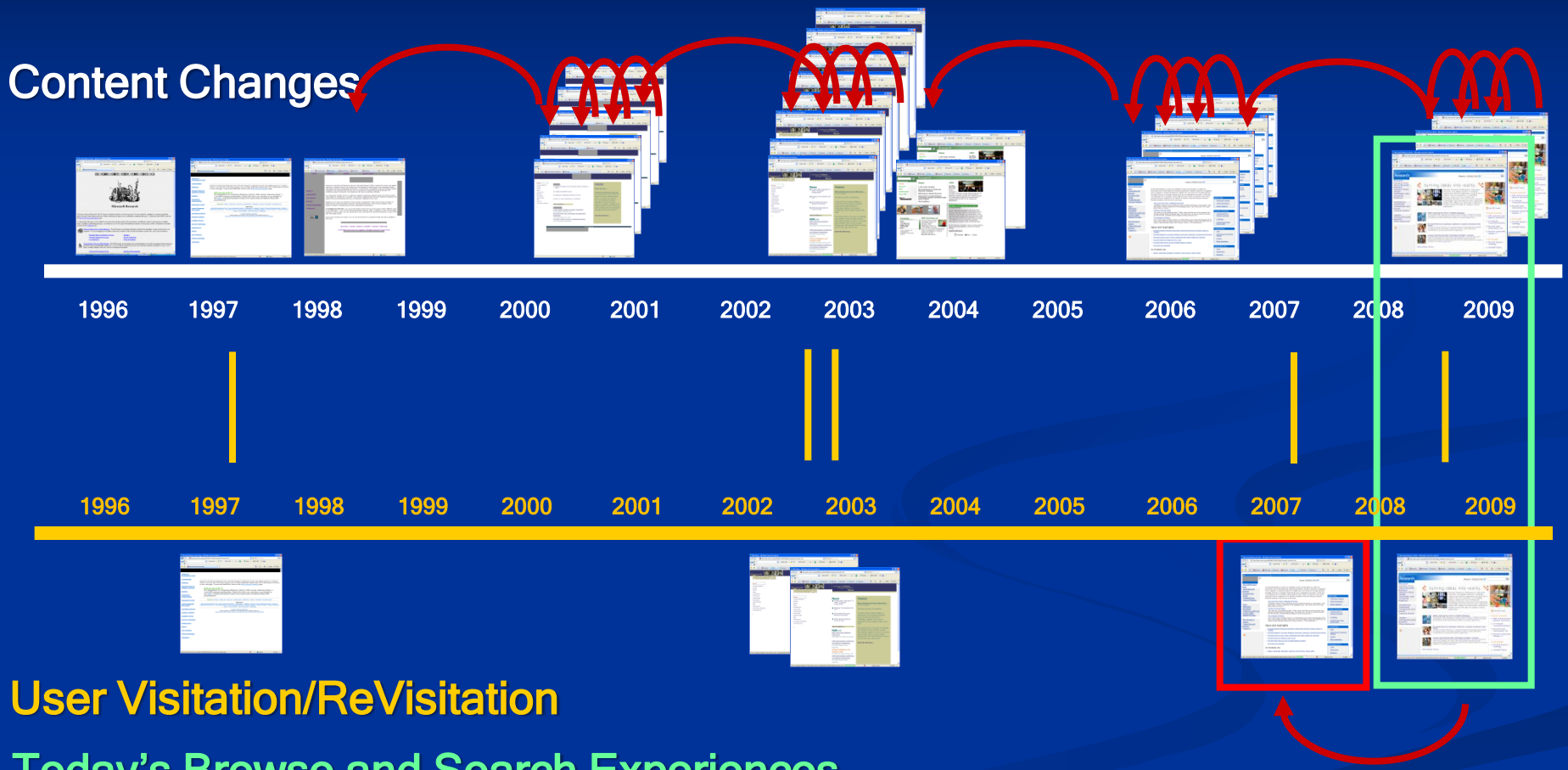
Digital Dynamics Easy to Capture

- Easy to capture



- But … few tools or algorithms support dynamics

Web Dynamics



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

[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

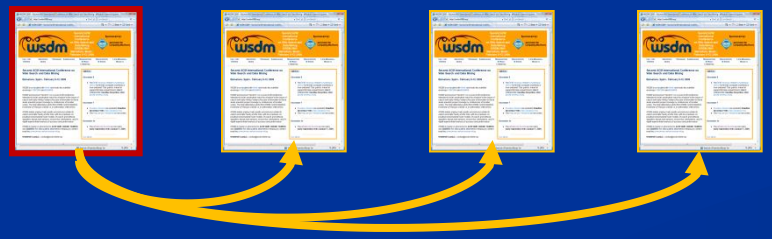
- Summary metrics
  - Number of changes
  - Amount of change
  - Time between changes
- Change curves
  - Fixed starting point
  - Measure similarity over different time intervals
- Within-page changes

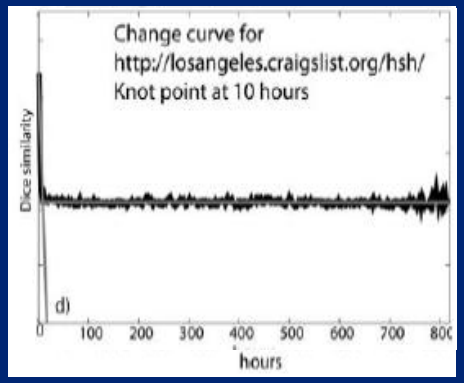
Measuring Web Page Change

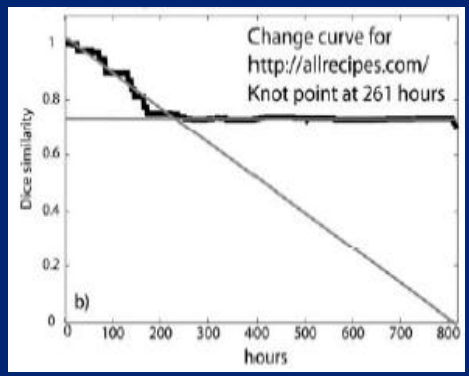
- Summary metrics
  - Number of changes
    - 33% of Web pages change
    - 66% of visited Web pages change
      - 63% of these change every hr.
  - Amount of change
  - Time between changes
    - Avg. Dice coeff. = 0.80
    - Avg. time bet. change = 123 hrs.
    - .edu and .gov pages change infrequently,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



