The evolution of IR (Part 2 – The Electronic Era)

1960 - 2010

David Hawking

IRF Conference, Vienna 31 May 2010
“One never notices what has been done; one can only see what remains to be done”

Marie Curie, 1894
(in a letter to her brother)
A golden era of information retrieval
Why not start at 1940?
Why start at all?

“Those who cannot remember the past are condemned to repeat it”

George Santayana, *the Life of Reason*, 1906
Sixties
• Punched card, paper tape, and keyboard
• Simultaneous read, compute and punch
• Large capacity core storage – up to 60,000 digits
• High internal processing speeds.
• Access time – 20 microseconds
• Multiplication (5 digits by 5 digits) - 4.96 msec.
• Solid state!!
Cranfield Experiments

e.g Cleverdon & Mills, 1963

- Test collection created from “In-situ” judgments in a real retrieval task
- 5-point relevance scale.
- 400 searchers surveyed
- 1500 papers fully judged.
Cleverdon '64 Evaluation Criteria

- Coverage
- Recall
- Precision
- Response time
- User effort
- Form of output

- Automatic indexing outperforms manually assigned terms. (Cranfield II report, 1966)
Gerard Salton

F. Wilfrid Lancaster
• The term “hypertext” coined and the Xanadu idea of universal interlinking (Ted Nelson, 1965)

• **The Mother of All Demos** (Doug Engelbart, 1968) – “NLS” – mouse, interactive text, video conferencing, email, hypertext and collaborative real-time editing.
Seventies
c. 1970 – Univac 1108

Storage prices (2009 USD Per gigabyte)

- RAM: $893M
- Drum: $65M

- Typical cost (1968 dollars): >$2M
- 131 kWords (36-bit) core memory: $823k
- 2 Mword magnetic drum, 92 msec ave latency: $96k
- Approximately 1 MIPS
Information Retrieval in 1971
- Relevance feedback (Rocchio, 1971)
- Bibliographic co-citation (Small, 1973)
- Relevance and utility (Cooper)
- IP protocols (Kahn & Cerf, 1973)
- Ethernet (Metcalf et al, 1975)
- Idf, Probabilistic ranking (Harter, Robertson, Sparck Jones, Maron)
- Vector space model (Salton et al)
- Porter stemmer (1979)
- Wordplex, Troff, Scribe, TeX
- Multi-processing operating systems.
- The WIMP interface (Alan Kay, 1972)
Eighties
• Deployment of Ethernet and laser printers
• SMTP (Jon Postel, 1982)
• The concept of the ideal machine for computer science: the 3M machine
• Personal computers
• WordStar, WordPerfect, MSWord
• WYSIMOLWYG DTP (Apple Mac)
• Unicode (Joe Becker, 1988)
• Archie (1989)
• World Wide Web (Tim Berners-Lee 1989)
Nineties
Remember 1991?

- Berners-Lee invented WWW two years ago
- Lycos web search was still 2 years off
- Harman and Candela – ranking over 1GB
- FreeWAIS took a week to index 1GB (Brewster Kahle)
- TREC-1 next year
- BM25 three years away
- ANU had the fastest computers in the world – outside countries starting with U or J!
c.1990 – Sun 3/80

- Up to 16MB RAM
- 104MB disk
- 25 MHz M68030
- c. $5k
1990s – What else?

• TREC (Donna Harman, 1992 – )
• UTF-8 (Ken Thompson, 1993)
• Mosaic browser (Andreessen, Bina, 1993)
• Lycos (Mauldin, 1993)
  – Alta Vista (Monier, Burrows …, 1995)
  – Infoseek (Kirsch, Li, Chang)
  – DirectHit (Culliss, Cassidy, 1998)
  – Google (Brin, Page, 1998)
• Query-driven advertising (Open Text, 1996)
• Effective e-commerce (Amazon, 1994)
• Recommender systems
Noughties
• The verb “to blog” coined (c. 2000)
• **Wikipedia (Wales/Sanger, 2001)**
• Flickr (Butterfield & Fake, 2004)
• Facebook (Zuckerberg, 2004)
• YouTube (Hurley, Chen, Karib, 2005)
• Twitter (Dorsey, Williams, Stone, 2006)
• **Evaluation (in general)**
  – NDCG (Jarvelin & Kekkalainen, 2002)
• **Geographical context in retrieval**
• Continued Moore's law
• Expansion of network connectivity & bandwidth
Mobile devices

From “Life on Mars”:

- “Call her on her mobile.”
- “Call her on her mobile what?”
In the last 50 years ...
What's grown?

