INFORMATION THAT COUNTS

Karen Sparck Jones

University of Cambridge

3/04

Fifty years ago :

problem:
rapidly growing scientific/technical literature
- how to organise, access ...

opportunity:
appearance of general-purpose digital computers

==>

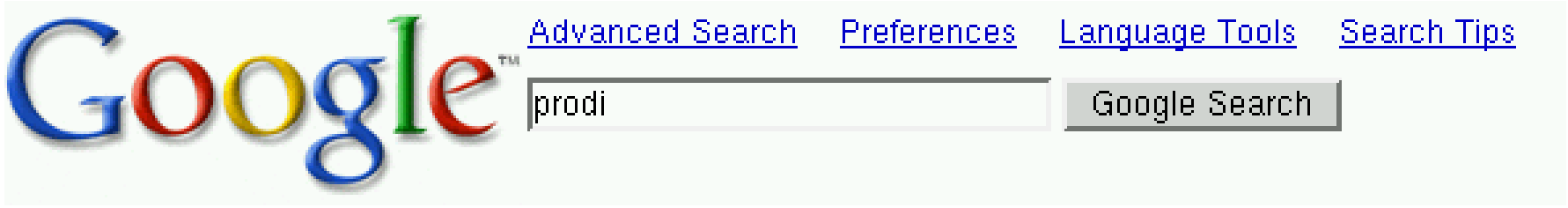
automated indexing and searching
automated summarising
automated translation

Now :

Web engines with

- **billions of pages**
- **tens of billions of searches**

automated indexing, summarising, translation ...



Biografías de Líderes Políticos CIDOB: Romano **Prodi** (Unión ... - [Translate this page]
Romano **Prodi** Unión Europea/Italia * 7 de agosto de 1939,
Scandiano, región de Emilia-Romagna (Italia). ...
Description: Breve biografía sobre este político del país.
Category: World > Español > Países > Europa > Italia > Sociedad
www.cidob.org/bios/castellano/lideres/p-018.htm - 24k - 1 Mar 2004 - Cached - Similar pages

And studió the race of Jurisprudence in the Catholic University of the Sacred Heart of Milan, where cum received the master's degree renders in 1961, and, after attending posgrado in the London School of Economics (LSE), in 1963 doctoró with a thesis on protectionism in the Italian industry in the first years of the unit of the country.

Its educational and investigating activity in the Catholic University of Milan and the Lombardo Institute of Economic Studies and Sociales began (ILSES), but its bond more hard was with the University of Bologna, first like professor associated in the chair of Political Economy of the Faculty of Jurisprudence and soon like director of Center of Economy and Industrial Poli'tica. There it was member of an academic group that emphasized in a the analysis of the Italian industrial development, the frame of his competition of market and its policies of enterprise fusion. In 1968 it exerted briefly of investigator in Stanford Research Institute, of the United States.

Google output, search: Prodi

Biografias de Lideres Politicos ... Romano Prodi (Union ...)

Romano Prodi Union Europea/Italia "7 de Agosto 1939,
Scandiano, region de Emilia-Romagna (Italia) ...

Comenz su actividad docente e investigadora en la Universidad
--
Catlica de Miln y en el Instituto Lombardo de Estudios Econmicos
y Sociales (ILSES), pero su vnculo ms fuerte fue con la
Universidad de Bolonia, primero como profesor asociado en la
ctedra de Economa Poltica ....

Its educational and investigating activity in the Catholic University
---
University of Milan and the Lombardo Institute of Economic Studies
and Sociales began (ILSES), but its bond more hard was with the
University of Bologna, first like professor associated in the
chair of Political Economy ....

what happened between then and now ?

where are we going ?

talk structure :

- **history**
- **current research**
- **observations**

1. HISTORY

language and information processing
a complex human activity -
interpreting, manipulating meaning

how to emulate well enough ?

information (document, text) retrieval
pressing, potentially tractable

not answering question, but giving user material
by content indicator match

```
How can carbon emissions be reduced ?
        CARBON EMISSION     CONTROL
```

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

apply ideas to other information management tasks :

automatic summarising -
extract key sentences by word scores

IC51 INTERNATIONAL CONFERENCE ON SCIENTIFIC INFORMATION
AREA 5 PG 103

THE ANALOGY BETWEEN MECHANICAL TRANSLATION AND LIBRARY RETRIEVAL

MASTERMAN M CAMBRIDGE LANGUAGE RESEARCH UNIT CAMBRIDGE ENGLAND
NEEDHAM RM CAMBRIDGE LANGUAGE RESEARCH UNIT CAMBRIDGE ENGLAND
JONES ES CAMBRIDGE LANGUAGE RESEARCH UNIT CAMBRIDGE ENGLAND

AUTO ABSTRACT

6 NOW, IN THE PRESENT STATE OF RESEARCH THIS ANALOGY CAN ONLY BE DRAWN AT ALL PRECISELY BETWEEN ONE FORM OF LIBRARY RETRIEVAL PROCEDURE, AND ONE FORM OF MECHANICAL TRANSLATION PROCEDURE.. THESE TWO ANALOGOUS PROCEDURES ARE THOSE, IN EACH FIELD, WHICH MAKE USE OF A THESAURUS.

