



*The 32nd Annual ACM SIGIR Conference*

*July 19-23 2009*

## **Salton Award Lecture**

# **An Interdisciplinary Perspective on IR**

Susan Dumais

Microsoft Research

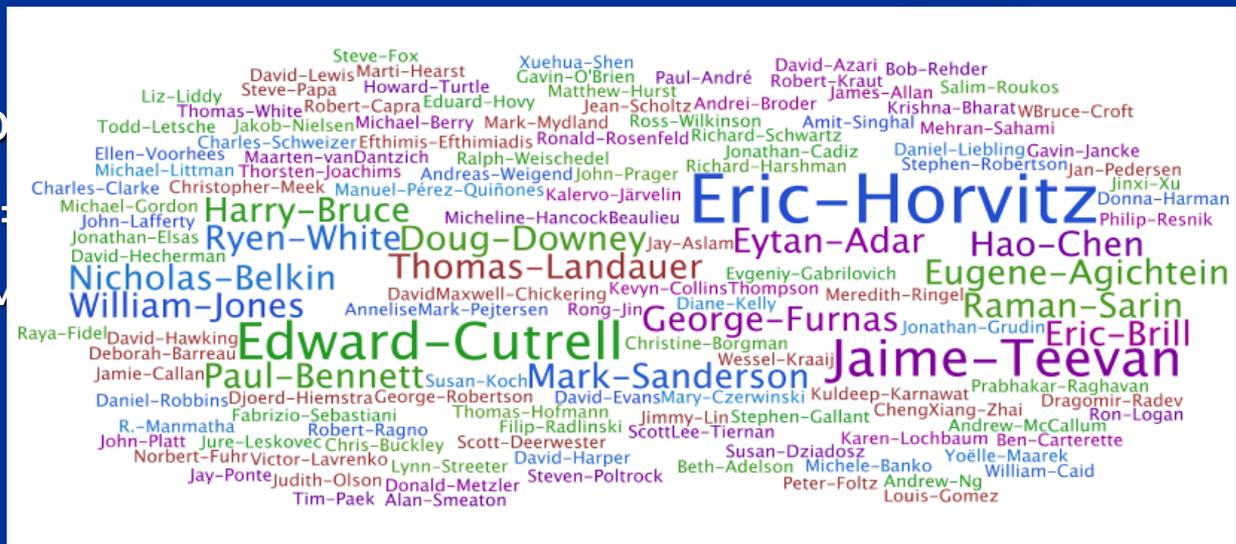
# Thanks!

- Salton Award Committee
- Many great colleagues
  - 1979-1997, Bell Labs/Bellcore
  - 1997-present, Microsoft Research
  - Many other collaborators ...

- Tremendous help from

- Salton number 1:

- Michael Lesk: SIGIR
- CHI 1995 Panel:



# Overview

## ■ Personal reflections

- My research is interdisciplinary, at the intersection of IR and HCI
- User-centric vs. system-centric
- Empirical vs. theoretical
- Evaluation via many methods
  - Test collections, field work, prototypes, deployment experiences, lab studies, etc.

## ■ My background

## ■ Common themes

- Understanding user, domain, and task contexts

## ■ Future challenges

- Dynamics, data and more

# Background

- Mathematics and Psychology
- HCI group at Bell Labs, 1979
- Introduction to IR, 198

- The problem(s) ...

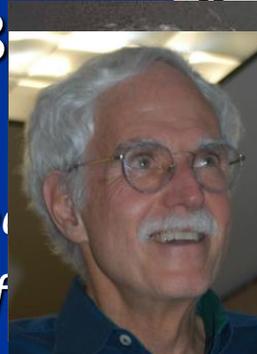
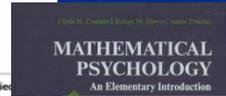
- *Human factors in database*
- *Describing categories of*
- *Verbal disagreement/Statistical semantics/Vocabulary problem*

- Some solutions & applications ...

- Rich aliasing / Adaptive indexing / Latent semantic indexing

- Closing the loop back to psychology ...