- My RAM (KB)
- My CPU (KIPS)
- My Storage (MB)
- Online info I can search (GB)
- No. of online searchers
What's shrunk?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU cost (USD / MIPS)</td>
<td>1</td>
<td>$10^{-7}$</td>
</tr>
<tr>
<td>RAM cost (USD / GB)</td>
<td>1</td>
<td>$10^{-10}$</td>
</tr>
<tr>
<td>Online storage cost (USD / GB)</td>
<td>1</td>
<td>$10^{-9}$</td>
</tr>
<tr>
<td>Prevalence of manual indexing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What are the **breakthroughs** of the last 50 years which most strongly underpin the present **golden age** of information retrieval?
Hardware

- Cheap, capacious online storage
- Cheap, powerful CPUs
- Ubiquitous networking (late seventies on)
Standards

- TCP/IP etc.
- Ethernet
- Markup languages: SGML, HTML, XML
- Unicode / UTF-8
- HTTP
- URIs
- robots.txt, sitemap.xml
- ISO date formats
Software

• Word processing (hundreds)
• Webserver (httpd, apache ...)
• Web browsers (Mosaic, Netscape, ...)

Efficient IR Algorithms

- Crawling
- Indexing
- Query evaluation
- Summarisation
- Spelling suggestions
Traditional IR

• Automatic indexing (Cranfield)
• Text-based relevance ranking – probabilistic, vector space, inference networks, language model.
• (Query-biased) summarisation
• Evaluation / tuning by test collection.
• Understanding searchers
Web IR

• Crawling (Lycos, 1993?)
• SPAM rejection
• Static + Dynamic ranking
• Anchortext and other annotations
• Building on user interaction data
  – spelling suggestions
  – related queries
  – learning to rank
  – evaluation by flights
Where are we now in IR?
“tree climbing with one's eyes on the moon”

Hubert Dreyfus, *What computers can't do*, 1972
“Fanaticism consists in redoubling your efforts when you have forgotten your aim”

George Santayana, the Life of Reason, 1906
“search is a solved problem”

My boss, CSIRO, 2000
• **Searching** isn't a problem – just type words into a search box and hit 'go'! Everytime you do it, a search happens – what more do you want?

• Unfortunately, depending upon who you are, what you are searching, what you want, and what words you type, you may decide that **finding** isn't quite as well solved as searching!
“Unsolved” Search Areas

1. Contextualised search from mobile devices
2. Specialised professional search – e.g. patent search
3. Workplace / Enterprise search
Needed: More machine-friendly users!

(or tools to help them become more considerate of the limitations of systems)
e.g. Spelling suggestion tools

• Suggestions may be useful even if words are correctly spelled:
  – Manchester Untied → Manchester United

• Suggestions based on whole query, not word-by-word

• Don't suggest queries which make no sense in the collection being searched

• Autocompletion: Guide users to the best query

• Context is king
e.g. Query expansion tools

• Manual rules:
  – Rego → [registration rego]
  – MOPEM → [“manually operated personnel egress mechanism” door]

• Related queries (automatic)
  – Based on co-clicking

• Contextual navigation (on-the-fly)
  – Finding superphrases in a deep result set

• Faceting (semi-automatic)
Go8 Quick Search

Other search options: People | Projects | Publications

Search for: University: ANU

Search term/s: chemistry

1 - 10 of 1,499 search results

Local crystal chemistry, structured diffuse scattering and inherently flexible framework structures
Publication type: Book chapter
University: The Australian National University

Chemistry in Stringland: One-Dimensional Complexes of Main-Group Metal ions with the Ligands NC<sub>2</sub><sub>n</sub> (<i>n</i>= 0, 1, 2, 3)
Publication type: Journal Article
University: The Australian National University

A coupled electron diffraction and rigid unit mode (RUM) study of the crystal chemistry of some zeotropic AlPO4 compounds
Publication type: Journal Article
University: The Australian National University

Polyazolylic chelate chemistry. 13. An osmaboratrane
Publication type: Journal Article
University: The Australian National University
Query Spike Alerting

Query Frequency: 1pq

Queries per Day

Date

May  Jun  Jul  Aug  Sep  Oct  Nov
• Thanks to:
  – Everyone whose work I have relied on
  – My son Jack for his “data exploding” montage
  – Mike Swanson for the Ned Kelly line

• Apologies:
  – For the errors and omissions which inevitably occur in a cursory historical romance like this.
  – To the friends whose great contributions I couldn't fit into this idiosyncratic tale
“Search is life”