Paul Haahr

Senior Staff Software Engineer and Manager, Search Quality
haahr@google.com

I joined Google in March, 2002. (I was on paternity leave from November, 2005 through March, 2006.)

Accomplishments & Responsibilities

Clicks Team (April, 2006 - *present*)

I'm the manager for logs-based ranking projects. The team's efforts are currently split among four areas:

- Navboost. This is already one of Google's strongest ranking signals. Current work is on automation in building new navboost data; recency and anti-grandfathering; and building an improved model of user behavior, taking into account impressions and display bias.
- Generalizing navboost: extending the system to help with (query, document) pairs for which we have no direct evidence.
- Mergeserver and navseek: These are really two sides of the same coin. The former is an experimental framework for comparing user behavior in the presence of changes to ranking or UI; the latter, a system for exploring new results for existing queries.
- Logging improvements for quality purposes.

Mustang Scoring (June, 2004 - *present*)

I was tech lead for the development of AScorer, the ranking code used by websearch in Mustang, and continue to lead maintenance of that code base. Recent (spring/summer 2006) work has focused on reliability and performance as well as the addition of support for Google Co-op and merged web/image search. We are now launching a project to extract a modular scoring library from AScorer that can be customized for arbitrary corpora.

AScorer was initially written by Yonatan Zunger and implemented scoring semantics designed by Amit Singhal. During the initial phase, I code reviewed all changes and helped design the structure. This included several rewrites, as we learned what would and would not work efficiently within Mustang. While we were launching Mustang, Amit and I started working directly on AScorer: Amit made structural and scoring improvements while I worked on performance issues and fixing bugs or divergences from existing Google.com behavior.

Engineering Badurls Lead (December 2002 - *present*)

Liason with the legal team and customer support for `badurl` removals, including DMCA requests. Our team handles all removals from web and image search that cannot (or should not) be handled by customer support. 90% of the job is drudgery which should be more automated, but 10% requires active involvement, investigation, or arguing with our lawyers; distinguishing the 90% from the 10% is sometimes hard.

Quality Bugs/Triage/Tools/Small Projects Team (May - November, 2005)

Managed a changing group of roughly half-a-dozen engineers within search quality. This project was chartered by me and Urs with two goals in mind: first, fix longstanding websearch problems that hadn't been getting sufficient attention and, second, provide a

good place for training new search quality engineers. The project appears to be a significant success on both fronts.

Officially, I ran the project until April, 2006, but, in practice, it was managed by Scott Huffman, with Steve Young serving as tech lead, during the time I went on paternity leave. I've been acting in a consulting role for it since then.

Under the umbrella of this team, I managed a few longer-running projects:

- Sitebreaker - Levent Ertoz. Levent took over development of the sitebreaker from me in spring 2005. He developed a new, much-improved approach for identifying how to break domains into sites and has shipped new sitechunker patterns for the spam team based on it. This has a major effect on our ability to deal with spam on freehosts.
- Rephil in scoring - Taku Kudo and Levent Ertoz. Attempted to improve ranking using Rephil as a replacement for the current using of Phil in scoring, improving several known deficiencies of the current implementation, notably the partial coverage of Phil over documents. Unfortunately, we were not able to find significant quality gains. Our hypothesis is that the fine-grained version of topicality matching offered by synonymy has similar wins but fewer losses than the gross topicality matching that Phil/Rephil provides; we suspect that the synonyms project has obviated this approach to using Phil in scoring.
- Capitalization-influenced ranking - Jim Meehan. This project moved with Jim into my group from Anna Patterson's in summer 2005. The code worked correctly but had little impact and had significant performance costs due differently capitalized versions of queries not being able to share cache entries.

Vertical Ranking: Travel (January, 2005 - October, 2005)

I was tech lead for a project Amit Singhal and I created to improve rankings for queries in vertical domains. Our first vertical was travel-related queries; the project consisted of me, Steve Young, Chuck Rosenberg, and Kevin Lacker.

The outcome of the project was mixed. We did not end up launching anything directly. However, our identification of issues with navigational queries including place names and several techniques pioneered by the group — navsmearing, geographic term smearing, A/B scoring — were picked up by the local navigational queries project, which had (mostly) in Q2 of 2006.

The only substantial code I personally wrote for this project was a Phil-based classifier for travel queries. The classifier was very successful at identifying queries which were travel-related (for example, knowing that [Paris Marriott] and [San Francisco] are likely about travel but [Paris Hilton] and [Baghdad] are not), but we failed to find significant techniques for improving non-navigational travel queries once we identified them.

Mustang GT Launch (June, 2004 - March, 2005)

Mustang GT is a reimplementation of Google's entire web-serving system on top of the Mustang framework. As of Spring 2005, Mustang GT was handling all Google.com traffic.