- *A solution to Plato's problem* [Psychological Review, 1997]



# From Verbal Disagreement to LSI

- Observed: Mismatch between the way that people want to retrieve information from a computer and the way that systems designers describe that information
  - The trouble with UNIX
  - Command names, menu and category descriptors, keywords
- Studied: How people describe objects and operations
  - Text editing operations, recipes, classified ads, etc.
  - Demo:
  - Data:

**TABLE I. Word-Object Data**

(a) Sample data from the text-editing study

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

(b) Sample data from the common object study

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

# From Verbal Disagreement to LSI

## ■ Findings:

- Tremendous diversity in the name that people use to describe the same objects or actions (aka, “the long tail”)

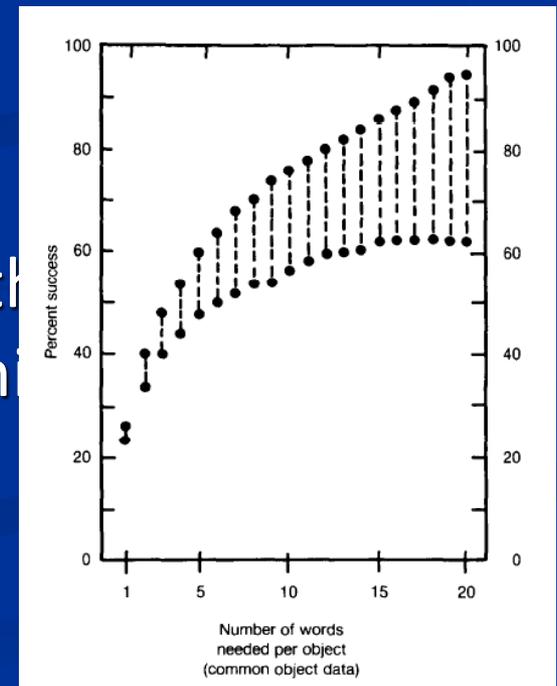
- Single keyword: 0.07 – 0.18 “repeat rate”

- Single normative keyword: 0.16 - 0.36

- Three aliases: 0.38 – 0.67

- Infinite aliasing:

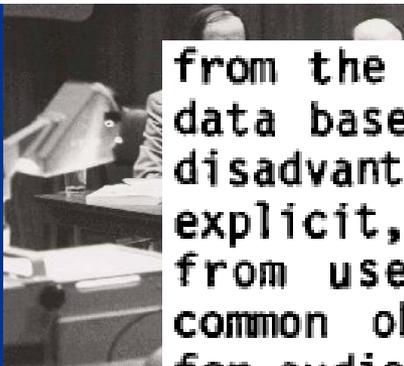
- Interestingly, we have referred to the verbal disagreement, vocabulary mismatch, and statistical semantics



# From Verbal Disagreement to LSI

## ■ CHI 1982 Paper ... 0<sup>th</sup> CHI Conference

STATISTICAL SEMANTICS: HOW CAN A COMPUTER USE WHAT PEOPLE NAME THINGS TO GUESS WHAT THINGS PEOPLE MEAN WHEN THEY NAME THINGS?



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

Photos from  
B. Shneiderman



in that they do not usually get immediate, and continuous feedback. Knowing how people describe objects and shift their descriptions for audiences of different levels of sophistication may help designers build systems whose information is accessible to the widest audience.

secretarial and high typing, but no computer a sample manuscript i. They were asked to of instructions for tually going to make it have the author's allowed us to observe ous names for common is used by non-pro- especially interested ge use. Do different cal units to describe operation? Does the give the same name to n addition, we also items used to specify as a function of the actors, words, lines, paragraphs) and type (i.e., insert, delete, replace, move, transpose) of text unit being changed.

(2) Three hundred thirty-seven college students gave short statements to specify verbal objects. They were given a list of common items like "Newsweek", "Empire State Building", and "motorcycle", and asked to

# From Verbal Disagreement to LSI

- Some solutions: ... with a lot of help from our friends
- Rich aliasing [Gomez et al. 1990]
  - Allow alternative words for the same
  - “Natural” in the world of full-text index command naming
- Adaptive indexing [Furnas 1985]
  - Associate (failed) user queries to destination objects
  - Add these queries as new entries in term-document matrix
  - Quickly reduces failure rate for common requests/tasks
- Latent Semantic Indexing [Dumais et al. 1988; Deerwester et al. 1990]
  - Model relationships among words, using dimension reduction
  - Especially useful when query and documents are short
  - Baker, Borko/Bernick, Ossario (1962-1966); Kohl (SIGIR 1978, p.1)



