

**INFORMATION RETRIEVAL RESEARCH
AND
DIGITAL LIBRARIES**

Karen Spärck Jones

University of Cambridge

8/03

IR research :

for fifty years

increasingly solid

but belated, uneven operational impact

research findings :

achievements so far ?

implications for digital libraries :

new scope for current methods ?

need for new methods ?

IR research characteristics :
(information = document = text)

starting point :

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

progress by :

developing evaluation methodology -
test design
performance measures

research findings :

indexing and searching with

derivative descriptions
distributional grounding
statistical techniques

systems based on these **WORK**

talk structure

1. some research history
2. what it's shown
3. where it leads

1 RESEARCH HISTORY

IR early concern in computing

initial goal :

automate existing strategies

BUT

novel strategies emerged

[linking information management tasks
eg retrieval, summarising]

Luhn 1957 +

computer support for indexing, searching :

 statistical word associations, weighting

extended to other IM tasks :

 summarising by statistically-based
 sentence extraction

OPENED PANDORA'S BOX

Maron & Kuhns

1960

development of theory for statistical approach :

probability of relevance

using

statistical term weights, associations

output ranking

statistical document attraction

(iterative feedback)

Cleverdon 1960 +

development of index, search method testing :

systematic controlled experiments

factors affecting performance

devices achieving performance

performance measures

established an evaluation paradigm

Salton (SMART)

1962 +

development of computational systems :

combining

text-statistical base

decompositional laboratory evaluation

best results with

abstract texts

word stems, weights, scores

relevance feedback

also, manual thesaurus

BUT collections very small, performance variation

the 1970s and 1980s :

building on 1960s ideas and findings -

more and different models
eg inference networks

bigger and varied test collections
eg 11500 documents, not 1000

extra and alternative performance measures
eg document cutoff

wider range and finer grain on

environment variables eg document types
system parameters eg clustering methods

SO

better understanding of retrieval itself
eg uncertainty as constraint on performance

more evidence for good methods
ie helpful statistics pervasive as tool

research mainstream (core indexing, searching) :

simple natural language good

statistical weighting good

relevance feedback good

ranked output good

clustering (term, document) no good

grammatical analysis no good

no gain from thesaurus/subject headings

eg Salton Medlars test

gain over boolean match

but limitations :

tests small scale

remote from users and use context

cut off from interaction

thirty years research outcome :

black box in opaque mathematical packaging

a world apart from operational systems

... why ?

the operational world :

automating libraries - cataloguing etc
MARC, OCLC ...

automating bibliographic services - databases,
retrieval
INSPEC, MEDLARS ...

going online
DIALOG, ORBIT, ESA, MEDLINE ...

exploiting text
LEXIS, STAIRS ...

[SCI exception]

entrenched conventional assumptions on core -
controlled language indexing
boolean matching

vital non-core concerns -
document delivery
file coverage
interface design
multiple languages
.....

but some incidental convergence with research -
natural language
full text

and some prototype research-style systems

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 ?

statistical sentence extraction

crude but may be useful, passages more so ?

statistics + lightweight NLP

(anaphor resolution, phrase extraction)

less crude and probably useful

statistics + non-trivial NLP

(select sentence parsing, text generation)

looks good :

eg Columbia's Newsblaster

issue : what is a summary for (and how evaluate) ?

Summarising -

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.

Search for:

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)

question answering (TREC, AQUAINT)
- quotation or construction ?

statistical passage (not snippet) extraction
crude but may be useful

statistics + very light NLP + Web
(alternative answer snippet strings)

surprisingly effective

statistics + heavy NLP

(typing, parsing, enhancing, unifying)

demonstrated very effective :

eg Language Computer Corp's QASTM

issue : what is an answer (and how evaluate) ?

Question answering - snippet responses [TREC]

What river in the US is known as the Big Muddy?

the Mississippi

Known as Big Muddy, the Mississippi is the longest
messed with . Known as Big Muddy , the Mississip
the Mississippi is the longest river in the
Mud Island,; Mississippi; ‘‘The; history; Memphis

?

The Mississippi’s brown, earth-laden waters are
very distinctive

[Home](#) > [Demos](#) > [Question Answering](#) > Internet Demo

Internet Demo

Open Search. Concise Answers.

Type your question:

...and get the answer **Power Answer**

Question Examples

- ▶ [When did the Challenger space shuttle explode?](#)
- ▶ [What percent of the earth air is oxygen?](#)
- ▶ [Which is the largest volcano in Europe?](#)
- ▶ [How many calories in a glass of wine?](#)
- ▶ [Who invented Coca Cola?](#)

Have WMD been found in Iraq ?

3. 06/09/03 : (PIPA) A striking finding in the new Program on International Policy Attitudes (PIPA) Knowledge Networks poll is that many Americans are unaware that weapons of mass destruction have not been found in Iraq...
lists.gp-us.org/pipermail/texgreen/2003-June/002367.html
5. Fish of Mass Destruction the Top of an Iceberg...
skog.de/ennukes.htm
9. Is it not ironic that the only WMD found in Iraq were never lost in the first place...
theblackhandside.net/2003/06/wmd_have_been_found.html

unifying model development : “language modelling”

the ngram revolution -
statistics and implicit NLP

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

works very well on some tasks,
interestingly on others

OBSERVATIONS ON STATE

research - engines - libraries relationship :

in libraries, automation preceded innovation
(eg OCLC)

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

[many retrieval researchers in computing]

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

in computing research, no baggage :

good effects -

rapid action on ideas

boundary crossing

bad effects -

ignorance of library experience

wheel reinvention

(ontologies ...)

can we improve on current state

with deeper notion of 'digital library' ?

3 RESEARCH IMPLICATIONS FOR DIGITAL LIBRARIES

(assuming infrastructure issues addressed -
document formats ...)

obviously,

exploit computing & Web developments :

welcome

new information supports eg mobile access

new information objects, eg lecture slides

new information cues eg URL links

adopt retrieval research findings -
statistical text-based strategies

import other technologies -
speech processing, natural language processing

BUT ALSO,

much more importantly ==>

systematically apply the general retrieval
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'

statistical methods are

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

statistical methods promote multi-task integration

generality encourages common perspective

simplicity encourages easy trials

eg retrieval and query-oriented summary



[Advanced Search](#) [Preferences](#) [Language Tools](#) [Search Tips](#)

cactus succulent propagation

Google Search

Web Images Groups Directory News

Searched the web for [cactus succulent propagation](#).

Results 1 - 10 of about 1,600. Search took 1.41 seconds.

[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 ...

Description: Growing **cactus** and succulents at home - includes growing guides, **propagation** techniques, news, forum,...

Category: [Regional](#) > [Europe](#) > ... > [Gardens](#) > [Plants](#) > [Tropicals and Exotics](#)

www.easycactus.co.uk/ - 37k - [Cached](#) - [Similar pages](#)



Web Images MP3/Audio Video Directory News

Advanced Family Filter: **off** Settings

cactus succulent propagation

FIND

[More Precision](#)

SEARCH: [Worldwide](#) [U.K.](#) RESULTS IN: [All languages](#) [English](#)

[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](#)

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.languagecomputer.com