

An old-timer's perspective on the Web Track(s)

David Hawking
dahawkin@microsoft.com

My perspectives on the Web track:

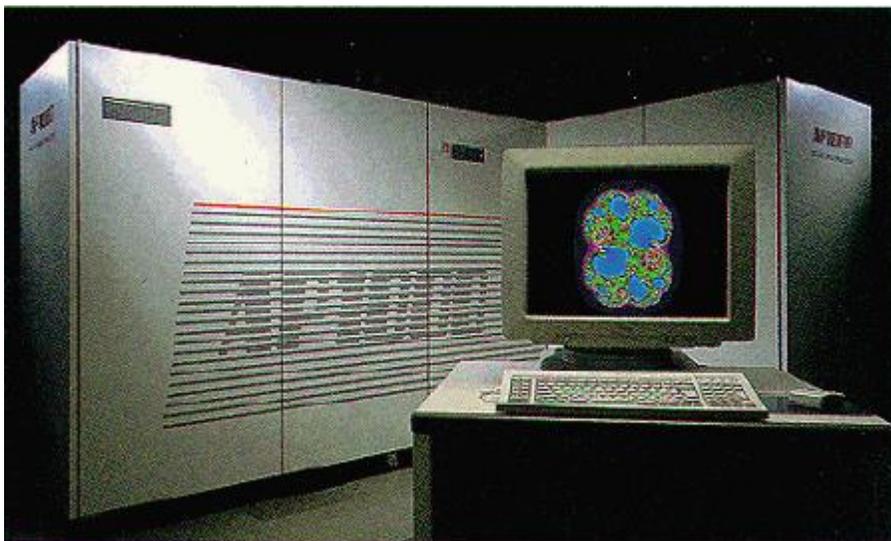
- TREC participant (10 years)
- Track co-ordinator / PC member (9 years)
- Small scale commercial enterprise and web search (14 years: CSIRO, Funnelback)
- Web search engine (3 years: Bing)

TREC-3 (1994) was my very first visit to the USA.

I expected something like the Grand Canyon ...



“The policy of this hotel is one of *aggressive friendliness*”.



PADRE: The PARallel Document Retrieval Engine
My TREC-3 Retrieval system ran on a 512-node
supercomputer, worth about \$12M!

Don't worry about inverted files; Stick it all in a corner of
memory and grep it.

Me and TREC-3



- System and network admin.
- Breakfast in Gaithersburg
- A solution in search of a problem.
e.g. reg. expressions
- The most naive attendee ever?
- TREC-3 highlights
 - BM25
 - Gerry Salton
 - Fulcrum
 - The “Sponsors”
- I drank in every talk
- On arrival I didn’t know a single attendee. On departure ...
- Donna, Stephen, Karen, Ellen, Chris, Amit, Bruce, Jamie, James, Sue, Alan, Charlie, Gord, Nick, Ross, Peter, Jacques, K.L, Doug, ...

At TREC-3 I joined a community and started a new career!

VLC & Web Tracks (1996 --)

(The late Paul Thistlewaite and Nick Craswell led the move to web)

- At TREC-3 I argued for ramping up the data set.
 - My motives were impure
 - But the end justifies the means!
- Obtaining data
 - Mark: Glasgow Herald / FT
 - Gord: USENET news
 - Aust/Can. govts, Universities
 - Brewster Kahle
- Obtaining judgments
 - Focus on early precision
 - Judging order
 - Concept based judging tool (the RAT)
- We tried to use TREC ad hoc methods and standard judging methods for web.
- That was a mistake
- But we learned more from that mistake than from any ten others 😊
- We applied TREC methods to evaluating Web search engines
- Infonortics Search Engines Meeting 2000

<http://newsbreaks.infotoday.com/NewsBreaks/Old-Economy-Info-Retrieval-Clashes-with-New-Economy-Web-Upstarts-at-the-Fifth-Annual-Search-Engine-Conference-17819.asp>

Old to New: Play by the Rules!

The "old economy" faction consisted primarily of participants in the Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA). In his presentation "Secrets of TREC," Chris Buckley from SabIR Research said that TREC's purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.

Each year, TREC focuses on several different problems of text retrieval. NIST establishes the testbed, providing a standardized set of documents and questions. Participants run their own retrieval systems on the data, and return to NIST a list of the retrieved top-ranked documents. NIST pools the individual results, judges the retrieved documents for correctness, and evaluates the results.