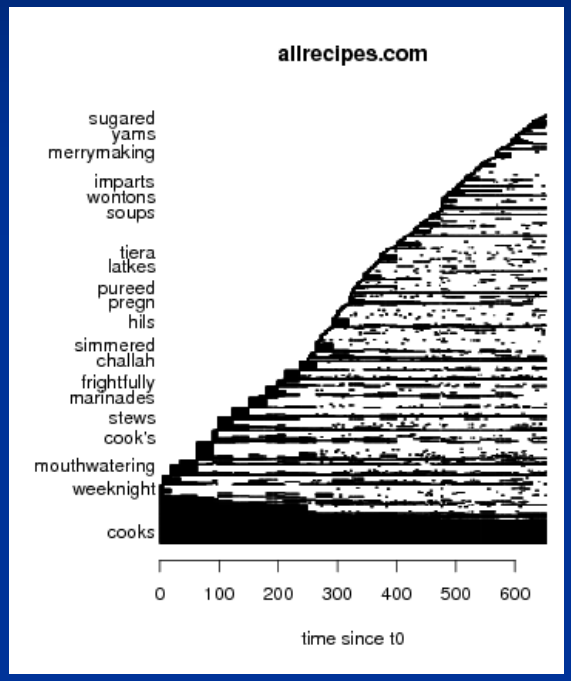


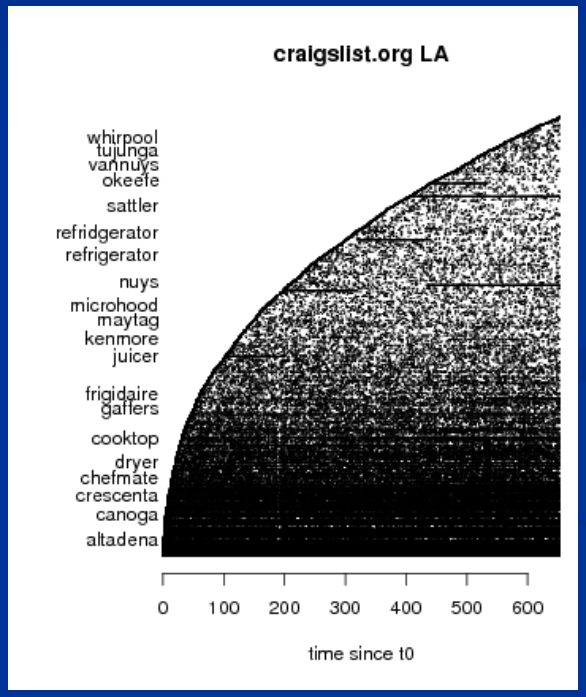


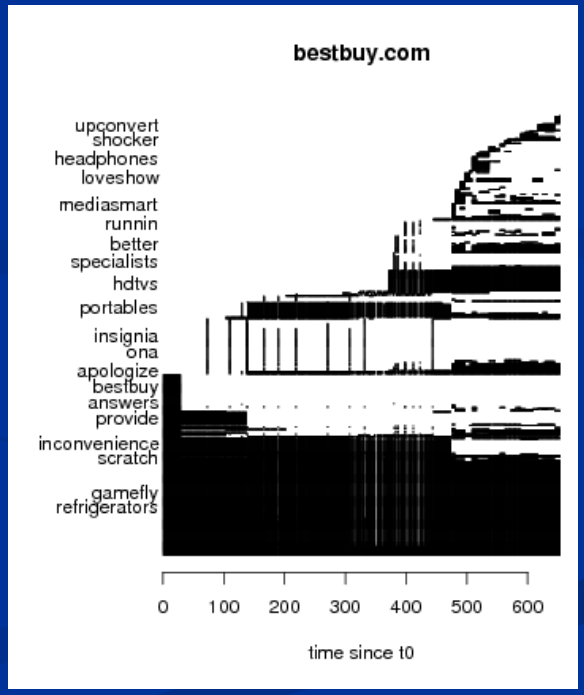
Measuring Within-Page Change

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

Example Term Longevity Graphs



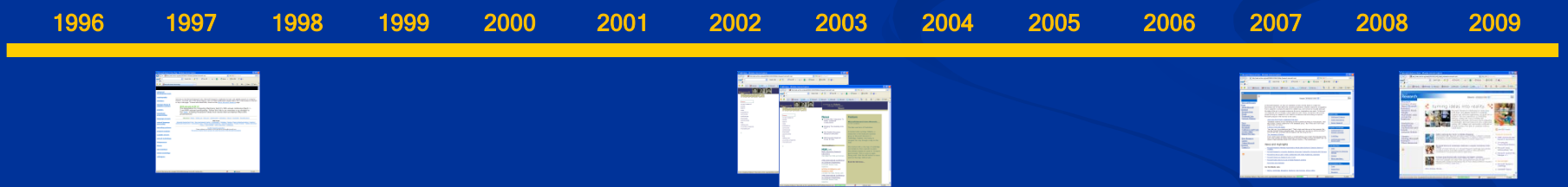




[Adar et al., CHI 2009]

Revisitation on the Web

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



User Visitation/ReVisitation

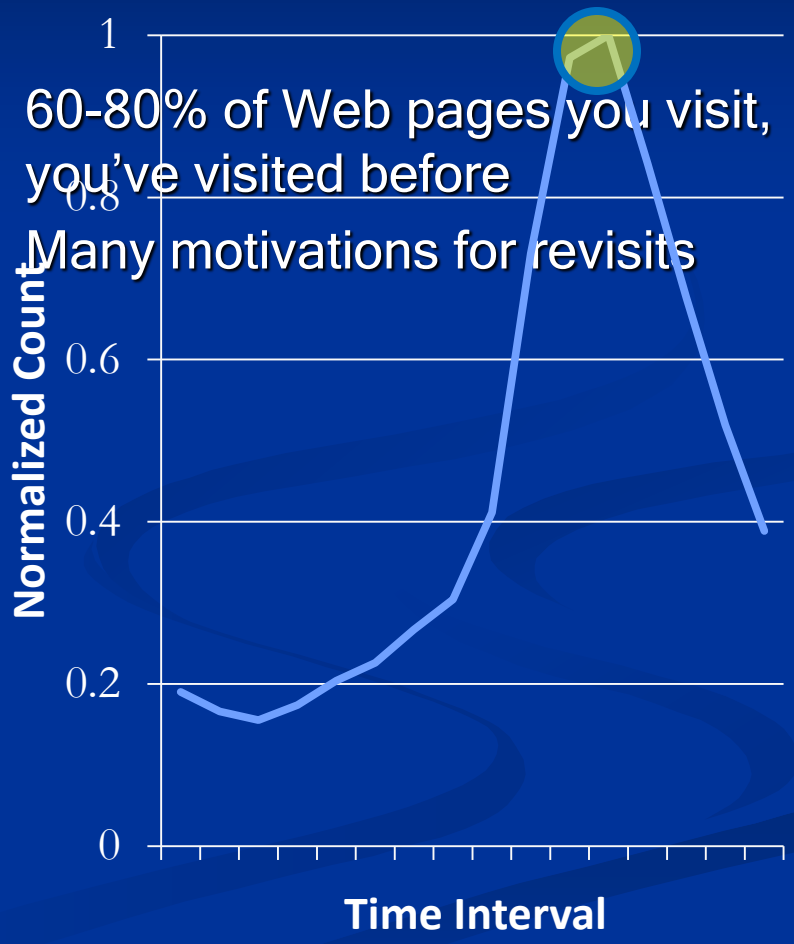
What was the last Web page you visited?

