Text REtrieval Conference (TREC)

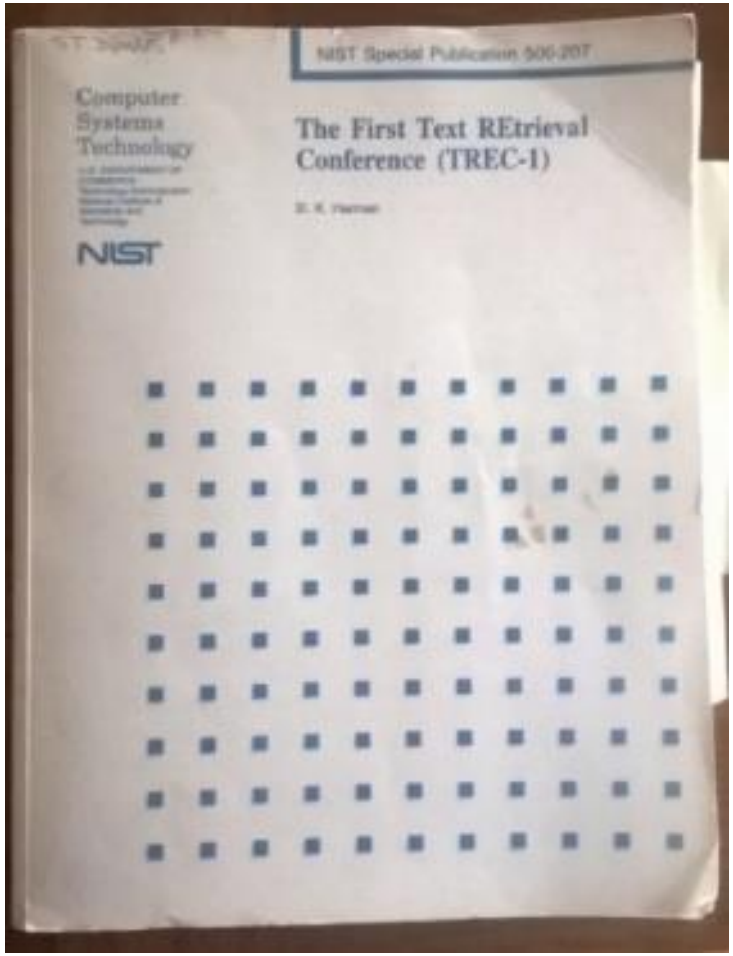
...to encourage research in information retrieval from large text collections.

REFLECTIONS ON TREC@25 AND BEYOND

Nov 15, 2016

Susan Dumais, Microsoft Research

Happy 25th <> TREC !!!



One of the most well-known and celebrated anniversaries is the **25th** wedding **anniversary**, also known as the silver **anniversary**. A milestone this important calls for classy **25th anniversary** gifts, so if you're going the traditional silver route, consider gorgeous jewelry or engraved keepsakes.

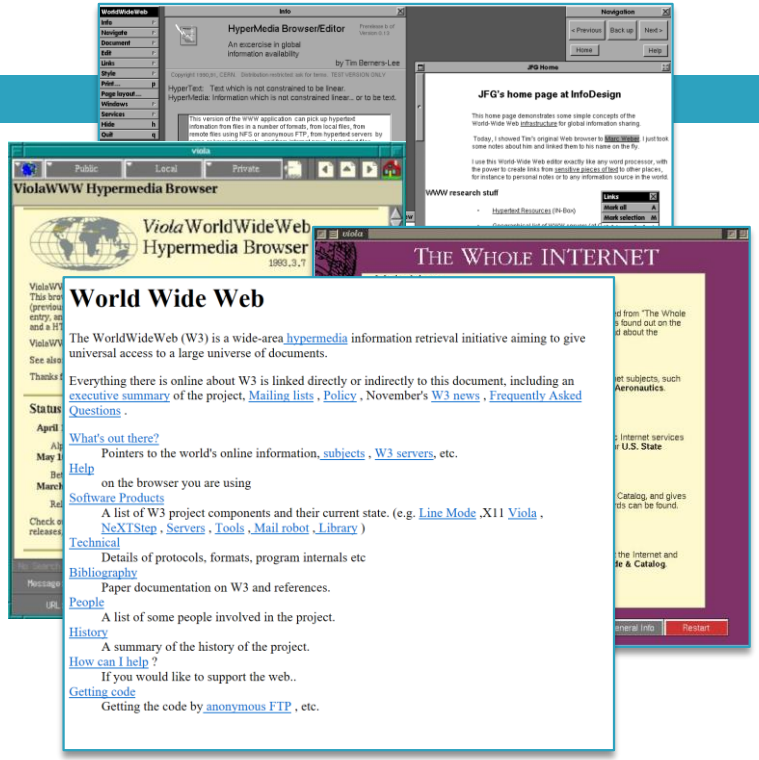
25th Anniversary - Gifts.com
https://www.gifts.com/**anniversary**/**25th-anniversary**/ob6bmJ