# From Verbal Disagreement to LSI

- Many applications and algorithms of LSI
  - Bell Labs directory of services, expert finding, reviewer assignment, handwritten notes, data evidence analysis, measurement of knowledge, literature-based discovery, IR & IF test collections
- Rich aliasing and Adaptive indexing in Web era
  - Full text indexing (rich aliases from authors)
  - Anchor text or Tags (rich aliases from other users)
  - Historical query-click data (adaptive indexing, with implicit measures)

# Common Themes

- The last 10-20 years ... amazing time to be involved in IR
- TREC and related evaluations
  - TREC-1 in 1992
- Search is everywhere – desktop, enterprise, Web
- Web search
  - Big advances in scale, diversity of content and users, quality of results (for some tasks), etc.
- SIGIR community has a lot to be proud of
- But ... many search tasks are still quite hard
  - Need to represent and leverage richer contextual information about users, domains, and task environments in which search occurs

# Web Search at 15

## What's available

### ■ Number of pages indexed

- 7/94 Lycos –
- 95 –  $10^6$  millions
- 97 –  $10^7$
- 98 –  $10^8$
- 01 –  $10^9$  billions
- 05 –  $10^{10}$  ...

### ■ Types of content

- Web pages, newsgroups
- Images, videos, maps
- News, blogs, spaces
- Shopping, local, desktop
- Books, papers, many formats
- Health, finance, travel ...

## How it's accessed



# Support for Searchers

- The search box
- Spelling suggestions
- Query suggestions
- Auto complete
- Inline answers
- Richer snippets
- But, we can do better



*Search in the future will look nothing like today's simple search engine interfaces, [Susan Dumais] said, adding, "If in 10 years we are still using a rectangular box and a list of results, I should be fired." [Mar. 7, 2007, NYTimes, John Markoff]*

# Search and Context

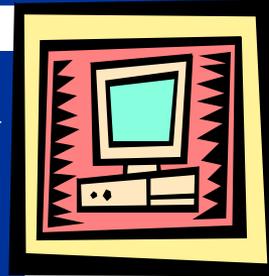
User  
Context



Query Words



Query Words



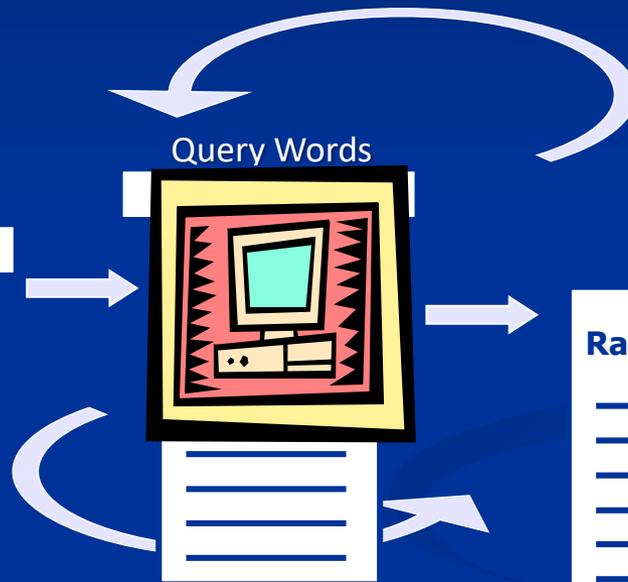
Ranked List



Document  
Context



Task/Use  
Context



## Systems/Prototypes

- New capabilities and experiences
- Algorithms and prototypes
- Deploy, evaluate and iterate

## Inter-Relationships among Documents

### Categorization and Metadata

Reuters, spam, landmarks, web categories ...