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

Measuring Revisitation

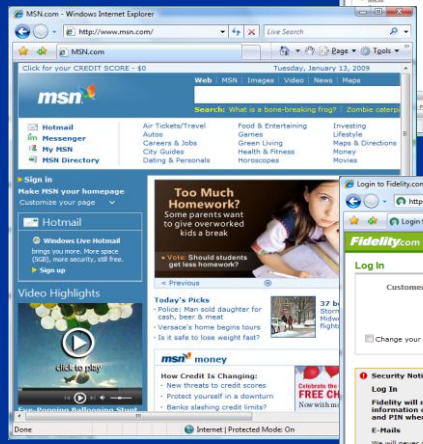
- Summary metrics
  - Unique visitors
  - Visits/user
  - Time between visits
- Revisitation curves
  - Histogram of revisit intervals
  - Normalized

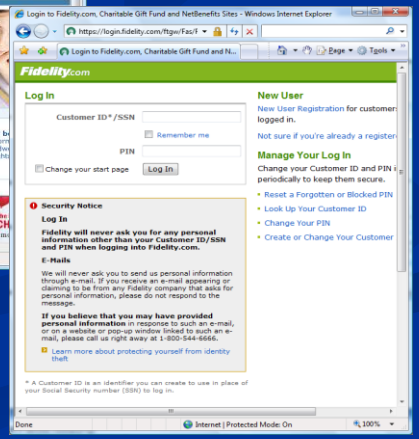
- 60-80% of Web pages you visit, you've visited before
- Many motivations for revisits



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

- 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]

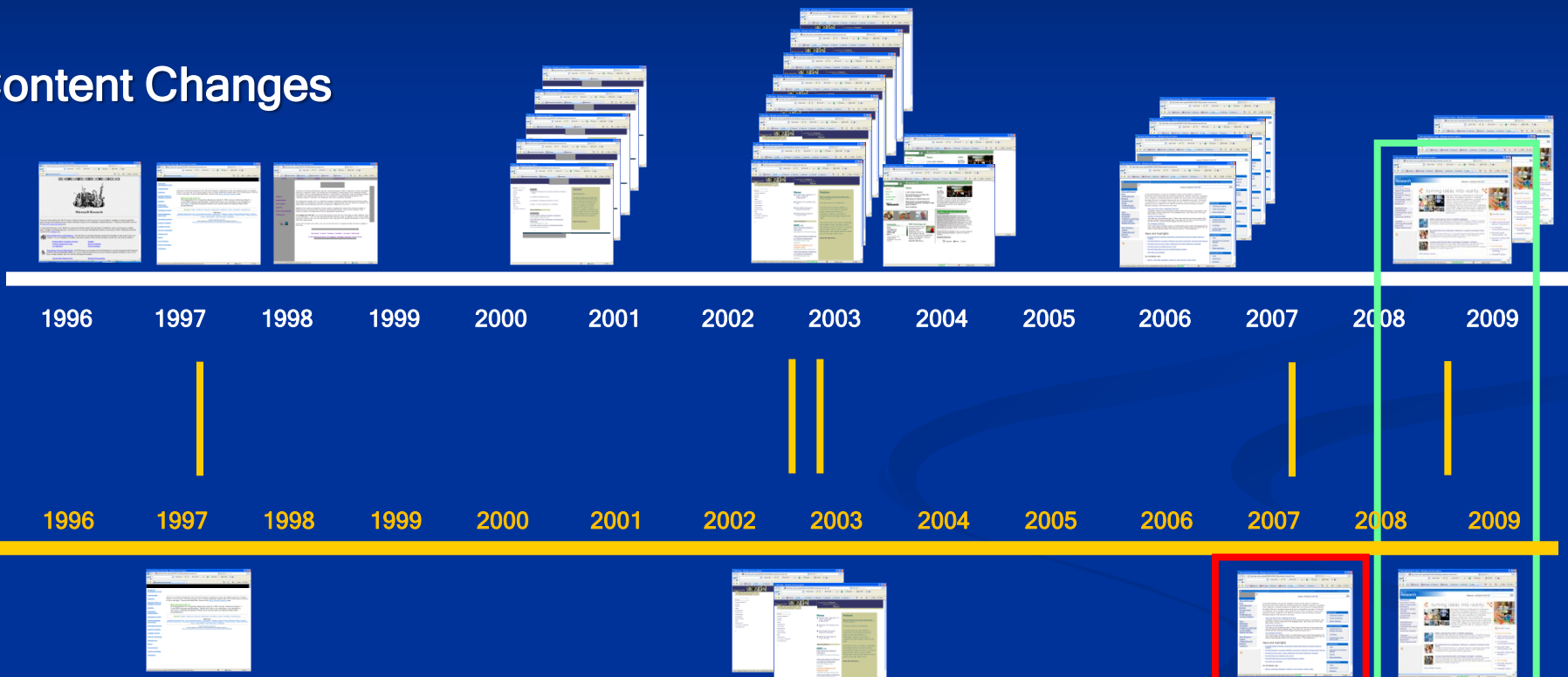
Revisitation and Search (Re-finding)

- 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%)