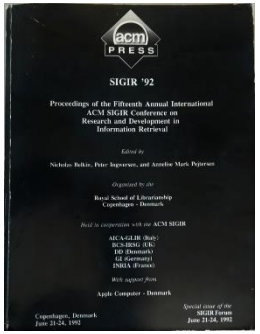
Outline

- Looking back 25 years to 1992
  - In web, search, and TREC-1
- Characterizing the evolving landscape
  - In TREC, search
- Predicting what's next
  - In search

25 Years Ago …

- Rudimentary Web browsers
  - 1990: WorldWideWeb
  - 1992: ViolaWWW & Erwise
- First web site in 1991
  - http://info.cern.ch/
- No Web search engines
  - Search (over web files) in 1990: Archie, Veronica & Jughead
  - Online info systems: Dialog, Medlars, LexisNexis, Westlaw
  - Most common: Online Public Access Catalogs (OPACs)
  - Research in search systems: 15th SIGIR; 1st CIKM



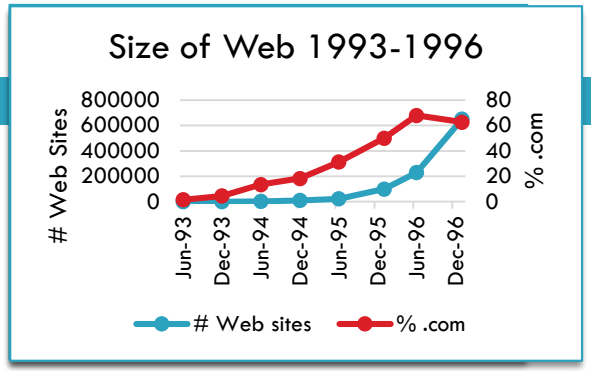


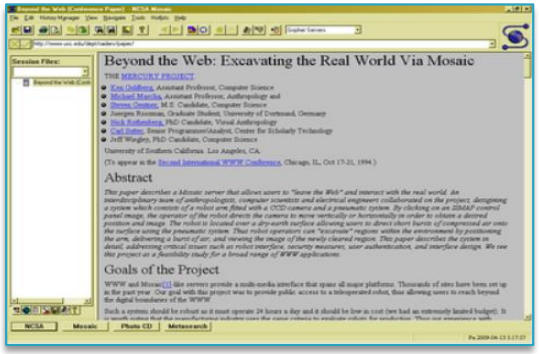
Research Search Systems

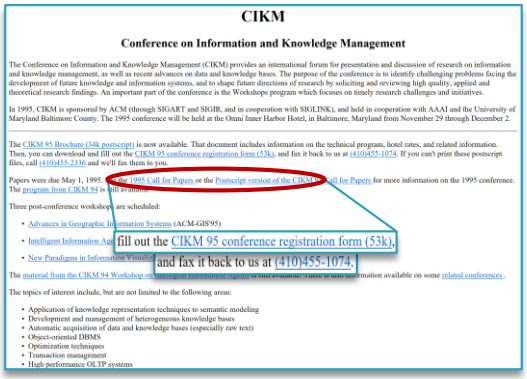
- Research search systems
  - SMART (1960s), Okapi (1980s), INQUERY (1990s), etc.
  - Ranked retrieval, relevance feedback, structure, NL
- Common evaluation collections, ~1-2k docs
  - TIME, MED, CRAN, CISI, CACM, WEST, etc.
- DARPA's TIPSTER program, Phase 1 (1991-1994)
  - Information retrieval, extraction, and summarization
- TREC-1 began in this context