Domain-specific features, time

### Interfaces and Interaction

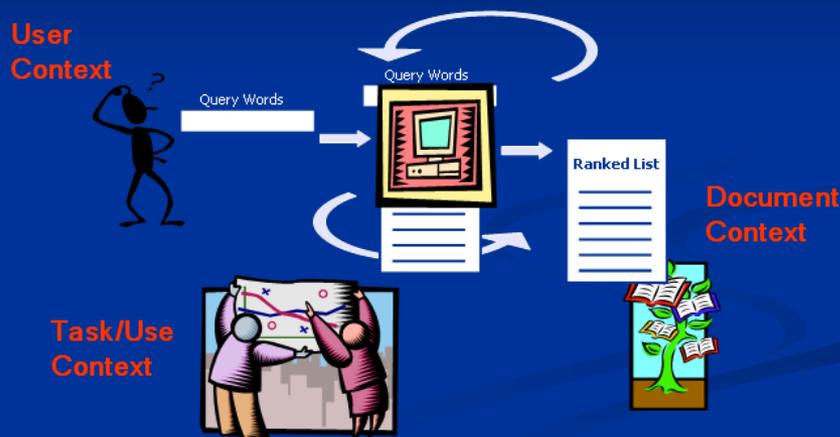
Stuff I've Seen, Phlat, Timelines, SWISH

Tight coupling of browsing and search

## Redundancy

## Temporal Dynamics

## Search and Context



## Modeling Users

Short vs. long term  
Individual vs. group  
Implicit vs. explicit

## Using User Models

Stuff I've Seen (re-finding)

Personalized Search

News Junkie (novelty)

User Behavior in Ranking

Domain Expertise at Web-scale

## Evaluation

- Many methods, scales
- Individual components and their combinations

# User Modeling

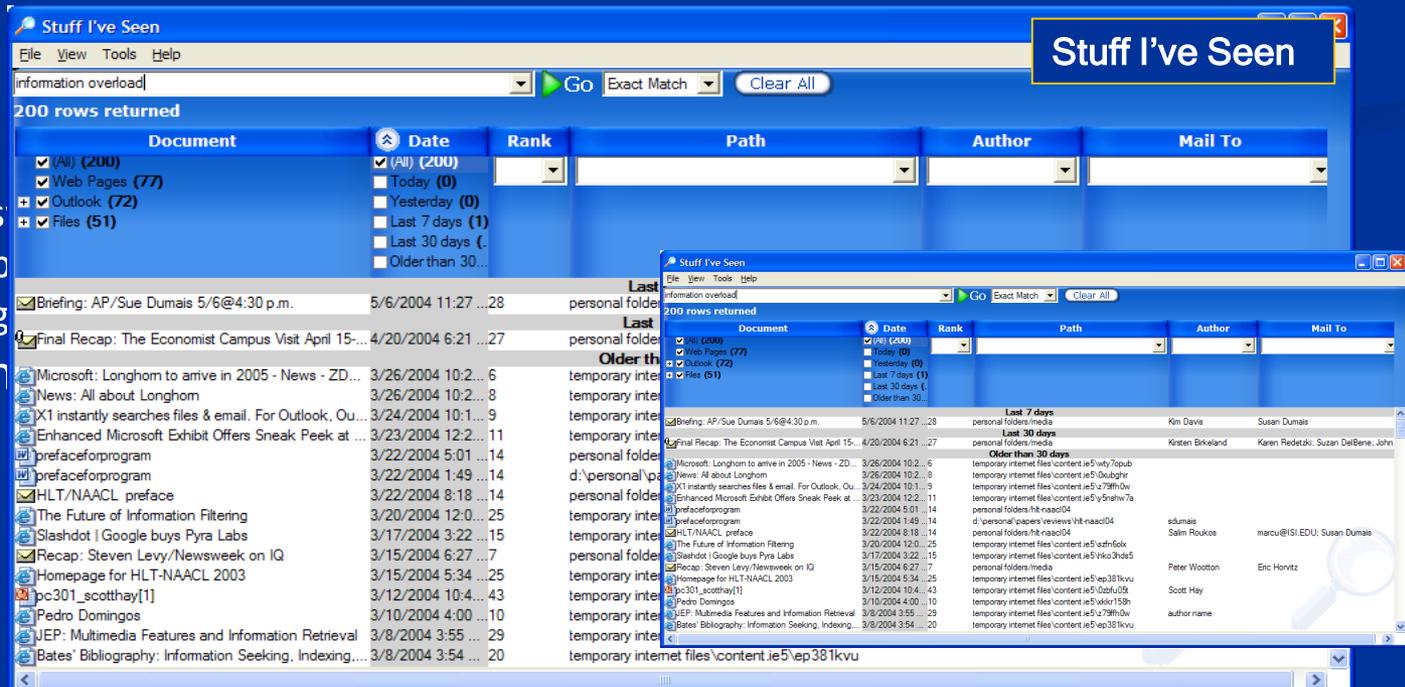
- Modeling searcher's interests and activities over time
  - Iterative and interactive nature of search
  - Within and across sessions
- Example applications
  - Re-finding (e.g., Stuff I've Seen, Web) [Dumais et al. 2003]
  - Personalization (e.g., PSearch) [Teevan et al. 2005]
  - Novelty (e.g., News Junkie) [Gabrilovich et al. 2004]
  - Domain expertise at Web-scale [White & Dumais 2009]
  - User behavior for Web ranking [Agichtein et al. 2006]
- Evaluation via explicit judgments, questionnaires, client-side instrumentation, and large-scale search logs, lab and field studies, etc.