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

Building Support for Web Dynamics

Content Changes

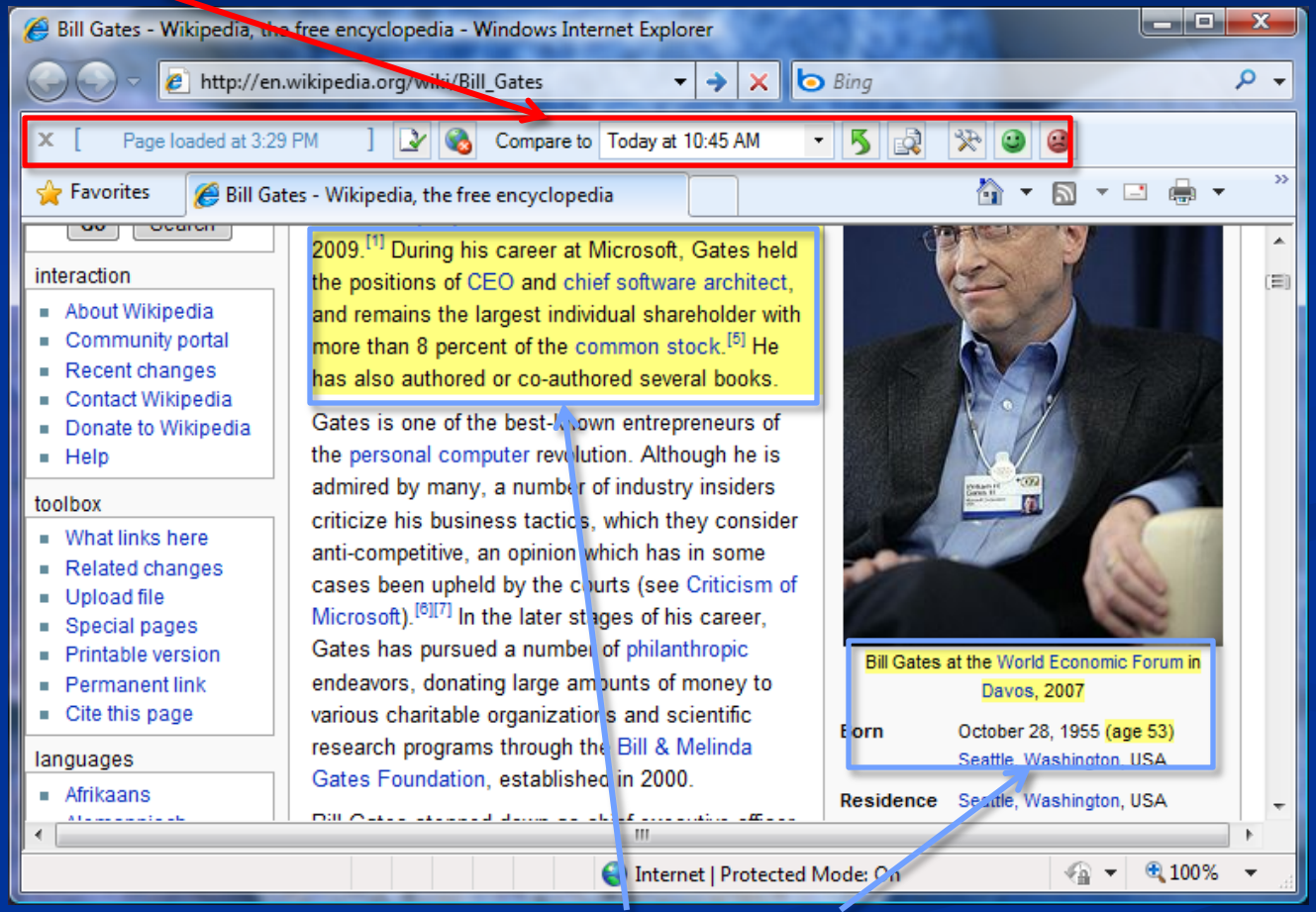


User Visitation/ReVisitation

[Teevan et al., UIST 2009]