Last year, 66 groups representing industry, academia, and governments in 16 countries participated. *"The benefits of TREC are the blind, independent evaluation, and TREC allows scientific evaluation of which techniques work best," said David Hawking, SIRO Mathematical and Information Sciences, Australia.*

However, of the 66 groups participating, "the search engine companies didn't play," said Hawking. "We tried to entice them to come along with TREC. They didn't. So we evaluated them anyway," he said. The results weren't pretty. With the exception of Northern Light and Google, major search engines TREC evaluated fared relatively poorly compared to others participating in the test.

Nonetheless, representatives from the major engines did not appear chastened by the results of the TREC evaluation.

<http://newsbreaks.infotoday.com/NewsBreaks/Old-Economy-Info-Retrieval-Clashes-with-New-Economy-Web-Upstarts-at-the-Fifth-Annual-Search-Engine-Conference-17819.asp>

New to Old: The Rules are Irrelevant!

On a panel hosted by TREC advocate David Evans (Claritech), search engine representatives spoke about improvements they've made to their services, and plans for the future. Evans also asked panelists to define relevance, and describe the metrics used to determine that a specific approach was working.

Citing studies of searcher behavior, Evans said, "We know that if people are willing to spend large amounts of time, eventually they will start to converge on perfect performance." Practice and persistence pay off—research shows that if a searcher spends just an hour working with an engine, performance improves dramatically. But most Web search engine users demand nearly instant results.

"Search engines are clearly caught in a corner because of the pragmatics of their enterprise. It's a difficult problem; we acknowledge that," said Evans.

The panelists' responses were forthright and pulled no punches, acknowledging that many design and implementation decisions were made quickly, often in response to user demand. "At the end of the day, it was seat of the pants," said Jan Pedersen, formerly head of search and directory for Go/Infoseek. "People had a hunch and it was implemented."

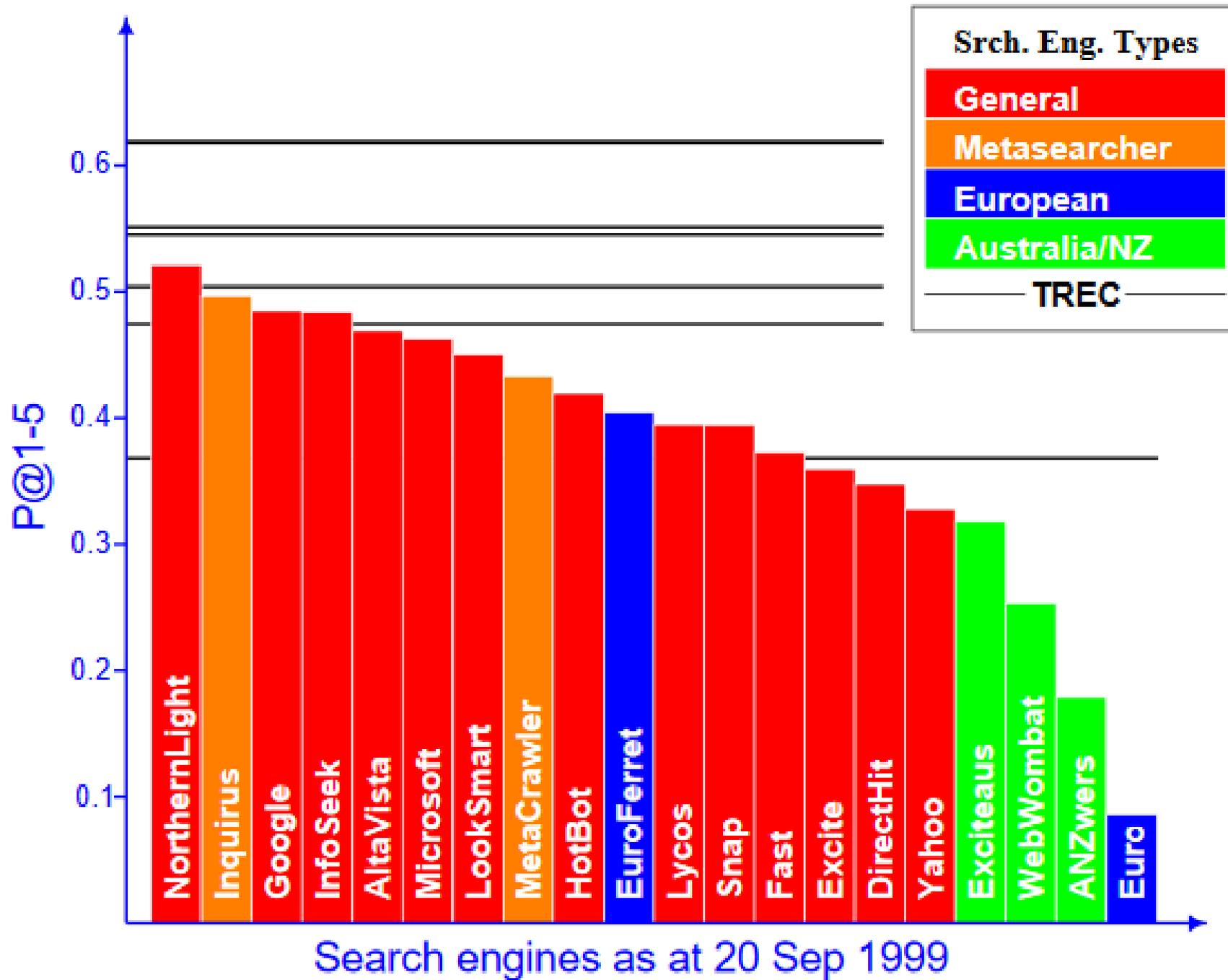
It gradually became clear why, despite diplomatic overtures, the major Web search engines didn't participate in the TREC evaluations. "We're constantly surprised by good ideas that don't help," said Marc Krellenstein of Northern Light.