# Re-Finding on the Desktop

- **Stuff I've Seen (SIS) [Dumais et al. 2003]:**
  - Unified access to many types of info (e.g., files, email, calendar, contacts, web pages, rss, im)
  - Index of content and metadata (e.g., time, author, title, size, usage)
  - Rich UI possibilities, because it's your stuff and client application
  - Demo:

- **Analysis:**

- Deployed
  - Query s
  - Result p
  - Ranking
- Questionn
- Log data,



# Re-Finding on the Desktop

- Research Results:
  - Short queries
    - Few advanced operators in initial query (<10%)
    - Many advanced operators via specification in UI (~50%) - filter; sort
  - Date by far the most common sort attribute (vs. best-match)
    - Importance of time, people, episodes in human memory
    - Few searches for “best match”; many other criteria
  - Need for “abstractions” – date, people, kind
  - Rich client-side interface
    - Support fast iteration/refinement
    - Fast filter-sort-scroll vs. next-next-next
- Interesting reviews from SIGIR 😊
- Practice: XP and Vista desktop search

# Re-Finding on the Web

- 50-80% page visits are re-visits
- 30-50% of queries are re-finding queries

Data from Teevan et al., SIGIR 2007

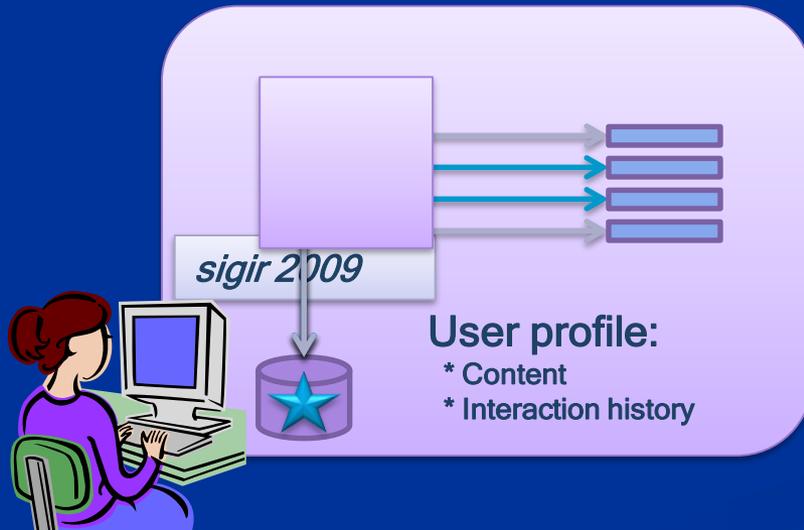
		Repeat Click	New Click
Repeat Query	33%	29%	4%
New Query	67%	10%	57%
		39%	61%

Total = 43%

- Big opportunity to support re-finding on Web
- Models to combine Web rank w/ personal history of interaction
- Interfaces to support finding and re-finding

# Personalization

- Today: People get the same results, independent of current session, previous search history, etc.
- PSearch [Teevan et al. 2005]: Uses rich client-side model of a user to personalize search results



# Personalization

- Building a User Profile
  - Type of information
    - Content: Past queries, web pages, desktop
    - Behavior: Visited pages, explicit feedback
  - Time frame: Short term, long term
  - Who: Individual, group
  - Where the profile resides:
    - Local: Richer profile, improved privacy [but, increasingly rich public data]
    - Server: Richer communities, portability
- Using the User Profile
  - Ranking
  - Query support
  - Result presentation

*PSearch*

# Personalization

- Ranking algorithm [Teevan et al. 2007]
  - Linear combination of scores from: content match, history of interaction, Web ranks