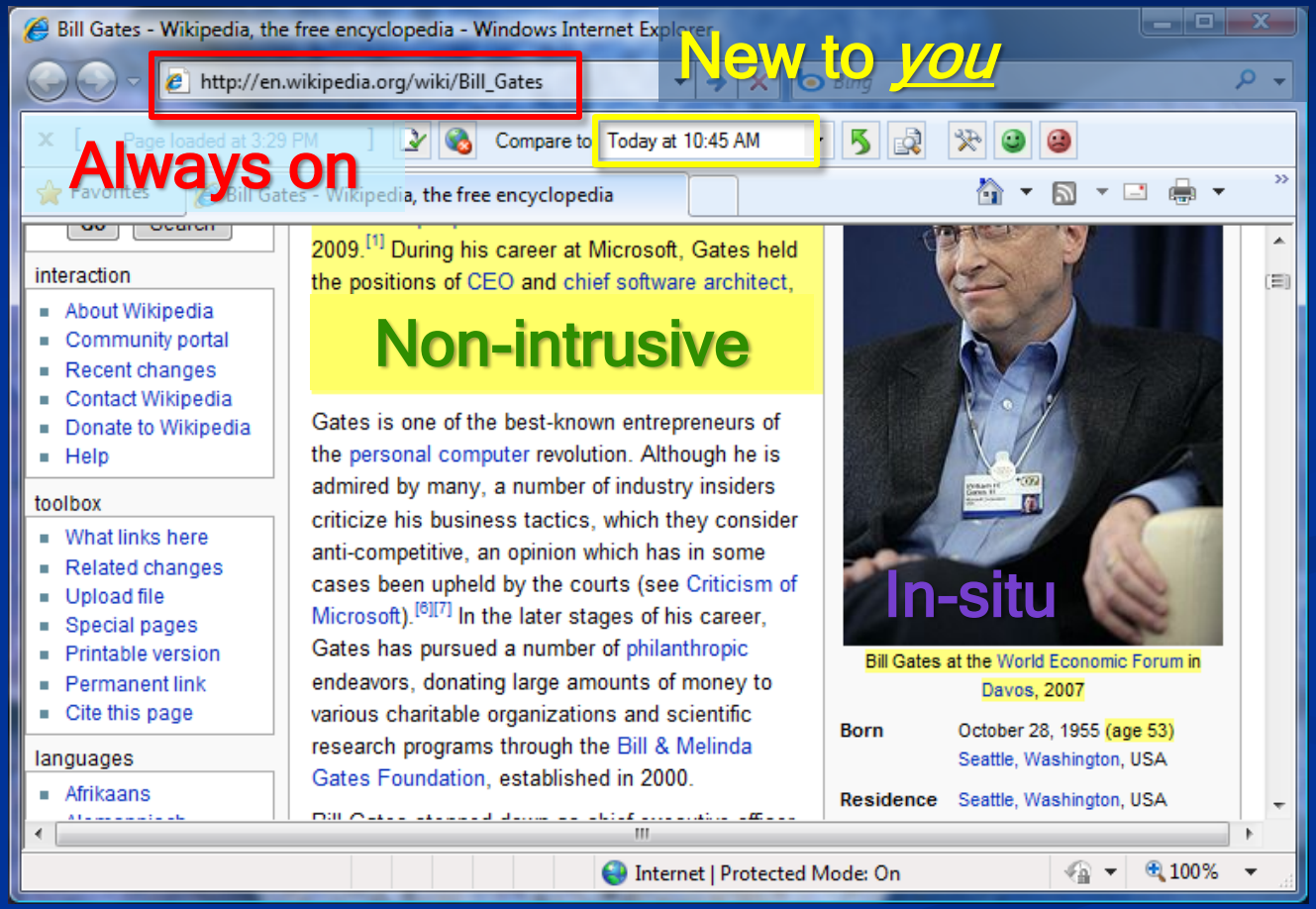
Diff-IE

Diff-IE toolbar



Changes to page since your last visit

Interesting Features of Diff-IE

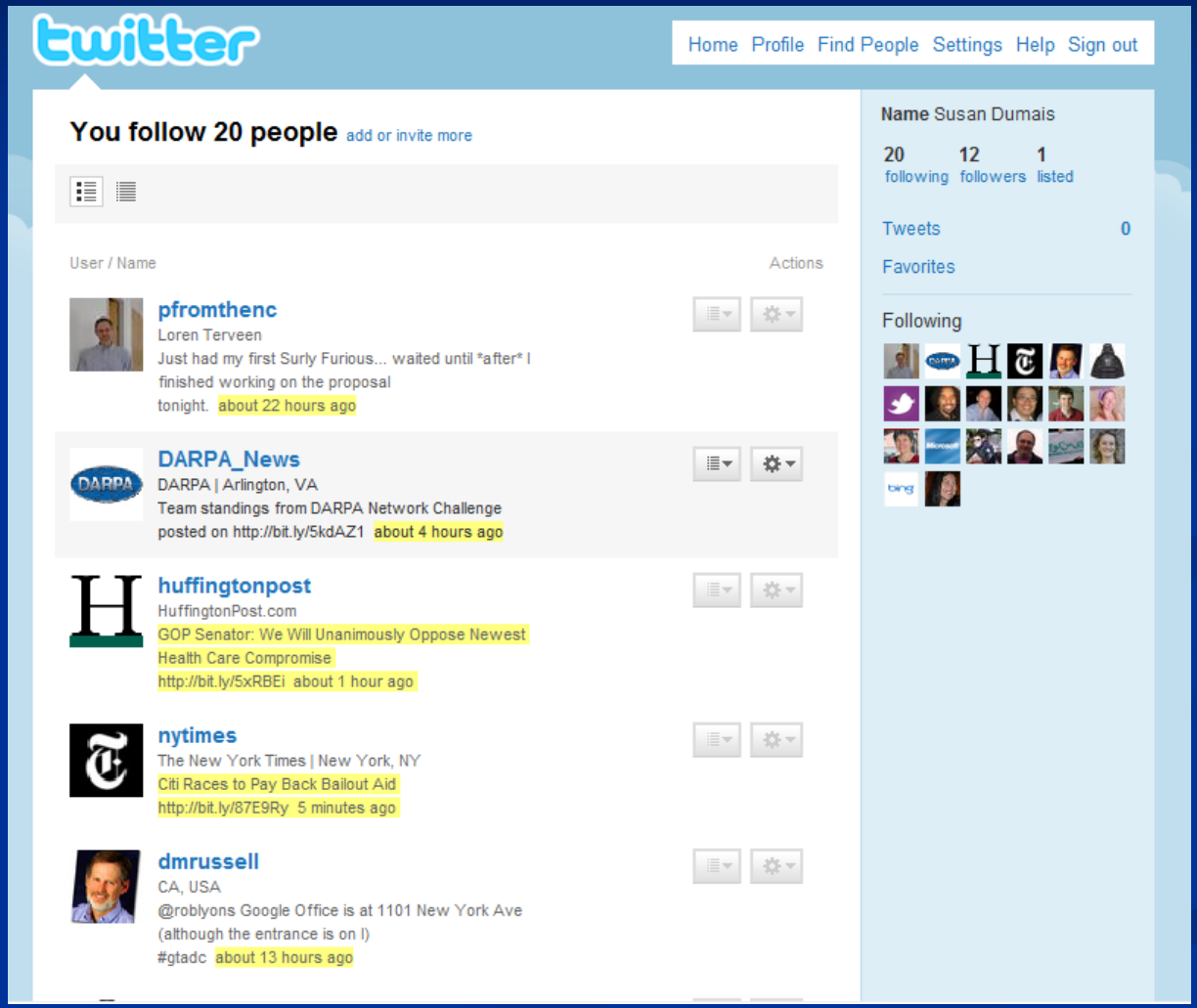


Try it: http://research.microsoft.com/en-us/projects/diffie/default.aspx

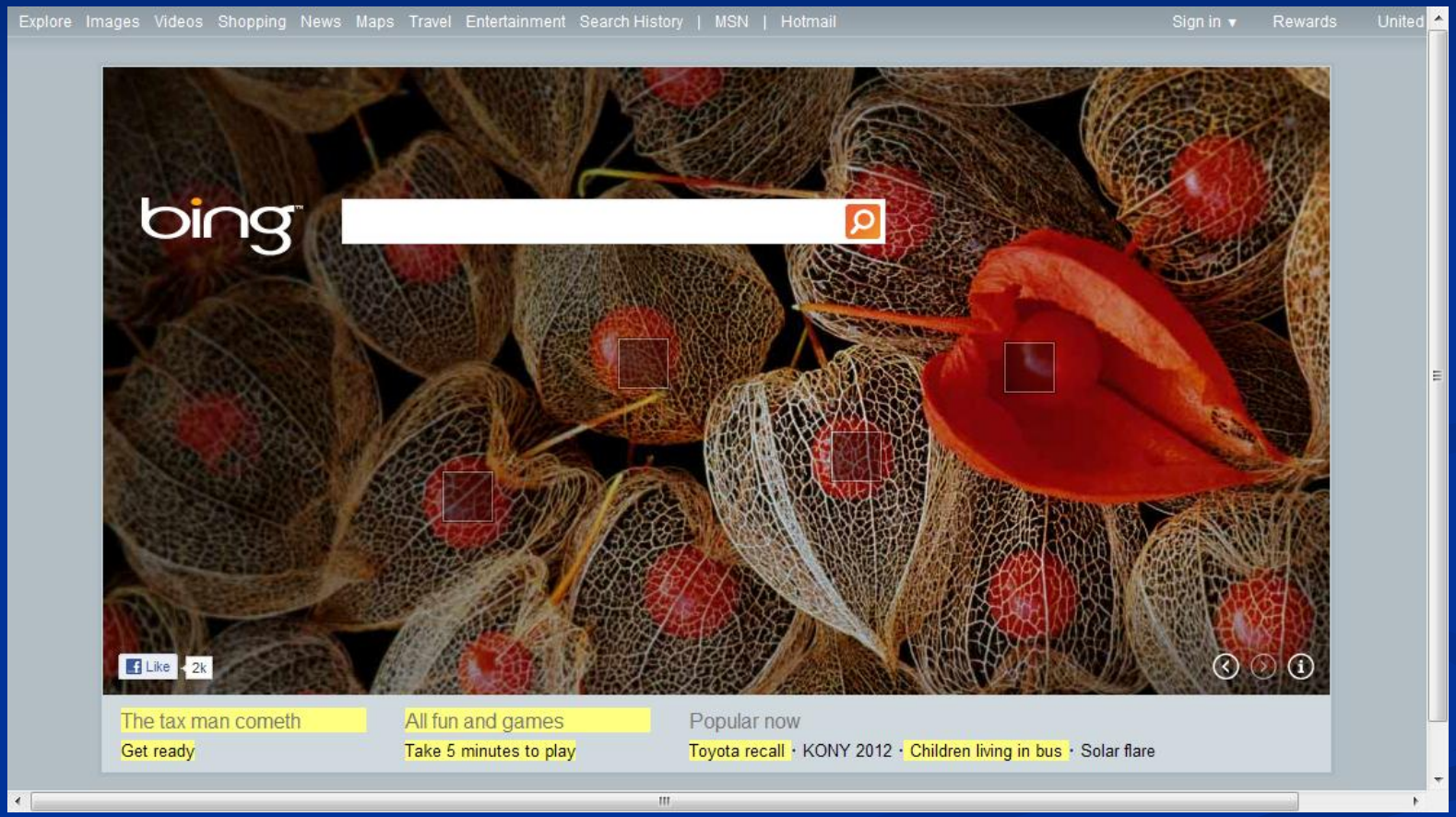
Examples of Diff-IE in Action

Expected New Content

Monitor

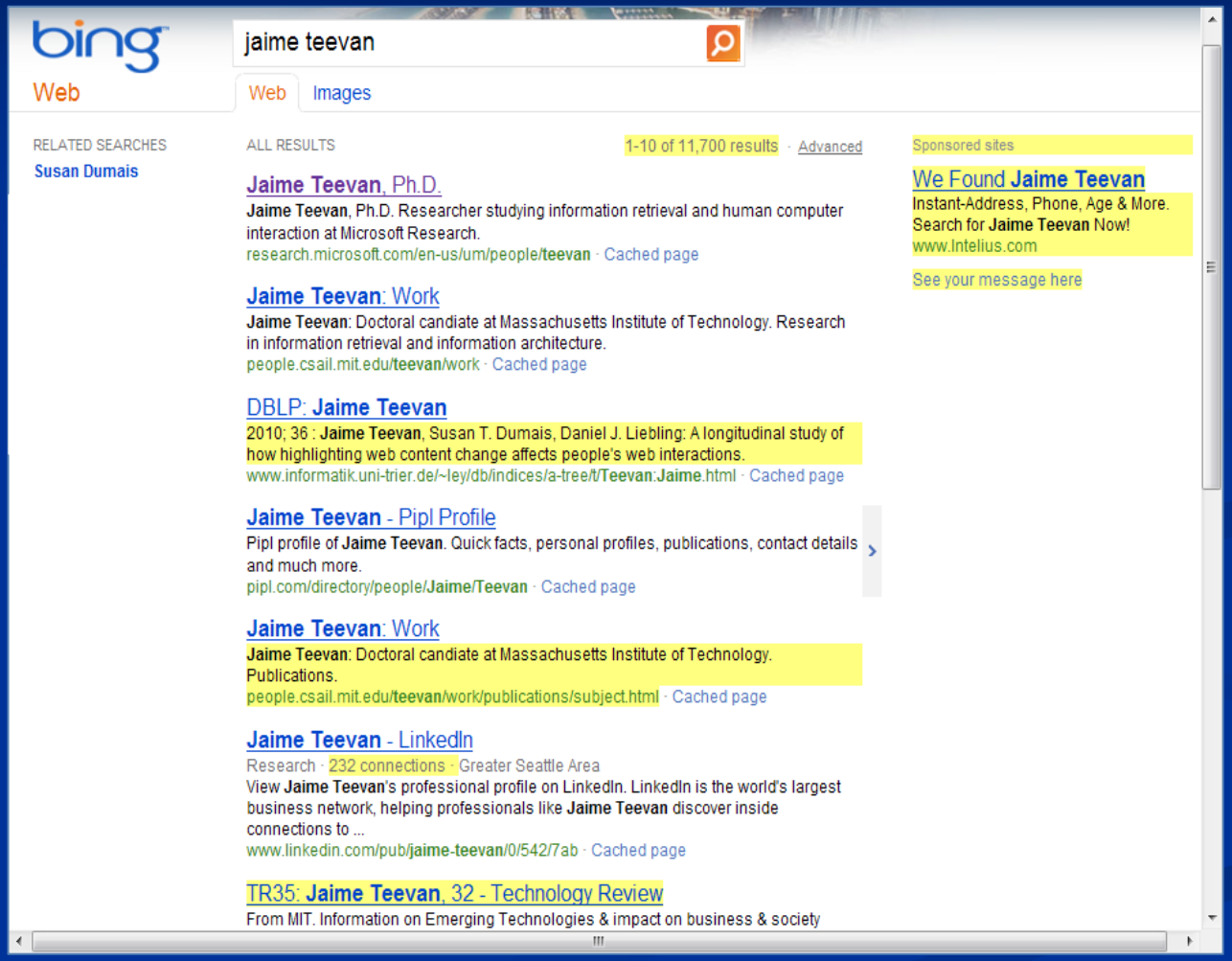

twitter
Home Profile Find People Settings Help Sign out
You follow 20 people add or invite more
User / Name
Actions
pfromthenc
Loren Terveen
Just had my first Surly Furious... waited until *after* I finished working on the proposal tonight. about 22 hours ago
DARPA_News
DARPA | Arlington, VA
Team standings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

Unexpected

Expected
New Content

Unexpected
Important Content

Edit

Attend to Activity

Understand
Page Dynamics

Monitor

Serendipitous
Encounter

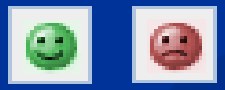
Unexpected
Unimportant Content

[Teevan et al., CHI 2010]

Studying Diff-IE

- Internal study of Diff-IE (3k people, 1+ months)
- Logging
  - Hash of URLs visited
  - Amount of change when revisited