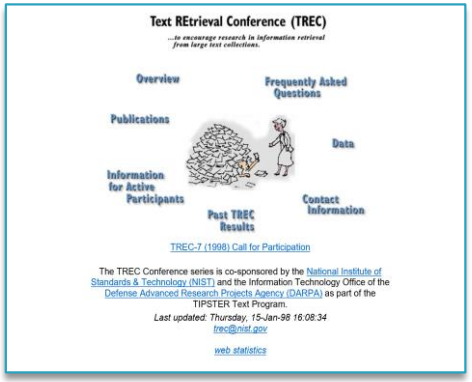
<25 Years Ago … The Web

- The Web was really tiny
  - 130 sites in June 1993
- NCSA Mosaic debuted in 1993
  - 1994 Netscape Navigator
  - 1995 Internet Explorer
- Web presence, ~1995-1997









<25 Years Ago … Web Search

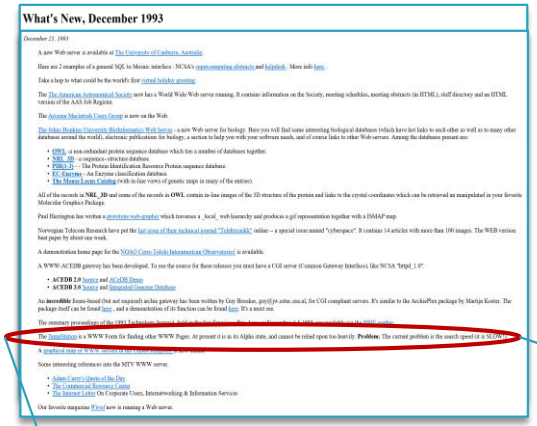
- Early Web search engines 1993-1994
  - Crawl, index, query form & ranking
- 1993 JumpStation,
- 1994 WebCrawler,

From 590i-request@cs.washington.edu Wed Apr 20 23:11:17 1994
Date: Wed, 20 Apr 1994 23:10:35 -0700
From: bp@cs.washington.edu (Brian Pinkerton)
To: 590i@cs.washington.edu
Subject: WebCrawler server up

I finally got a forms query interface hacked together for the WebCrawler index. Give it a try, at

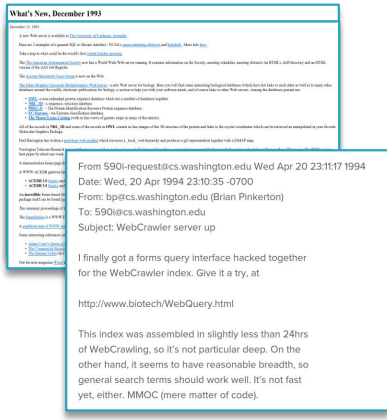
http://www.biotech/WebQuery.html

This index was assembled in slightly less than 24hrs of WebCrawling, so it's not particular deep. On the other hand, it seems to have reasonable breadth, so general search terms should work well. It's not fast yet, either. MMOC (mere matter of code).