Google CEO Larry Page had the most heated response to the TREC advocates, at one point calling the entire formal evaluation process "irrelevant." "I don't believe that binary relevance rankings are useful," said Page. He's convinced that surviving and thriving in the crucible of the Web is sufficient measure of success." "All of us could think of things to do that would make things better if you gave us infinite resources," he said.

Indeed, the TREC testing process seems akin to standing in the midst of a stampeding herd of elephants, taking a snapshot, and trying to draw meaningful conclusions about the veracity of the herd. The camera might be top-notch, the photographer first-rate, and the interpreters brilliant. Nonetheless, by the time conclusions are reached the herd will have changed course numerous times, individuals in the herd will have grown stronger or weaker, and even the environment through which the herd is stampeding will have changed dramatically.

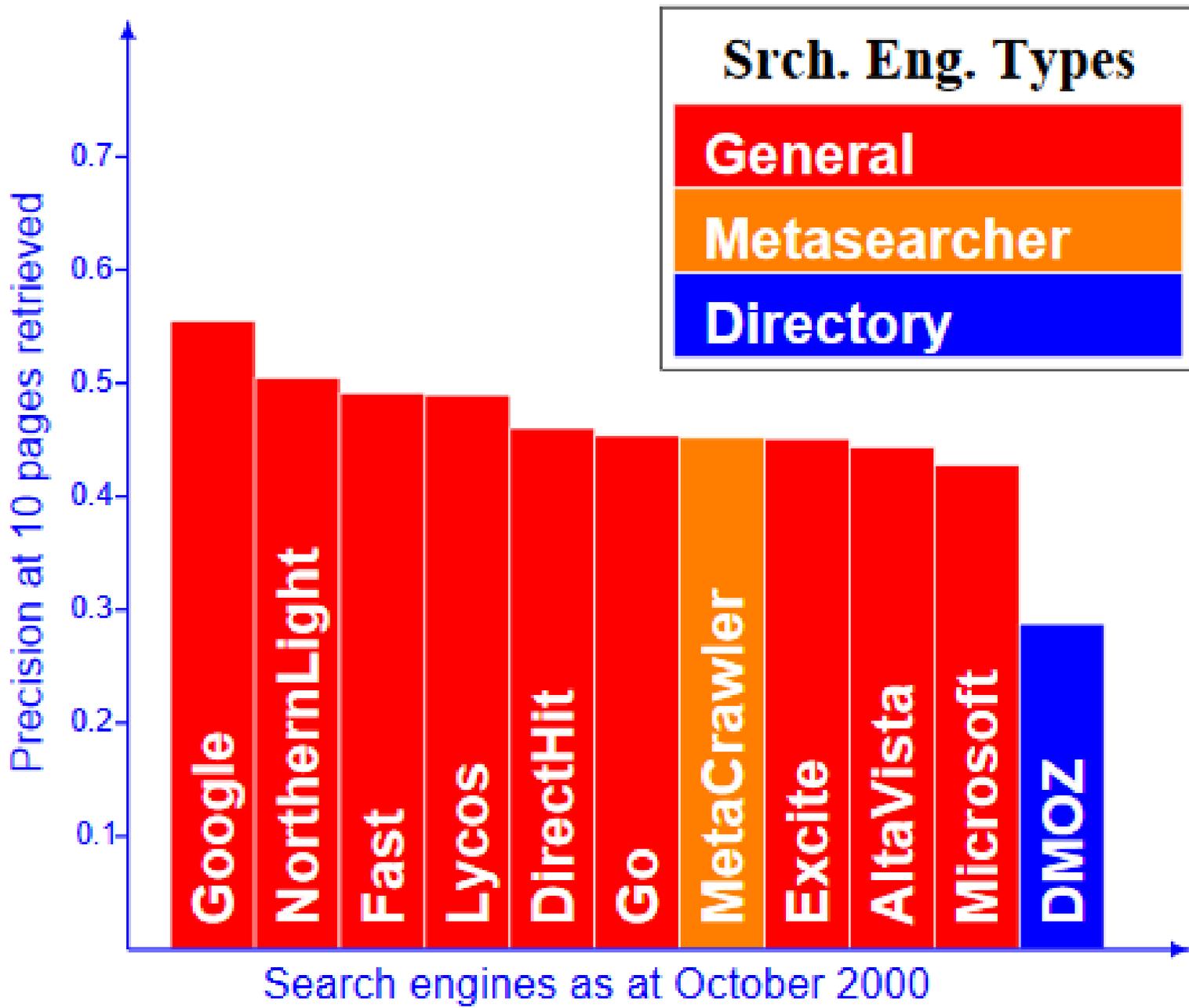
On balance, both the "old" and "new" economy participants made valid, thought-provoking observations, and scored meaningful points against their counterparts. The charged dialog only underscored both factions' passion and commitment to providing the best possible results for searchers.

My conclusion: We were absolutely right about the importance of principled evaluation but they were right in criticising our method of web evaluation.