- Feedback buttons

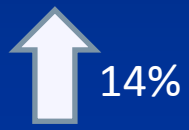


- Survey
  - Prior to installation
  - After 1 month of use

- Experience interview

Diff-IE Results

- People revisit more w/ Diff-IE
  - 1st week: 39.4% of visits are revisits
  - 4th week: 45.0% of visits are revisits
- Perception of change increases
- Amount of change seen increases
  - 1st week: 21.5% revisits changed, by 6.2%
  - 4th week: 32.4% revisits changed, by 9.5%
- Diff-IE is driving visits to changed pages
  - It supports people in understanding change



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 improve 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

$$P(D \mid Q) \propto P(D) \; P(Q \mid D)$$

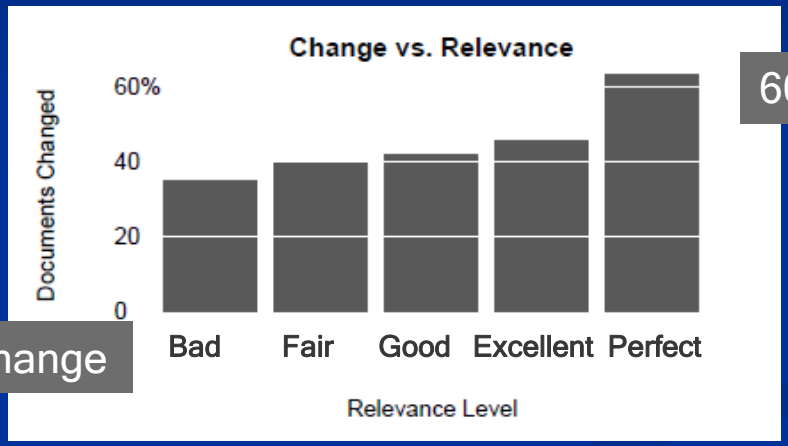
Change prior Term longevity

Temporal Retrieval Expt 1

- 18K queries, 2.5M returned documents
  - Crawled docs weekly, for 10 weeks
  - Judged docs for relevance, only once
    - 5-point scale: Perfect/Excellent/Good/Fair/Bad
- Subset of 2k “navigational” queries
  - Queries that have a “Perfect” judgment
  - Assume these relevance judgments are consistent over time
- Measure changes in nDCG