I joined Mustang GT a few months after the project had started, to help get the system into production. During the launch process, I acted as a coordinator for quality and evaluation issues.

Sitebreaker (March - May, 2004; December, 2004)

Develeped the Sitebreaker, an automated system for indentifying domains used for web-hosting and the boundaries between different sites within the domains. The sitebreaker accompanies the "sitechunker," which applies the rules discovered by the sitebreaker. The sitechunker, which was developed for Anchor++ and Independence Rank, is now used widely within quality when developing signals that operate at a coarser grain than individual web pages.

Quality War Room (March, 2004 - May, 2004)

I joined a group of quality engineers (plus Jeff and Sanjay) camping out in Larry and Sergey's office for a few months as we tried to make a rapid improvement in quality. (Lots of good came out of the war room, but it wasn't rapid in the end.)

Most of my work in the war-room was involved with integrating new features into the segment indexer (including page and link classifiers, anchor contexts, and pagerank variants), code reviewing, and diagnosing changes between production segindexer and the war-room builds.

During this time, I also moved the segment indexer and our pre-Trawler crawling code from google2 to google3 to ensure that indexing would build in google3.

Domain and Anchor++ Demotion (January, 2004)

Added a *twiddler* to gws which demotes later-occurring results from a given domain or anchor++ cluster. This promotes diversity in our results and gives a substantial improvement to some of our most-spammed queries.

Segment Indexer (September, 2003 - February, 2004)

Part of the team that moved the segment indexer from the development version written by Jeff and Sanjay into production. My involvement included:

- Generating LCA (Local Context Analysis) data from the composite documents.
- integration of independence rank.
- Improvements to SSTable and MapReduce, including splitting of SSTables such that multiple values for the same key are processed in the same map piece, which was necessary for correct operation of the segment indexer.
- Diagnosis and debugging of quality differences between the segment indexer and the old, BART-based indexing system. This included adding audit trails to some segindexer data structures, building tools to examine indices, helping other people build such tools, writing unit tests, looking at results and composited documents, and was probably the bulk of my time on the project.
- Established the *segment indexer live by Halloween* goal, which we did not meet. (Two major holidays later, it was up and running.)

URL→queries data (March - September, 2003)

This was my 20% project and an intern's summer project. Developed the idea of using our result logs (a sample of queries including result URLs and scoring) to build maps from URLs to queries, for use as document summaries.

Query refinement suggestions derived from this data (work with Steve Baker) were tested as a 1% UI experiment on Google.com for English-language users and follow-up version was used in the Chameleon project. As part of this effort, I:

- built tools for gathering, cleaning, and processing the data.
- came up with the method of using the URL→Queries data for clustering and query refinement suggestions.

- wrote a design doc for the refinement server.
- worked with Steve to design the refinement suggestion algorithm.
- implemented the refinement server, support for query refinements in gws, and oversaw the launch.
- wrote an invention disclosure which was the basis for a patent filing (System and Method for Providing Search Query Refinements).
- wrote logs analysis programs (in sawzall) to evaluate the effects on users of the experiment.

The data is also being used in the development of "more-like-this/less-like-this" relevance feedback (work by Radhika Malpani) and result clustering (work by Mehran Sahami, Sugato Basu, and Alex Mizil). Potential future applications of this data include content ad targeting and identifying related pages.

Link-based quality (December, 2002 - February, 2004)

Developed, with Amit Singhal, "Independence Rank," a query-independent ranking based of sites based on the link graph, which is intended to be more difficult to manipulate than PageRank. Developments included:

- the underlying network flow function.
- programs for quickly computing and visualizing the ranking from a link graph on a single machine.
- a simple system for breaking domains into sites.
- quality evaluation of the effects of adding Independence Rank and Anchor++ to scoring.
- integration with production, including per-doc data for the index server (ongoing).
- one invention disclosure on the ranking function.

Work done in coordination with Martin Kaszkiel (project lead; developer of Anchor++ and the site clustering used by Independence Rank) and Matt Cutts.

Mergeserver (June, 2002 - September, 2003)

Technical lead. The mergeserver was for a long time Google's sole automated quality evaluation tool for scoring and ranking changes. My contributions included:

- **figuring out how to integrate the merged results idea into Google's architecture.**
- **building the mergeserver as a derivative of the mixer.**
- **designing the original and current experiment frameworks.**
- **writing the mergeserver design document.**

Original design and implementation with Radhika Malpani, Mizuki McGrath, and Simon Tong. Significant redesign and refactoring work with Corin Anderson.

Bits & Pieces

(Not all of these are accomplishments.)