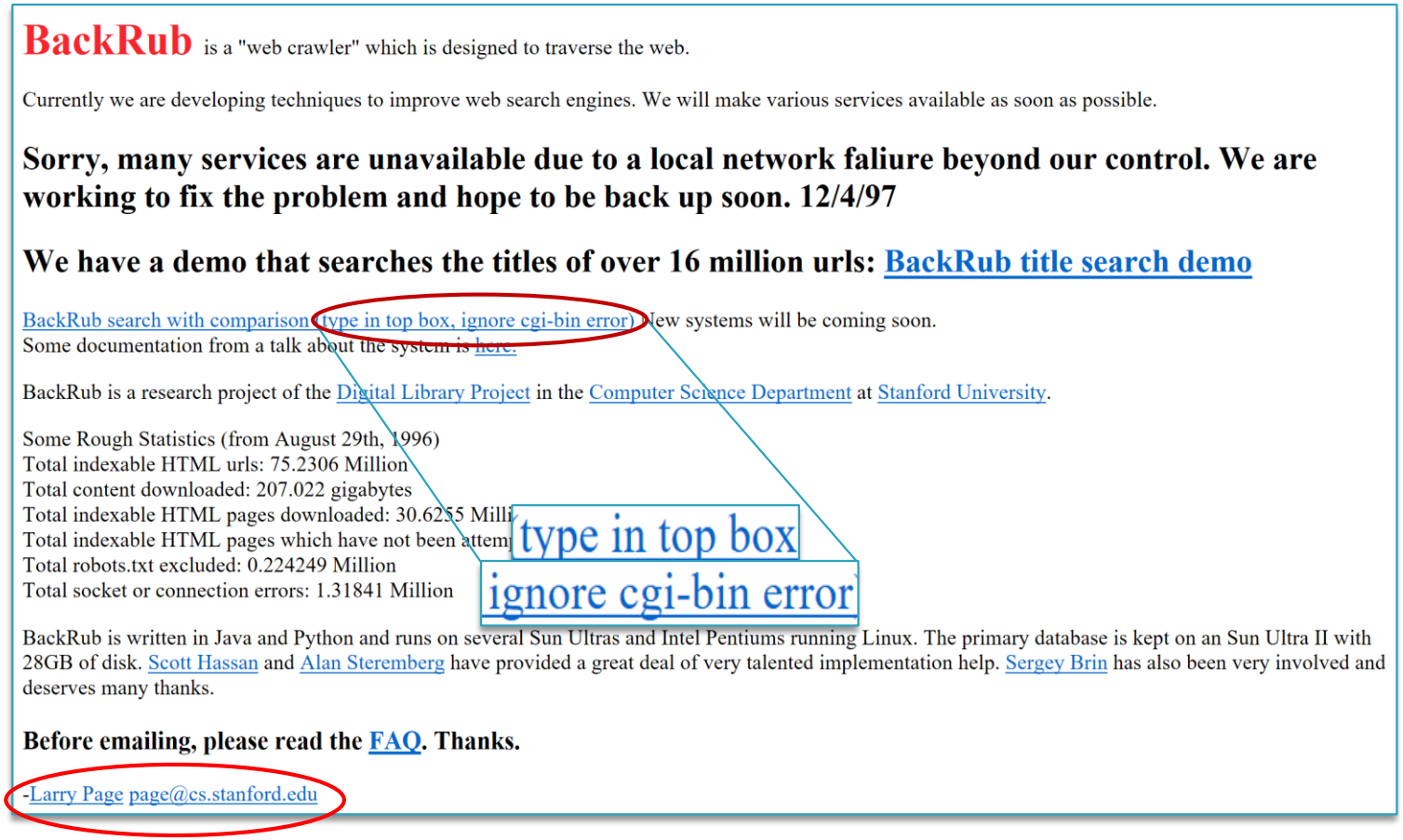
<25 Years Ago … Web Search

- Early Web search engines 1993-1994
  - Crawl, index, query form & ranking
- 1993 JumpStation, WWW Worm, RBSE
- 1994 WebCrawler, Go, InfoSeek, Lycos
  - 2.7k web sites, 50-100k pages, 1.5k queries *[today: 100000x]*
- 1995 AltaVista, Excite, Yahoo!



<20 Years Ago … Web Search

- 1994-1998 NSF Digital Libraries Initiative



BackRub is a "web crawler" which is designed to traverse the web.

Currently we are developing techniques to improve web search engines. We will make various services available as soon as possible.

Sorry, many services are unavailable due to a local network faliure beyond our control. We are working to fix the problem and hope to be back up soon. 12/4/97

We have a demo that searches the titles of over 16 million urls: BackRub title search demo

BackRub search with comparison (type in top box, ignore cgi-bin error) New systems will be coming soon.
Some documentation from a talk about the system is here.

BackRub is a research project of the Digital Library Project in the Computer Science Department at Stanford University.

Some Rough Statistics (from August 29th, 1996)
Total indexable HTML urls: 75.2306 Million
Total content downloaded: 207.022 gigabytes
Total indexable HTML pages downloaded: 30.6255 Milli
Total indexable HTML pages which have not been attem
Total robots.txt excluded: 0.224249 Million
Total socket or connection errors: 1.31841 Million

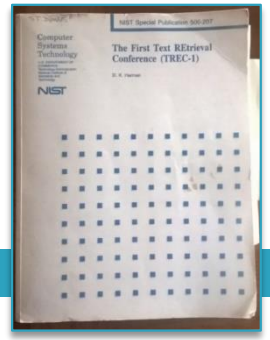
type in top box
ignore cgi-bin error

BackRub is written in Java and Python and runs on several Sun Ultras and Intel Pentiums running Linux. The primary database is kept on an Sun Ultra II with 28GB of disk. Scott Hassan and Alan Steremberg have provided a great deal of very talented implementation help. Sergey Brin has also been very involved and deserves many thanks.