Relevance and Page Change

- Page change is related to relevance



60% “Perfect” pages change

30% “Bad” pages change

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

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

Change prior

Relevance and Term Change

- Terms patterns vary over time

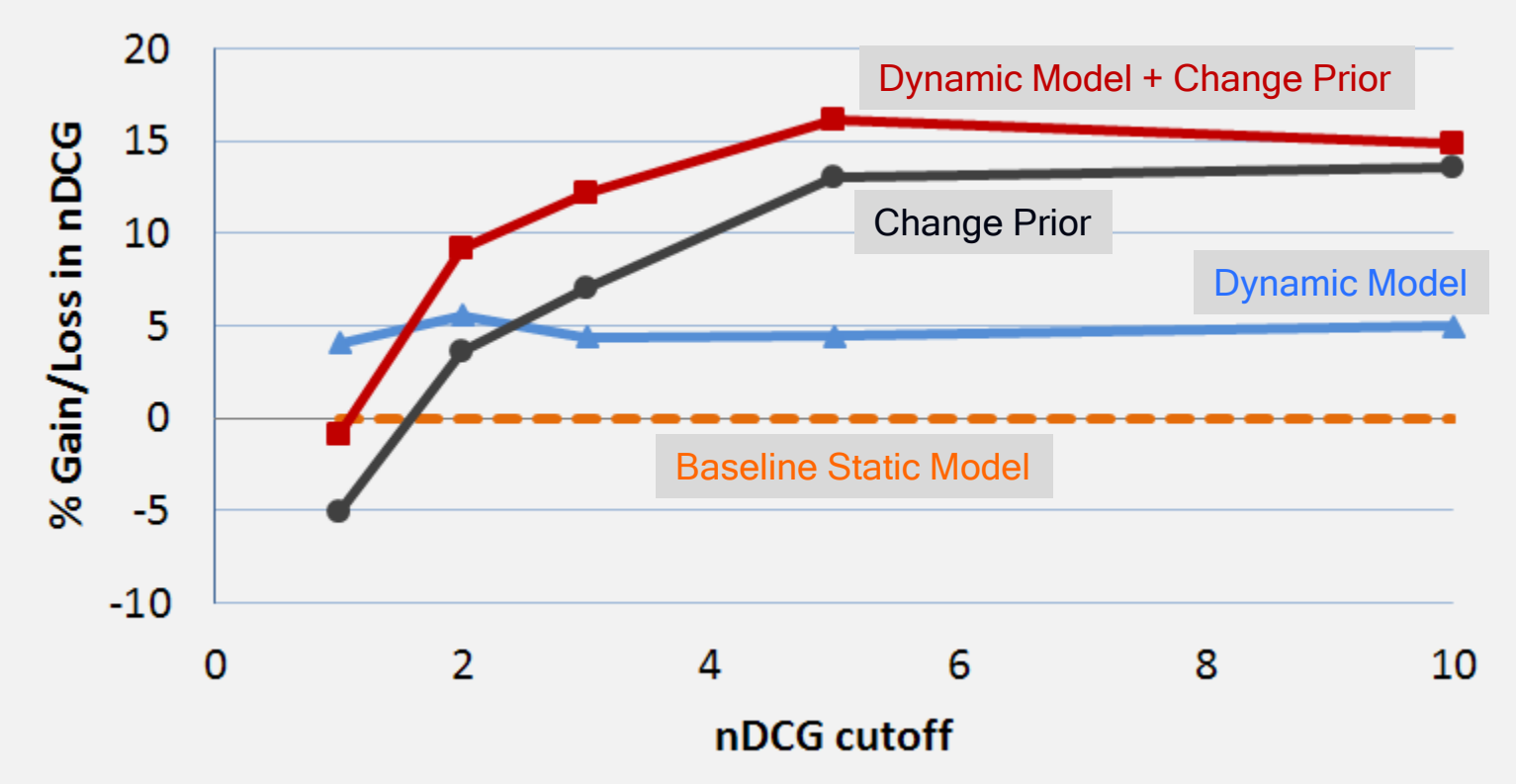
- Represent a document as a mixture of terms with different "staying power"
  - Long, Medium, Short

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

$$P(D|Q) \propto P(D) P(Q|D)$$

Term longevity

Temporal Retrieval Results



[Radinsky et al., WWW2012, submitted]

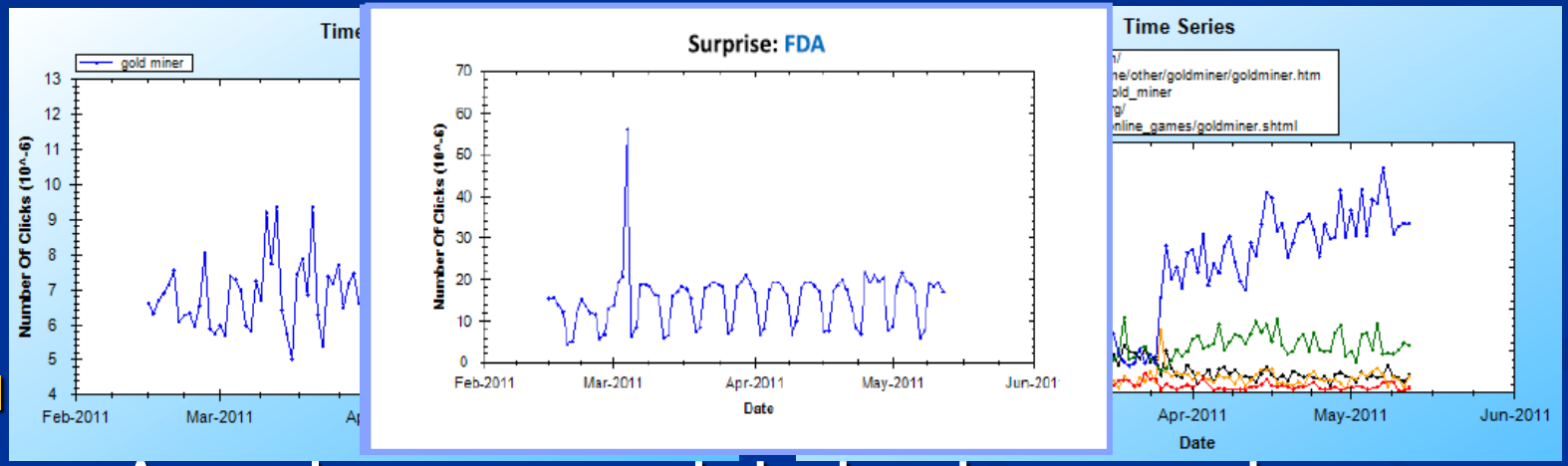
Temporal Retrieval Models 2
(behavior-based)

- Initial evaluation: assumed relevance is "static" over time
- But, relevance often changes over time
  - E.g., *SIAM SDM* -- in 2012 vs. in 2011
  - E.g., *US Open 2012* -- in June (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 relevance judgments, query frequency, interaction data, and page content over time

Relevance over Time

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

Temporal Retrieval Expt 2

- Data
  - Queries and clicked URLs, over 4 months
  - 

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

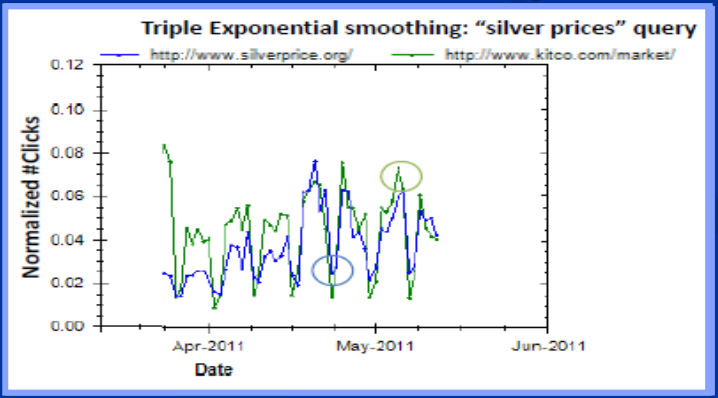
Time Series Modeling

- Model search behavior as time series
  - Assume that the series of behavioral observations $Y_1 \ldots Y_n$ 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 time t, then a semi-linear state space model is defined by:

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

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

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



Experimental Details

- Learn: Time series models of user behavior
  - Can be query or URL dependent
- Predict: Future query and click behavior
- Ranking models
  - Predicted clicks as the only feature for ranking
  - Temporal (+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; Win/Loss

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 content + 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, w using our temporal models. Statistically significant differences based on a paired t-test performing algorithm ($p < .05$) are shown in bold.

