INFORMATION RETRIEVAL RESEARCH :

OLD IDEAS

CURRENT CHALLENGES

NEW POSSIBILITIES

Karen Sparck Jones

University of Cambridge

5/04

IR * research * :

for fifty years
increasingly solid
but limited operational impact

meanwhile ...

the promise of digital libraries -
quality information at your fingertips
(if you do Boolean search)

the actuality of Web engines -
information at your fingertips
(if you ask for 'Britney Spears')

IR research :
(information = document = text)

systems oriented -

focus on core tasks - indexing and searching
take context as implicit in requests,
documents, assessments
demand effectiveness

apply laboratory evaluation
formulate test design
develop performance measures

research findings :

- **indexing and searching with**
  - **derivative descriptions**
  - **distributional grounding**
  - **statistical techniques**
- **systems based on these WORK**

how use these

- **for digital libraries ?**
- **for Web engines ?**

talk structure :

1. **some research history**
2. **research state**
3. **research directions**

1 RESEARCH HISTORY

IR 1950s :

problem - growth of publications
opportunity - arrival of computers

==> automated indexing and search

key idea :

can't capture meaning
so use word patterns

key ideas :

HP Luhn late 1950s
computer support for human indexing -
look at
text word cooccurrences
text word occurrences

surface words signals for concept labels to apply -
frequent cooccurrence marks topic
`density, mass, measurement vs density, argument`
`PHYSICS RHETORIC`

frequent occurrence marks importance
`density x 10 vs density x 2`

==> forget the labels, just use the word facts :

associated word classes supply matching keys
(substitution or addition)

`mass, measurement, determination`

[query] [document]

relative frequency differentiates matching value

simple ideas, but they had something going for them

development for retrieval :

theoretical underpinning -

Maron 1960
get probability of relevance via statistics
rank search output by probability
also rerank via document associations

experimental evaluation -
test methodology :

Cleverdon early 1960s
performance measures eg recall, precision
test collection design

systematic strategy comparisons :

Salton / Sparck Jones / Robertson 1960s - 1970s

- **establishing statistically-based techniques -**
  - **simple word stems**
  - **tf - idf - rf weights**
  - **iterative feedback (implicit associations)**

- **work as well as human subject indexing**
- **advantages of search-time indexing**
- **well-suited to automation**

- **BUT experiments very small**

1980s more, bigger experiments on same lines
confirming results, supporting theory

BUT

all about system design, not user concerns
(though minimising user effort)

user studies separate strand :
needs, behaviours

difficult, laborious observation
challenging, costly experiment

especially on system-user interaction

meanwhile, operational bibliographic services

automating abstracts journals (and catalogues)

many other legitimate concerns eg speed

but

- **conventional controlled indexing (thesauri ...)**
- **constrained coordinate term search**

1980s word search, some full text, a little ranking

- **but Boolean model dominant**

quality assumptions :

research -

- **quality control on file input**
- **seriousness filter on user community**

services -

- **quality control on file input**
- **seriousness filter on user community**

+

- **quality enhancement by file-time indexing**
- **seriousness enhancement by expert advisor**

end users not always good

- **but have domain experience**

2 RESEARCH STATE

the 1990s revolution :

major change in environment -

a) Information Technology developments
b) Natural Language (Information) Processing developments

IT :

- machine power, connections
- bulk, varied stuff
- multimedia

* the Web *

NL(I)P :

- task systems
- component tools
- shared techniques

* evaluation programmes *

effects on

- IR research
- research / real world relations

the Web :

- **huge, mixed data**
  - **(not just 'proper papers')**
- **vast, varied clientele**
  - **(not just 'serious users')**
- **spread, assorted search types**
  - **(not just 'regular topics')**

thoroughly eclectic engines

some key inputs from mainstream IR research

evaluation programmes - DARPA,NIST,ARDA etc
speech recognition, information extraction ...