Before emailing, please read the FAQ. Thanks.

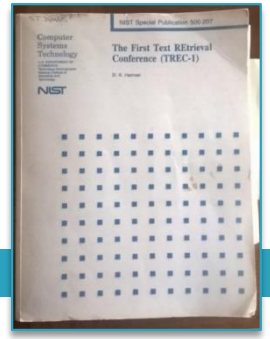
-Larry Page page@cs.stanford.edu

TREC-1: Nov 4-6, 1992



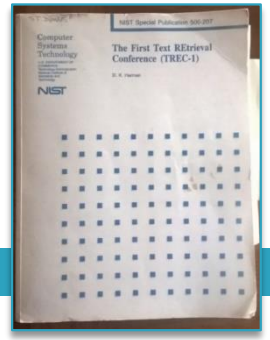
- Co-sponsored by NIST and DARPA (TIPSTER)
- Scale-up Cranfield-style tradition of IR experiments
  - 741k Docs (2 CDs, 2 Gb text), 50 queries adhoc & routing
  - Full text of documents (AP, WSJ, ZD news; Federal Register; DOE)
  - Lots of growing pains for systems and evaluation methods
- Participants: 25 groups, 92 people
  - Some from TIPSTER program, but most not
  - Harman, Buckley, Voorhees, Salton, Cooper, Robertson, Croft, Dumais, Fuhr, Spärck-Jones, Belkin, Allan, Hersh, Moffat, Zobel, Liddy, Callan, …
- Community
  - Some competition, but a real workshop w/ lots of discussion and learning
  - Binders with many preliminary analysis and system details

TREC-1: Nov 4-6, 1992

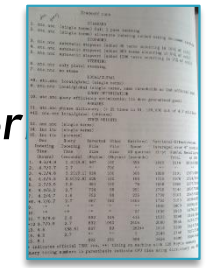


- Wide variety of software maturity and system hardware
- Software
  - Many groups modified IR systems that had existed for decades, but others built from scratch
  - E.g., PARA Group (M. Zimmerman)
    - Routing using Gawk to do line at a time regexp matching reading from the CDs. 11 days for each CDRom of data.
- Hardware
  - Many groups used Sun Sparc or DEC workstations
    - Typical configuration: 8-64 Mb RAM / 25-66 MHz clock rate *[today: 100-1000x]*
  - But also, TRW's *Fast Data Finder* (M. Mettler)
    - Hardware device for high-speed pattern matching on a stream of 8-bit data

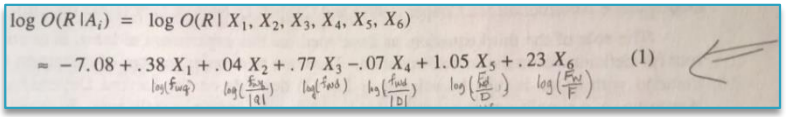
TREC-1: Nov 4-6, 1992



- A few of my favorite results



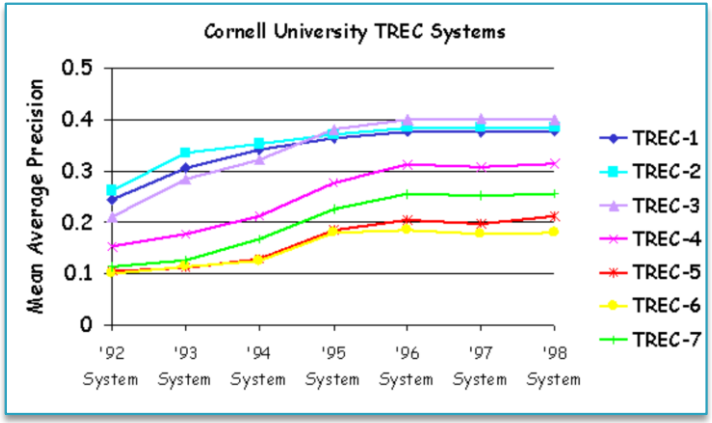
  - SMART (Buckley, Salton, Allan). *Retrieval with locality infor*
    - Local and global matching. Conducted 30 experiments!
  - Okapi (Robertson, Walker et al.). *Okapi at TREC*.
    - Probabilistic best matched system designed for interactive retrieval. F4 probabilistic global weight. (BM25 debut two years later.)
  - Berkeley (Cooper, Gey, Chen). *Staged logistic retrieval*.
    - Early "machine learned" ranking algorithm. 6 term frequency features.