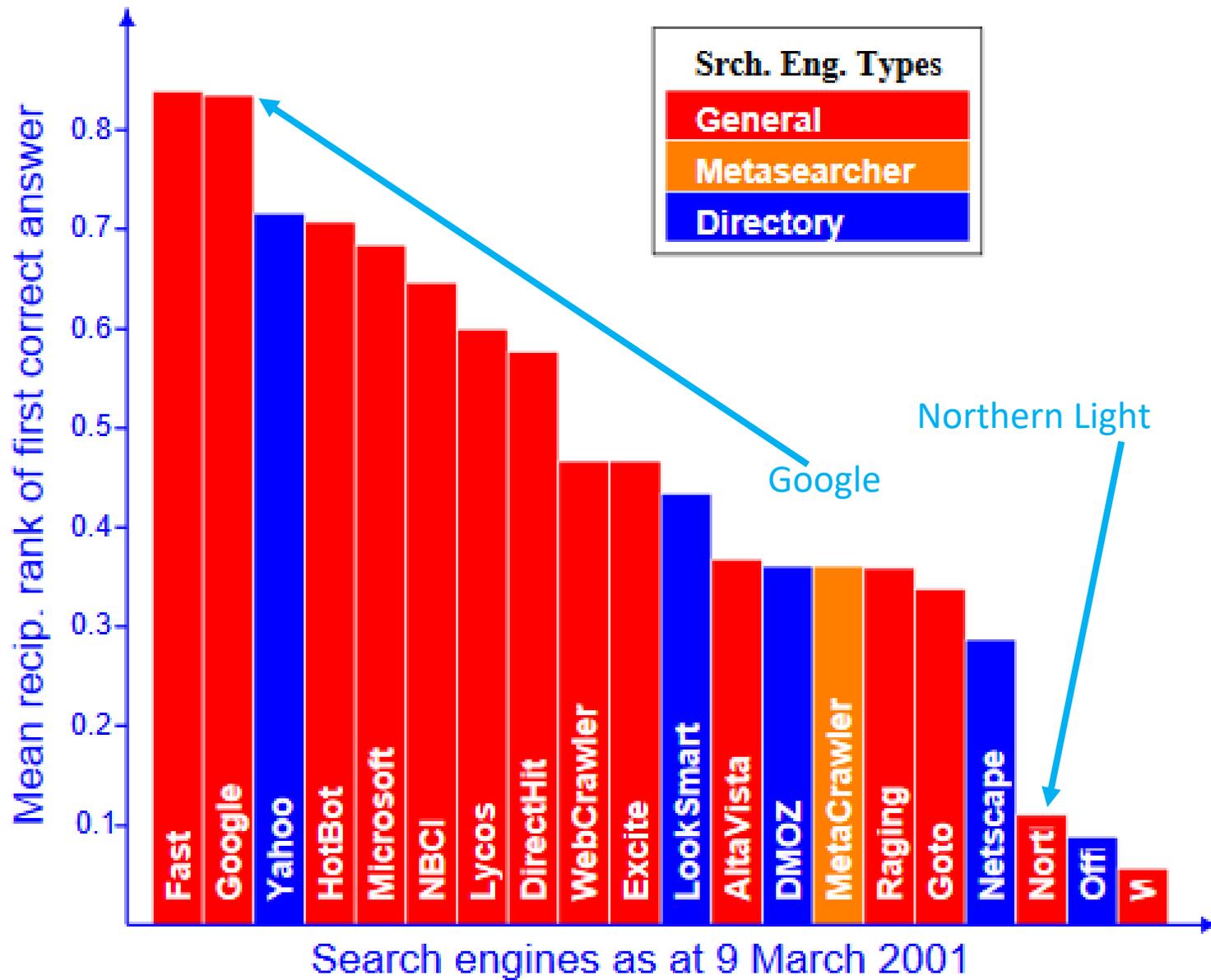
TREC ad hoc judging

Comparison using 54 queries selected from search engine logs, whose intent was judged to be clear. E.g. *“Find information about American anarchists”*



Online service finding

E.g. "Where can I buy flowers online"



Finding airline
homepages

E.g. "United Airlines" →
<http://www.united.com/>

Figure 1: Mean reciprocal rank achieved by the engines.

Table 1: Summary of VLC and Web Track evaluations 1996 - 2003.

Year	No.	Track/Task	Coll.	Topics	No. Partic.
1996	TREC-5	Pre-VLC	CDs 1-4	251-300 (From Ad Hoc)	4
1997	TREC-6	VLC	VLC	301-350 (From Ad Hoc)	7
1998	TREC-7	VLC	VLC2	351-400 (From Ad Hoc)	7
1999	TREC-8	Large Web	VLC2	50/10000 (From SE NLQ logs)	8
		Small Web	WT2g	401-450 (Joint w. Ad Hoc)	17
2000	TREC-9	Large Web (Online Srvcs)	VLC2	50/10000 (From SE NLQ logs)	5
		Main Web	WT10g	451-500 (Rev. Eng. SE)	19
2001	TREC-2001	Web Topic Relevance	WT10g	501-550 (Rev. Eng. SE)	29
		Homepage Finding	WT10g	EP1-145 (Random target selection)	16
2002	TREC-2002	Topic Distillation	.GOV	551-600 (NIST engineered)	17
		Named Page Finding	.GOV	NP1-150 (NIST engineered)	18
2003	TREC-2003	Topic Distillation	.GOV	TD1-TD50 (NIST engineered)	23
		Mixed Named/Homepage	.GOV	NP151-450 (NIST engineered)	19
		Interactive (Topic Dist.)	.GOV		2

From the TREC Book, 2004

Bob Travis & Andrei Broder: "Web search quality v. informational relevance". Infonortics SEM, April 2001

http://web.archive.org/web/20030930170629/http://www.infonortics.com/searchengines/sh01/slides-01/travis_files/v3_document.htm