- Best-performing queries

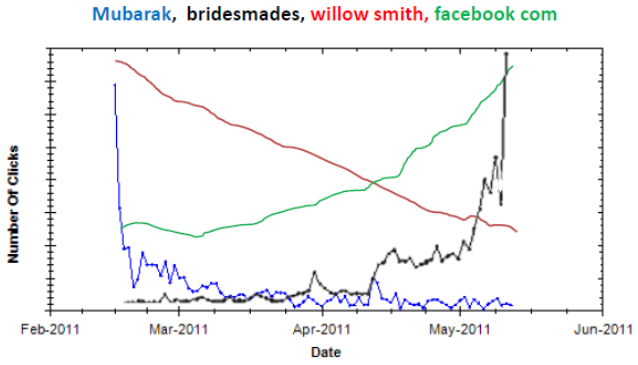
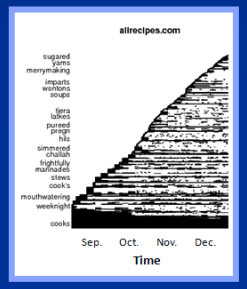
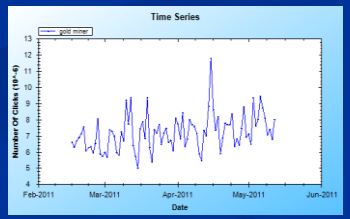


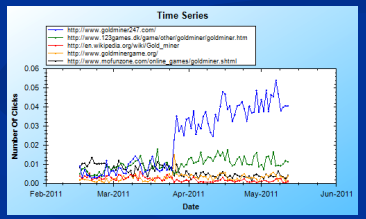
Figure 6: Dominant query shapes for queries where temporal model yielded better rankings than baseline rankers.

Temporal IR Summary

- Improve Web search by modeling temporal dynamics
- Content-based models
  - Rate of page change
  - Detailed term-level changes
- Behavior-based models
  - Query frequency over time
  - URL click patterns over time
- Ongoing work
  - Combing content and behavior features
  - Detecting surprise and periodicity
  - Modeling events







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

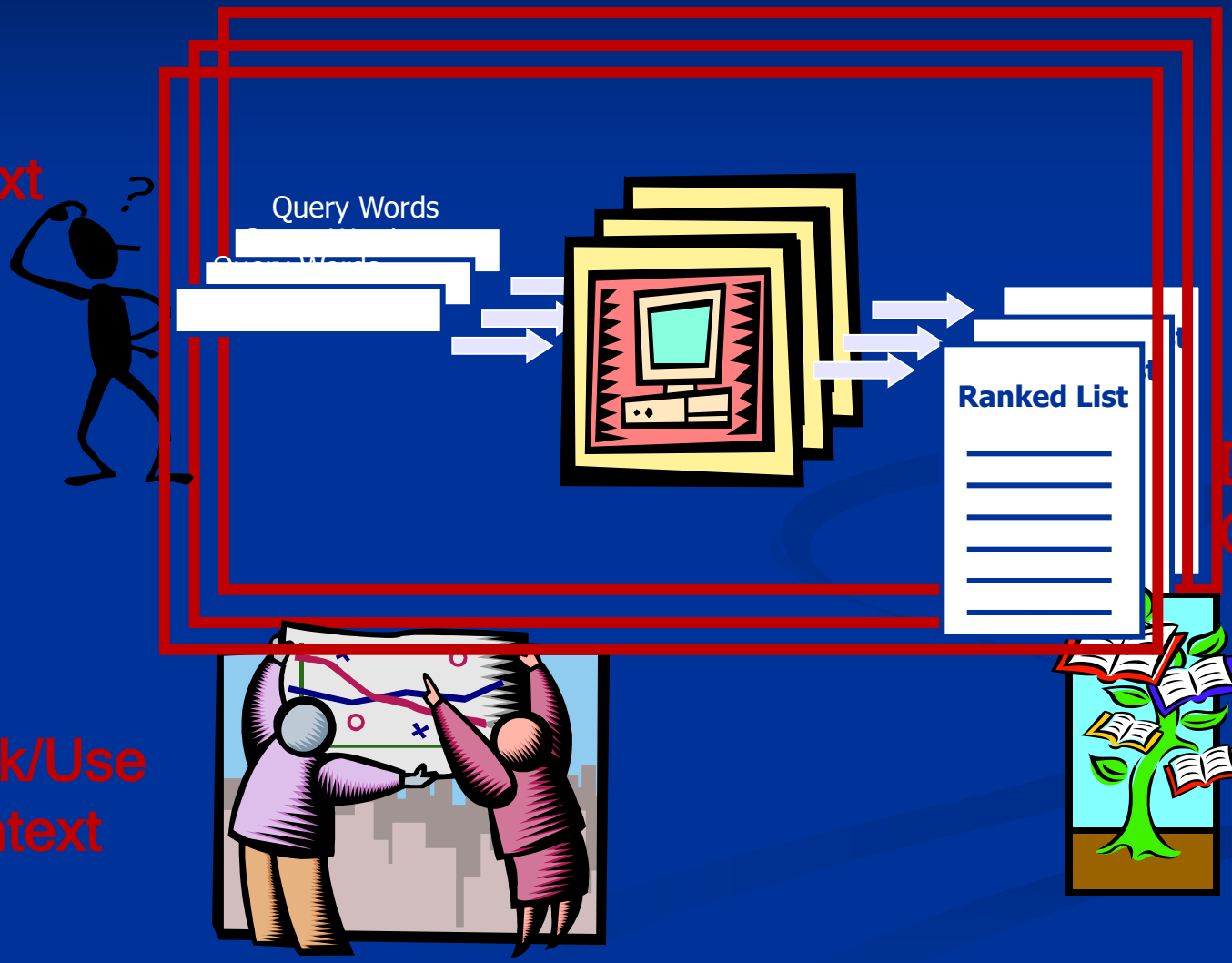
People revisit and re-find Web content

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

Opportunities and Challenges

- 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, individual, tasks, etc.

Think Outside the (Search) Boxes



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