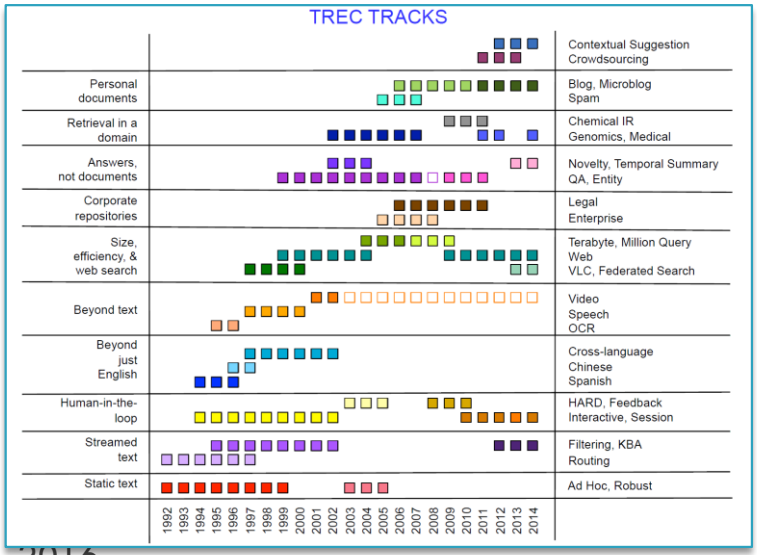


  - Bellcore & HNC – *Reduced dimensional representations*
    - LSI linear algebra; MatchPlus "neural" model

TREC Over the Years

- Participation remain strong
- Other forums started – CLEF, NTCIR, FIRE …
- Systems improve
- Tasks/tracks evolve





Benefits of TREC

- Provides rigor in evaluating search
  - New evaluation methodologies and metrics
  - Spawned other evaluation forums (CLEF, NTICR, FIRE)
- Develops shared (reusable) test collection
  - NIST evaluation for many programs (TIPSTER, MUC, MEMEX)
  - Incubated new search challenges
- Shapes research and practice in search
  - Research and publications
  - Practice (e.g., InQuery Infoseek, BM25 Bing, Watson IBM, legal, use of evaluation methods and hiring IR people)

TREC and Search Research

- TREC on the Web
  - *TREC retrieval* (274k); *TREC SIGIR* (235k), *wt10g* (142k)
- Use of TREC Corpora at SIGIR

- 1993 (37 papers): first TREC papers
  - *Overview of the First Text REtrieval Conference (D. Harman)*
  - TREC (4+2); CACM (5); others Medline, news, ency (15)

- 1998 (39 papers):
  - TREC (19+2); CACM (1); others (11)

- 2003 (46 papers):
  - TREC (23+5); others (16)

For the purposes of our experiments, … two very important but hard-to-find features: somewhat lengthy full-length texts and pre-determined relevance judgments for a set of queries.

Limitations of TREC

- Researchers/reviewers/funding agencies look at where the light (i.e., data) is
- Not clear what space of queries, documents and tasks we are sampling from
  - Sometimes lags search industry/practice
  - Scaled in number of documents, but not in queries
- Limited focus on end-to-end search tasks and search user experience
  - Gap between offline metrics and online experiences

Looking Where the Data Is

- Shared data sets and evaluation methods
  - Important for progress of IR
    - They are abstractions; not always applicable
  - "Streetlight effect" creates an observational bias
    - Illuminates only a small portion of the IR world
    - Supports some kinds of research, but not others
- Rapidly changing information landscape
  - New applications require new models, algorithms, etc.
    - E.g., Web @ TREC; Surprises in early web search