Text Retrieval Conferences (TREC)

systematic, controlled tests
many cycles

very large collections
many participants

==> rich comparisons
solid results

for classic topic search, confirms previous research

example : TREC data experiments
(Robertson, Walker, Sparck Jones)

150 requests, 370 K documents, full text

	precision at rank 10	
	10 terms	4 terms
unweighted terms	.11	.15
basic weighted	.52	.47
relevance weighted, expanded	.61	.51
assumed relevant	.57	.46

enlarging the envelope :

other languages, across languages -
eg Chinese
statistical methods work

other document types, cues -
eg homepages, links & URLs
statistical methods fine for topics

other media, mixed media -
eg speech, images
statistical methods on speech good
[image evaluation complexity]

Speech recognition - Av Word Error Rate = 10.7
speed 10 x real time

15.6 % WER
H: in the final hours of his administration president
S: in the final hours of his administration president
H: clinton WIPED the record clean for business
S: clinton WIPE the record clean for business
H: *** MAN GLEN BRASWELL the founder of a
S: MEN GLENN BROWSE WELL the founder of a

9.4 % WER
H: i have not seen a justification for some of the
S: i have not seen a justification for some of the
H: pardons that SEEM to be irregular and IF THEY be
S: pardons that SEEMED to be irregular and IT MAY be

example : TREC speech retrieval experiments
(Jourlin, Johnson, Sparck Jones, Woodland)

50 requests, 21 K news stories in 28K items

	mean av precision			
	11 words		3 words	
	HUM	SR	HUM	SR
known boundaries -				
basic weighted	.38	.35	.43	.40
blind feedback	.43	.37	.47	.44
partext feedback	.40	.38	.48	.45
unknown boundaries -				
basic weighted		.26		.29
partext feedback		.38		.42

further enlarging the envelope - other tasks

summarising (DUC)

- selection or condensation ?

simple statistical methods -

sentence extraction [Luhn] :

for highlighting

statistics with NLP -

select sentence parsing, text generation :

for reviewing

eg Columbia's Newsblaster

Columbia Newsblaster

Summarizing all the news on the Web

Friday, August 8, 2003
Articles from 08/05/2003 to 08/08/2003
Last update: 8:46 AM EST

Search for:

Go

in summaries

U.S.
World
Finance
Sci/Tech
Entertainment
Sports

View Today's Images

View Archives

About Newsblaster

About today's run

Newsblaster in Press

Academic Papers

Schwarzenegger joins race to replace California's Gov. Davis (U.S., 37 articles)

Gov. Gray Davis says counties will disenfranchise thousands of voters by opening fewer precincts during the Oct. 7 recall election, but election officials say opening all the polling spots would risk chaos because of a shortage of poll workers. Should California's senior solon, Democratic Senator Dianne Feinstein, abandon her reluctance and let her name be entered on the ballot for governor if Davis actually is recalled in the election now set for Oct. 7.

ACTOR–turned–candidate Arnold Schwarzenegger ended the suspense yesterday and said he would run in California's recall election, awarding Republicans his marquee value in their campaign to oust Davis. Schwarzenegger announced last night that he will be a Republican candidate in California's recall election this fall, a decision that startled political leaders around the state and that profoundly changes the landscape of the tumultuous campaign. Another Democrat, Democratic Insurance Commissioner John Garamendi, will also take out papers to run, his press secretary said early Thursday. As the state moves toward its historic recall election, the California Supreme Court has been asked to decide five separate legal challenges on the matter including a suit filed by Davis seeking to delay the Oct. 7 election.

Other stories about Schwarzenegger, Davis and Recall:

- Profile: Arnold Schwarzenegger (9 articles)

Columbia Newsblaster

Schwarzenegger joins race to replace California's
Gov. Davis (US 37 articles)

Gov. Gray Davis says counties will disenfranchise thousands
of voters by opening fewer precincts during the Oct. 7 recall
election, but election officials say opening all the polling spots
would risk chaos because of a shortage of poll workers.
Should California's senior solon, Democratic Senator Dianne
Feinstein, abandon her reluctance and let her name be
entered on the ballot for governor if Davis actually is recalled
in the election now set for Oct. 7.
ACTOR-turned-candidate Arnold Schwarzenegger ended
the suspense yesterday and said he would run in California's
recall election, awarding Republicans his marquee value in
their campaign to oust Davis. Schwarzenegger announced .....