- When to personalize

- Personalization works
- Models for predicting the query and query

- Evaluating personalization

- What's relevant for
- Explicit judgments
- Implicit "judgments"
- Linking explicit and

The screenshot shows a Microsoft Internet Explorer browser window displaying search results for 'www2005'. The search results include links to 'The 14th International World Wide Web Conference 2005', 'WWW2005 Call For Paper', and 'Web Engineering.org Community'. A 'Curious Browser' pop-up window is overlaid on the search results, asking 'Did you find the information you needed at this search result?' and providing three response options: 'Yes' (green smiley), 'Sort of' (yellow neutral), and 'No' (red frowny). The 'Yes' button is highlighted. Below the pop-up, a blue box contains the text: 'Curious Browser Study (~4k) \* 45% w/ just click \* 75% w/ click + dwell + session'. The browser's address bar shows 'http://search.msn.com/results.aspx?srch=1058FORM=AS55q=www2005'.

# Categorization and Metadata

- Algorithms and applications
  - Reuters, Web - fast SVM algorithm [Dumais et al. 1998, 2000]
  - Junk email [Sahami et al. 1998]
    - Domain-specific feature engineering
    - Constantly changing content (both ham and spam)
- Using metadata for ranking [Bennett et al.]
- Using metadata in UX
  - Tight coupling search & browse – e.g., SIS, Phlat [Dumais et al. 2003]
  - Faceted-metadata in many verticals -> Web? [Teevan et al. 2008]
  - Information theoretic models of search/navigation [Downey et al. 2008]
- Leveraging relations among documents

# Future Challenges

- Dynamic information environments [Adar et al., Elsas et al.]
  - Content changes (e.g., news, blogs, lifelogs ... much more general)
  - People re-visit, re-query, re-find
  - IR opportunities ... crawling, doc and user representation, ranking, etc.
  - Interesting historically and socially
- Data/Evaluation
  - Data as valuable resource
  - Large-scale log data
  - Operational systems and a “Living Laboratory”
  - IR opportunities ... representations, ranking, etc.
- Thinking outside the traditional IR boxes
  - Better understanding of users and application domains
  - Collaborations across disciplinary boundaries

# Information Dynamics

Microsoft Research Homepage

1996



2009

SIGIR 2009

# Information Dynamics

My Homepage

2008

1998

1999

**Susan Dumais**

Senior Researcher, [Adaptive Systems & Interaction Group](#), Microsoft Research  
E-mail: [sdumais@microsoft.com](mailto: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 in July 1997. I look forward to working on a wide variety of information access and management issues, including collaborative filtering, interfaces for combining search and navigation, and user/task modeling. Stay tuned for new developments here.

Prior to coming to Microsoft, I worked on a statistical method for concept-based retrieval known as Latent Semantic Indexing work on the [Bellcore LSI page](#).

**What's New:**

- Forbes article by William Baldwin on our anti-Spam work. [Spam killers](#), *Forbes*, Sept 21, 1998, 254-255
- S. T. Dumais, J. Platt, D. Heckerman and M. Sahami (1998). [Inductive learning algorithms and representations for text](#) appear in: *Proceedings of ACM CIKM98*, Nov. 1998.
- M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz (1998). [A Bayesian approach to filtering junk e-mail](#). (Postscript) *Learning for Text Categorization*, July 27, 1998, Madison, Wisconsin.
- S. T. Dumais (1997). [Tightly coupling structure and search](#). (Powerpoint slides) *SIGIR'97 Workshop on information description of the workshop and position papers can be found at [UMass](#).*
- S. T. Dumais (1997). [Data mining and the Web](#). (Powerpoint slides) Panel at *Knowledge Discovery and Databases (KDD)*, August 14-17, 1997.

**Susan Dumais**

Principal Researcher, [Adaptive Systems & Interaction Group](#), Microsoft Research

E-mail: [sdumais@microsoft.com](mailto:sdumais@microsoft.com)  
Mail: One Microsoft Way, Redmond WA 98052-6399, USA

**We're Hiring at MSR and LiveLabs ...**

We're looking for great folks to advance the state-of-the-art and influence new products in the search arena. We have internships and permanent positions in several areas including: internet search, desktop search, personalization, and novel interfaces for search.