Reproducibility and Generalization

- Reproducibility
  - "Data": Shared queries-documents-relevanceJudgments
  - "Methods": Careful description of algs and methods
    - Rifkin & Klautau, JMLR'04, "In defense of one-vs-all classification"
- Generalizability
  - New queries … what space are we sampling over?
  - Variation in queries … coverage limited with small N
  - New collections/tasks … again, what's the space?
  - In practice these differences are often bigger than algo diffs
- Opportunity for TREC to help generalizability

Search Over the Years

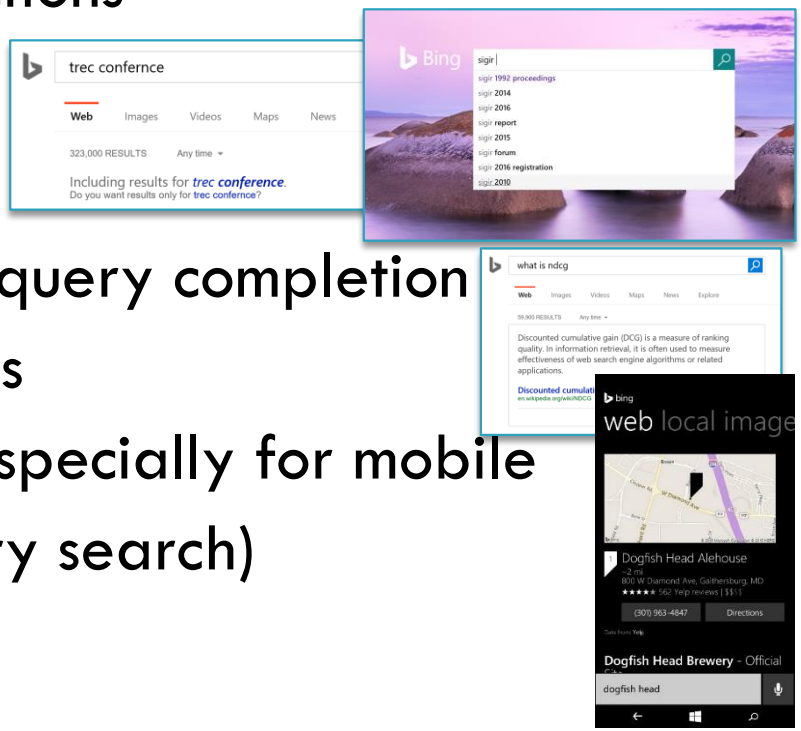
- Today search is everywhere
  - A billion web sites
  - Trillions of pages indexed by search engines
  - Billions of web searches and clicks per day
- Search is a core fabric of everyday life
  - Diversity of tasks and searchers
  - Pervasive (web, desktop, enterprise, apps, mobile, etc.)
- More important now than ever

How Did We Get Here?

- Early web search systems
  - Content + Links + Behavior (anchor text, queries, clicks)
- Surprises in early web search
  - Queries were short
  - Navigation was common
  - Queries were not independent
  - Amazing diversity of information needs (“long tail”)
  - Adversaries are prevalent
- Ongoing innovations in algorithms and UX

How Did We Get Here? (cont'd)

- New algorithms and content
  - Content: images, videos, news, maps, shopping, books
  - Entities and knowledge graphs
  - Machine learned ranking functions
  - Contextualization
- Enhanced UX capabilites
  - Spelling correction, real-time query completion
  - Inline answers and entity cards
  - Spoken queries and dialog, especially for mobile
  - Proactive notifications (0-query search)



What's Next in Search?

- Web search does very well at some things, but miserably at others
- In many other settings, search is much worse

- To make continued progress, we need to:
  - Understand entities and relations (from "strings" to "things")
  - Represent and leverage context
  - Understand dynamic environments in which docs, queries, and relevance change over time
  - Go beyond ranking to also encompass query articulation, results presentation, organization, and summarization

What's Next in Search?

The difficulty seems to be, not so much that we publish unduly in view of the extent and variety of present day interests, but rather that ***publication has been extended far beyond our present ability to make real use of the record.*** *The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships.*

Vannevar Bush, 1945

Summary

- Search has improved dramatically in the last 25 years
  - TREC evaluation methods, data sets, and community are an important part of that
- But there's still a long way to go
- Search is more important now than ever

Thanks!

- Questions?
- More info:
  http://research.microsoft.com/~sdumais