Other stories about Schwarzenegger, Davis and Recall:
Profile: Arnold Schwarzenegger (9 articles)

evaluation issues :

complex objects, contexts, tasks

Stockbrokers are reporting a ‘spectacular’ increase in online trading as private investors storm back into the market after five successive quarters of declining business.

? Private traders storm back to markets.
? Large increase in online trading.
? Spectacular increase in private investor trading.
? Online private traders back after long break.

question answering (TREC, AQUAINT)

- quotation or construction ?

statistics for passage response

word/phrase focused extract

for reading

statistics with some NLP

sentence parsing, exact snippet selection

for application

eg Yang and Chua

question answering example - Yang and Chua :

Where did Dr King give his speech in Washington ?

In the 35 years since Dr Martin Luther King Jr
delivered his ‘‘I have a dream’’ speech at
the Lincoln Memorial, how have economic and
social questions changed for African Americans ?

==>

Lincoln Memorial

evaluation issues :

correct, adequate, useful information ?

What is the longest river in the United States?

the Mississippi
the mississippi River

? 2,348 Mississippi

? At 2,348 miles, the Mississippi River is the longest river in the US.

? The Mississipi stretches from Minnesota to Louisiana.

pervasive role of statistics :

background data gathering

eg lexicon construction

foreground text processing

eg sentence selection

combine in unifying NLIP model ==>

“language modelling ”:

statistics for implicit NLP - the ngram revolution

essential idea -

given a corpus of paired discourses A and B
correlate A features - B features
(features eg word sequences, sets)

then given a new A, derive a B

speech transcr	A = sound	B = text
translation	A = source	B = target
summarising	A = document	B = abstract
retrieval	A = request	B = rel document

probabilistic modelling with ngrams :

- predict new B-word from old A/B-words
  - (unigrams)
- predict new B-sequence from old B-sequences
  - (bi/trigrams)

retrieval needs sets, other tasks sets and sequences

train for probabilities

works well on some tasks, interestingly on others

summarising example - Banko et al :

‘President Clinton met with his top Mideast advisors, including ...., in preparation for a session with ... Israel PM Netanyahu tomorrow. Palestine leader Arafat is to meet with Clinton later ....’

==> clinton to meet netanyahu arafat

3. RESEARCH DIRECTIONS
(& LIBRARIES AND ENGINES)

in libraries, automation preceded innovation
(eg OCLC)

innovation forced by computing researchers
(eg the Web, AltaVista)

implications for *digital* libraries ?

libraries' slow takeup of research ideas :

good reasons -

- **unproven, disruptive, costly ...**
- **other factors dominate perceived performance**

bad reasons -

- **general inertia**
- **not-invented-here syndrome**

good ? bad ? reason

- **professional hostility**

Web engines rapid takeup of research ideas :

good reasons -

- **built by computer scientists**
  - **without preconceptions**
- **novel technology environment**
  - **free of traditional constraints**

bad reasons -

- **ignorance of library experience**
- **arrogant wheel reinvention**
  - **(ontologies ...)**

engines

- **applied statistics from the start :**
  - **AltaVista tf * idf weighting, ranking**
    Google link statistics
    - **(and lots else by now ...)**

but perceived lack of quality

- **does this matter ?**
- **engines a huge success**

lessons for *digital* libraries ?

what is a digital library ?

a souped up catalogue system ?
(C U Library has ‘relevance ranking’)

like ScienceDirect ?

lots of multimedia stuff ?

some special purpose database ?

“a Web engine isn’t a digital library”

HUH ?

strategies for digital library quality :

- **control input - but risky**
- **concentrate on cataloguing - but marginal**
- **provide safe searching - but constraining**
- **organise knowledge - but who can ?**

strategies for quality AND utility :