- Intern candidates can apply online at: <http://web.archive.org/web/20070206172001/http://research.microsoft.com/aboutmsr/jobs/internships/default.aspx>
- Permanent employment opportunities at: <http://web.archive.org/web/20070206172001/http://labs.live.com/>

**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 work on a wide variety of information access and management issues, including personal information management, web search, question answering, information retrieval, text categorization, collaborative filtering, interfaces for improved search and navigation, and user/task modeling.

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 \(now Telcorda\) LSI page](#).

**Workshops, Collaborations and Papers:**

- as Latent Semantic Indexing. You can find pointers to this

200

# 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 ...

# Dynamics and Search

- Improved crawl policy
- Improved ranking using [Elsas and Dumais]

## Welcome – Please join us in Boston | SIGIR 2009

The SIGIR 2009 conference opens in just over a week in Boston 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.

**New content:** 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.

[sigir2009.org](http://sigir2009.org)

- Some are always on the

- Show change in snippets
- More general browsing

Welcome -- Please join us in Boston | SIGIR 09

http://www.sigir2009.org/

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)

#### 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%

http://www.bing.com/search?q=sigir+2009&go=&form=QRE&format=rss

# Data and Evaluation

- Data as a critical resource
- Shared IR data resources typically consist of
  - Static collection of documents and queries
  - Judgments of Q-Doc in isolation
  - Judgments with limited context (just the current query)
  - Judges (who are usually not the searcher)
    - ... and these resources often shape the questions we ask
- Search is an inherently interactive and iterative process, so user interaction data, is an especially important resource for the IR community
  - Large-scale log data
  - Operational system as an experimental platform

# Data and Evaluation

## ■ Large-scale log data

- Understanding how user interact with existing systems
  - What they are trying to do; Where they are failing; etc.
- Implications for: models, and interactive systems
- Lemur Query Log Toolbar – developing a community resource !

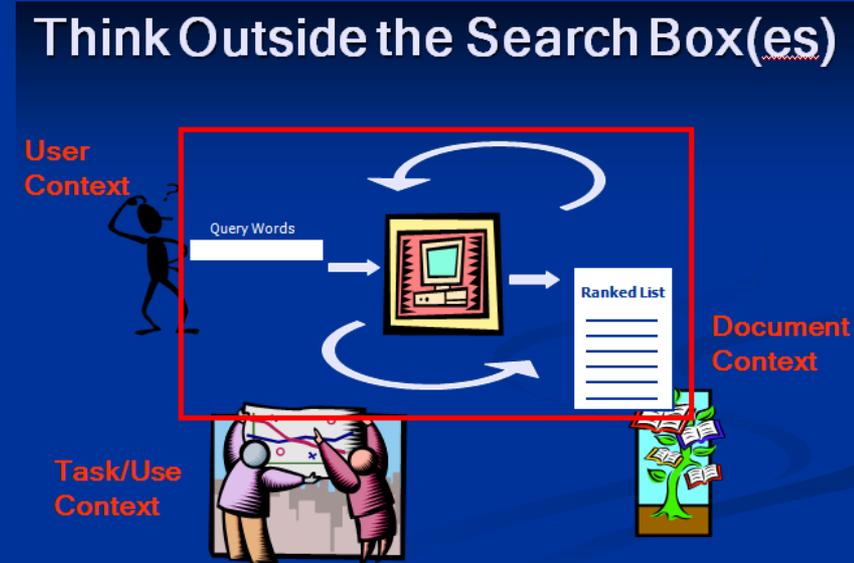
## ■ Operational systems as an experimental platform

- Can also conduct controlled experiments *in situ*
  - Interleave results from different methods [Radlinski & Joachims 2005]
  - A/B testing -- Data vs. the “hippo” [Kohavi 2008]
- Important in: linking offline and interactive results, understanding effect sizes, relations among results (and other page components), etc.
- Can we build such a “Living Laboratory”?

## ■ Replicability in the face of changing content, users, queries

# Opportunities

- Continued improvements in representation and ranking
- Think outside the traditional IR boxes !!!
  - Develop a better understanding of users, and their tasks
  - Design and evaluate interactive systems to support this
- Importance of
  - New data resources
  - Interdisciplinary perspective



# Thanks (again)!

Bell Labs

MSR, CLUES (Context, Learning and User Experience In Search)