Year	No.	Track/Task	Coll.	Topics	No. Partic.
2004	TREC-13	Web: Mixed TD, NP, HP	.GOV	225	18
		Terabyte: Ad Hoc	.GOV2	50 (NIST created)	17
2005	TREC-14	Enterprise: known item	W3C	125	15
		Enterprise: discussion	W3C	59 (participant created)	14
		Enterprise: expertise	W3C	? (participant created)	9
2005		Terabyte: Ad Hoc	.GOV2	50 (NIST created)	18
		Terabyte: Efficiency	.GOV2	50,000 (SE logs)	13
		Terabyte: Named page	.GOV2	252 (participant created)	13
2006	TREC-15	Enterprise: discussion	W3C	59 (NIST created)	10
		Enterprise: expertise	W3C	? (participant created)	23
		Terabyte: Ad Hoc	.GOV2	50 (NIST created)	20
		Terabyte: Efficiency	.GOV2	50,000 (SE logs)	8
		Terabyte: Named page	.GOV2	12 x 11 (participant created)	11
		Blog: opinion finding	BLOG06	50 (NIST from SE logs)	10
		Blog: open	BLOG06	50 (NIST from SE logs)	10
2007	TREC-16	Enterprise: overview	CERC	50 (created by 9 communicators)	16
		Enterprise: expertise	CERC	ditto	15
		Million Query:	.GOV2	10,000 (SE logs)	10
		Blog: opinion finding	BLOG06	50 (NIST from SE logs)	20
		Blog: opinion polarity	BLOG06	50 (NIST from SE logs)	11
		Blog: distillation	BLOG06	50 (NIST from SE logs)	9
2008	TREC-17	Enterprise: overview	CERC	77 (actual science enquiries)	14
		Enterprise: expertise	CERC	ditto	11
		Million Query:	.GOV2	40,000 (SE logs)	9?
2009	TREC-18	Web: Ad Hoc	ClueWeb09	50 (NIST, from SE Q clusters)	25
		Web: Diversity	ClueWeb09	ditto	18
		Million Query:	.GOV2	10,000 (SE logs)	9?
2010	TREC-19	Web: Ad Hoc	ClueWeb09	50 (NIST, from SE Q clusters)	< 23
		Web: Diversity	ClueWeb09	ditto	< 23
		Web: Spam	ClueWeb09	ditto	3
		Session:	ClueWeb09	150 query, reformulation	10

Web and efficiency
evaluations 2004 - 2010

Year	No.	Track/Task	Coll.	Topics	No. Partic.
2011	TREC-20	Web: Ad Hoc	ClueWeb09	50 (NIST, from SE Q clusters)	<= 16
		Web: Diversity	ClueWeb09	ditto	<= 16
		Crowdsourcing:	?	?	?
		Session:	ClueWeb09	76 sessions	13
		Microblog: real time ad hoc	Tweets2011	(NIST) 50?	59
2012	TREC-21	Web: Ad Hoc	ClueWeb09	50 (NIST, from SE Q clusters)	<= 12
		Web: Diversity	ClueWeb09	ditto	<= 12
		Crowdsourcing: TRAT	TREC-8	TREC-8	7
		Crowdsourcing: IRAT	ImageCLEF	90 topics, 20k images	2
		Session:	ClueWeb09	98 sessions	10
		Microblog: real time ad hoc	Tweets2011	(NIST)	<= 40
		Microblog: tweet filtering	Tweets2011	(NIST)	<= 40
Contextual suggestion:	ClueWeb12 or Web	34 profiles	14		
2013	TREC-22	Web: Ad Hoc	ClueWeb12	50 (NIST, from SE Q clusters)	<= 15
		Web: Risk-sensitive	ClueWeb12	ditto	<= 15
		Crowdsourcing: ad hoc	ClueWeb12	ditto	4
		Session:	ClueWeb12	87 sessions + 46 training	13
		Microblog: RT ad hoc	Tweet service	(NIST) 50?	20
		Contextual sugg.:	ClueWeb12 or Web	562 profiles, 50 contexts	19
		Federated web: resource sel.	FedWeb 2013	200 (organizers)	9
		Federated web: result merging	FedWeb 2013	200 (organizers)	6
2014	TREC-23	Web: Ad Hoc	ClueWeb12	50 (NIST, from SE Q clusters)	9
		Web: Risk-sensitive	ClueWeb12	ditto	5
		Session:	ClueWeb12	1257 sessions (AMT)	11
		Microblog: temporal ad hoc	Tweet service	(NIST) 55	21
		Microblog: tweet timeline	Tweet service	(NIST) 55	13
		Contextual sugg.:	ClueWeb12 or Web	299 profiles, 50 contexts	17
		Federated web: resource sel.	FedWeb 2014	60 (organizers)	6
		Federated web: result merging	FedWeb 2014	60 (organizers)	6
		Federated web: vertical sel.	FedWeb 2014	60 (organizers)	7
2015	TREC-24	Microblog: push notification	Tweet service	(NIST) 51	14
		Microblog: email digest	Tweet service	(NIST) 51	16
		Contextual sugg.: Live	merged POI crawls	?	6
		Contextual sugg.: Batch	merged POI crawls	?	17
		Live QA:	up to participants	1087 unanswered Yahoo answers	14
		Tasks: Understanding	ClueWeb12	50 queries, Freebase IDs	5
		Tasks: Completion	ClueWeb12	50 queries, Freebase IDs	3
		Tasks: Ad Hoc	ClueWeb12	50 queries, Freebase IDs	2
2016	TREC-25	?	?	?	