- **learn from the Web engine's hospitality**
  - **welcome objects, attitudes**
- **exploit research findings**
  - **especially statistical methods**

ie apply general retrieval research lesson :

use statistical data as far as you can
[and seek further] -

there are bulk language data for the asking

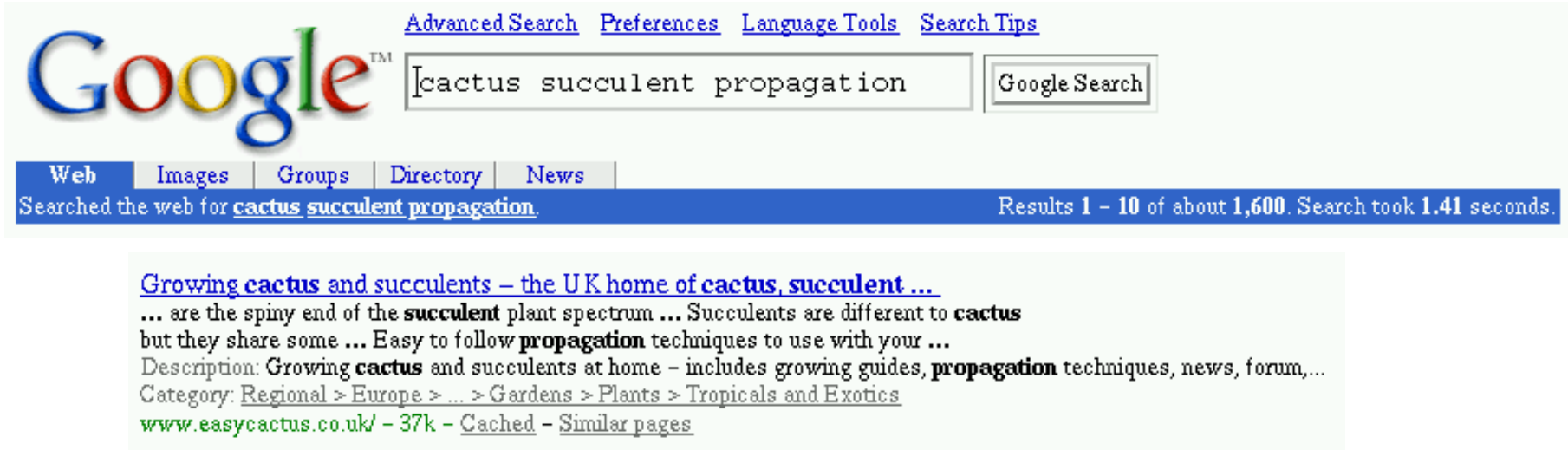
there are general, available processing methods
(pattern matching, classification, learning)
for 'finding like things'

==> training for better quality

statistical methods

- **good for some tasks**
  - **eg document retrieval, speech recognition**
- **adequate for some 'near' tasks**
  - **eg indicative summarising, selective extraction**
- **helpful for some complex task subtasks**
  - **eg question answering, multi-text summarising**
- **encourage multi-task integration**
  - **generality helps common perspective**
  - **simplicity givess easy trials**
    - **eg retrieval and query-oriented summary**

GOOGLE / ALTAVISTA



Growing cactus and succulents – the UK home of cactus, succulent and lithops info and shopping
... and succulents at home – includes growing guides, **propagation** techniques, news, forum, events and **cactus** shopping ... are the spiny end of the **succulent** plant spectrum and they come in a vast ...
www.easycactus.co.uk * Related Pages
More pages from www.easycactus.co.uk

query : cactus succulent propagation

Google -

Growing cactus and succulents - the UK home of cactus, succulent ...
... are the spiny end of the succulent plant spectrum ... Succulents are different to cactus but they share some ... Easy to follow propagation techniques to use with your ...

AltaVista -

Growing cactus and succulents - the UK home of cactus, succulent and lithops info and shopping
... and succulents at home - includes growing guides, propagation techniques, news, forum, events and cactus shopping ... are the spiny end of the succulent plant spectrum and they come in a vast ...

www.easycactus.co.uk/

TAKE-HOME MESSAGE :

statistical methods work through redundancy

all use of language has redundancy

so

statistical strategies are sound basic tools for information management

Sparck Jones et al, Info Proc and Mgmt 36, 2000

Jourlin et al, TR 517, Comp Lab, U of Cam, 2001

www-nlpir.nist.gov/proj_act.html

newsblaster.cs.columbia.edu

www.ic-arda.org/infoExploit/aquaint/index.html

Yang and Chua, Proc TREC 2002, 486-491

Banko et al, Proc ACL 2000