- A bunch of small contributions to webspam over the years. In 2005, the biggest of these was "scraperscraper," a tool which finds spam sites constructed by scraping search engine result pages.
- Wrote invention disclosure for Method for distinguishing meaninful stop words in keyword-based retrieval systems and guided Steve Baker's implementation of the idea.

- Code reviewed Craig Silverstein's query parser (Spring 2003) and Robert Griesemer and Rob Pike's sawzall language implementation (Summer 2003).
- Introduced a declarative registration system for ripper passes. (March 2003.)
- Wrote a parser for ASCII-format protocol buffers. (January - March 2002, with Corin Anderson.)
- Google2 Files for reading and writing gzip-encoded files. (January 2003.)
- Tools for manipulating DNS zone files, including extracting domain-ownership changes from and providing lookups from nameserver to domain names. (December 2002.)
- Investigated using preparsed repositories in the docserver and wrote an abandoned design doc for the idea. (April - July 2002, with Steve Glassman.)
- Dabbled with the global work queue design and implementation, after Jeff Dean started on it, but handed it over to Percy Liang who made it work. (April - May 2002.)

Other Work Activities

Google3/Components committee (April, 2003 - Summer, 2004)

Helped design the components model for Google's new development and release environment (i.e., `google3`). Did the first teasing apart of `google3/base` from the rest of the `google2` tree. Ported a few hundred thousand lines of code from `google2` to `google3`. Act as an "approver of last resort," by virtue of being in the `google3/OWNERS` file.

Recruiting

I've served on a hiring committee since September, 2004, as well as on the hiring intergrouplet and an "Eng 2100" working group run by business operations. I gave recruiting talks at Georgia Tech and Rice University in September 2004. Plus, the same demanding pace of interviews as other engineers carry.

Mentoring & Training

I gave the "Welcome to Google Engineering" for interns every week during summer 2003. I mentored Paul Tucker, Mike Dixon, Steve Baker (when he was an intern), Misha Dynin (on his transfer to the quality group), and Steve Young. I gave a Google 101 talk for all of Google Santa Monica when they first visited Mountain View. I gave a couple of tech talks in the NY office to explain what happened in the quality war room and how to read `deb=` output on Google.com. In summer 2005, I lead a quality "fixit day" on query debugging, which included a updated talk on how to diagnose queries.

Public relations

Participated in "meet the engineers" events at WebmasterWorld's New Orleans PubConference and Search Engine Strategies San Jose. Gave a talk on Google technology at Adobe in December 2004.

New Hires faction (May - September 2002)

Wrote, with Will Nevitt, our first New Engineer Checklist.

Employee referrals

Frederick Roeber, Adam Dingle, Stan Chesnutt, Evelyn Gee. (And a host of others who didn't make it.)

Employment History

Independent consultant

Software development at Network Appliance (1996-1997) and ~~Agile Storage ClariStor~~ OnStor, Inc. (2001-2002).

Xigo, Inc. (2000-2001)

Software engineer and engineering manager.

Cashcade.com (1999)

Founder of a financial services startup that ended without accomplishment.

Jive Technology (1997-1999)

Co-founder of a Java bytecode compiler startup that should have stayed a consulting business.

Harlequin Ltd. (1994-1996)

Software engineer, Dylan compiler group.

Kaleida Labs, Inc. (1993-1994)

Software engineer, ScriptX team.

Adobe Systems Incorporated (1990-1993)

Software engineer, type group.

Polygen Corporation (1987)

System administrator and software engineer.

Oracle Corporation (1986-1987)

System administrator and software engineer, porting group.

(For details, see my external resume.)

Education

Princeton University (September, 1984 - January, 1986; September, 1987 - June, 1990)

Bachelor of Arts in Computer Science, Summa Cum Laude.
Elected member of Sigma Xi engineering honor society.
Won Sigma Xi book prize for outstanding independent research.
Served as a teaching assistant for CS217 (Programming Systems) and CS320 (Compiler Design) and as a staff member in the system administration group.

Google connections: took an operating systems class from Rob Pike; was a teaching assistant when Karl Pfleger took the sophomore programming systems class; my junior project was an implementation of Rob's NewSqueak language for which Sean Dorward wrote an RPC stub generator.

Published Work

Kim Barrett, Bob Cassels, Paul Haahr, David A. Moon, Keith Playford, and P. Tucker Withington, *A Monotonic Superclass Linearization for Dylan*, OOPSLA 1996 Conference. This paper introduced the C3 Linearization, which is used by Python, among other languages, in the specification of multile inheritance.

Paul Haahr and Byron Rakitzis, *Es: A shell with higher-order functions*, Proceedings of the Winter 1993 Usenix Technical Conference.

Paul Haahr, *Montage: Breaking Windows into Small Pieces*, Proceedings of the Summer 1990 Usenix Technical Conference. Won best student paper prize.