Web and efficiency
evaluations 2011 - 2016

- They didn't understand/trust the measures
- They wanted to see the difference on **their** data
- Unless relevance ranking is **really bad**, other factors are more important.

CSIRO / Funnelback customers never cared about TREC results!

“Our technology performs well on TREC. For details please see the TREC overview.”

What other factors?

- Can you index all our data – Documentum, Sharepoint, E-vault?
- Speed of index refresh
- Quality of extraction from binary formats
- Platform your engine runs on
- Will your company still be around in five years?
- Cost (sometimes).
- Analytics
- Document level security
- Manual control
- Look and feel – templating, responsive design
- Faceting
- Query suggestion
- Related searches
- Duplicate suppression, diversity
- Summary quality
- Character sets
- Geospatial search
- “Relevancy”

Many of these things are indeed important aids to helping users to locate and consume information or to access services

C-TEST

- Side-by-side is great for convincing and explaining, but no good for tuning
- Can't afford complete judgements, so find the topic distillation answers for a set of queries
 - Business critical
 - Random sample
 - Many organizations have resources to do this.
- Funnelback: Tuning for best out-of-the-box performance
 - Multiple data sets with different characteristics
- Customer: Tuning to optimize for their environment

C-TEST administrator interface

Success rate: **81.82%**

Search quality score: **73%**

Delete Row

Suggest Query

	Query	Correct answers	Score	Info
☰	events	http://www.rmit.edu.au/events	100%	
☰	research	http://www.rmit.edu.au/research/	100%	
☰	news	http://www.rmit.edu.au/news/	100%	
☰	maps	http://www.rmit.edu.au/maps/	100%	
☰	information retrieval	http://www.rmit.edu.au/about/our-education/academic-schools/computer-science-and-inf...	100%	
☰	contact	http://www.rmit.edu.au/contact/	100%	
☰	predestination	http://www.rmit.edu.au/events/all-events/special-events/2015/april/predestination/	0%	?
☰	privacy	http://www.rmit.edu.au/utilities/privacy/	50%	
☰	alumni	http://www.alumni.rmit.edu.au/NetCommunity	100%	?
☰	childcare	http://www1.rmit.edu.au/childcare	0%	?
☰	disability	http://www.rmit.edu.au/life-at-rmit/support-for-students/	50%	
☰				

Evaluation / Experimentation at Bing

- The idea of complete judgments!
- The idea of evaluating on 20 YO data!
- A web search engine has many many components, which all need tuning and evaluation
- Offline judging
- User interaction logs for ML
- Judging has to be done in context of location and personal history
- Main focus is on online evaluation - flighting.
- The experimental infrastructure is very complex – very large numbers of simultaneous experiments
- Highly sophisticated statistics and experimentation to avoid false conclusions and control interaction effects
- Ronny Kohavi is the guru (See talk at Dublin SIGIR)



Lessons learned from the Web Track and beyond

- Methodologically sound evaluation and comparison is very important
- TREC doesn't always get that right
- There are many different IR problems
- And correspondingly many "best approaches" and "appropriate evaluations" – hence the necessary proliferation of TREC tracks.
- At ANU, CSIRO, Funnelback, and Microsoft, I've found it reasonably easy to create test collections, including judgments
- Unfortunately most are non-sharable.
- TREC has helped and will continue to help understanding of how to do Web and Enterprise retrieval
- But it's extremely difficult to model commercial enterprise and web reality with a TREC-style test collection.

- **My thanks** to NIST and to all the people and organizations who made possible the VLC and Web tracks that I was involved with.
- **Congratulations and thanks** to those who've taken the Web Track and its heirs and successors on to great heights in the years since I “ran off the tracks”.