10 WE PROPOSE, THEN, THAT A CONCEPTUALLY BASED, THESAURUS TYPE OF LANGUAGE CLASSIFICATION SHOULD BE USED FOR A COMPLETELY GENERALISED RETRIEVAL PROCEDURE, THIS CLASSIFICATION PROCEDURE BEING, BY ITS NATURE, INTERLINGUAL.

14 TRANSLATION SPECIALISTS, AND, IN PARTICUALR, LINGUISTS DENY EVEN THE POSSIBILITY OF THE ANALOGY BY MAINTAINING THAT ANY CLASSIFICATION OF LANGUAGE BASED ON A THESAURUS CAN, AT BEST, ONLY HOPE TO TRANSLATE SEMANTIC MEANING, WHEREAS LANGUAGE IS PRIMARILY A SYSTEM OF GRAMMAR AND SYNTAX.., AND BOTH OF THESE ARE NOTORIOUSLY MONOLINGUAL.

18 THE OBJECT OF THIS PAPER IS TO REFUTE THIS CRITICISM BY SHOWING HOW A TYPE OF RETRIEVAL PROCEDURE, BASED ON A THESAURUS ALREADY BEING USED FOR THE EXPERIMENTAL TRANSLATION OF SEMANTIC MEANING, MIGHT ALSO BE EXTENDED SO AS TO TRANSLATE GRAMMAR AND SYNTAX.

automatic translation -

identify input word senses for output equivalent

`The farmer cultivates the field.`

words repeating `AGRICULTURE` **concept select**

`field = land`

construct thesaurus from text corpus

apply to new text

language and information processing without
explicit reference to meaning

using

word distribution data
statistical techniques to interpret, apply

how can anything so simple work ?

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 techniques -**
  - **simple word stems**
  - **tf - idf - rf weights**
  - **iterative feedback**

- **work as well as human subject indexing**
- **well-suited to automation**

- **BUT experiments very small**

retrieval research to late 1980s :

consolidation - methods, results

BUT no real impact on operational systems :

emphasis on machine files eg Chemical Abstracts
index by subject headings, boolean search

other tasks :

summarising - very little work
no full text available

translation - some work
no good data, methods to build thesauri

2. CURRENT RESEARCH

major developments in the 1990s :

large-scale task evaluation programmes

progress with language processing systems
analysis for meaning, synthesis from meaning

demand for task applications

vast text data, powerful machines,
statistical programs

THE WEB

(US) evaluation programmes -
text retrieval, fact extraction, summarising ...

Text Retrieval Conferences (TREC)

systematic, controlled tests
many cycles

large full text files
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

tf (dl), idf, rf as weight components

spreading the technology :

other task forms eg filtering

other languages eg Chinese, mixed languages

other data types eg Web pages, cues eg URLs

other media eg (automatically transcribed) speech

statistical methods work

research ideas spread to real systems :

Web engines new challenge, new opportunity -

no strategy preconceptions

adopted statistical ideas -

weighting (tf, idf)
ranking

[but mixed with much else ...]

other tasks :

summarising -

selection or condensation ?

statistical sentence extraction

crude but may be useful, passages more so ?

statistics + light NLP

(anaphor resolution, phrase extraction)

less crude and probably useful

statistics + medium NLP

(select sentence parsing, text generation)

looks alright :

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 -

statistics for passages - perhaps useful
statistics + light NLP for snippets - passable
statistics + heavy NLP for exact quotes - maybe good

snippets:

```
Where is the Taj Mahal ?

The Taj Mahal is in [ India ] ...
The Taj Mahal by the Jumna in India ...
The Taj Mahal, finer than any tomb in Persia,
  is in India
There are fine monuments in India. There are ...
  ... Taj Mahal.
```

statistics everywhere : what are they doing ?

discriminating preferentially within a large noisy mass

one word set, word string better than another

as an interpretation / representation of Y *in relation to* X

eg document words in relation to query words

unifying statistical model development :

"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 sets, sequences)

then given a new A, derive a B

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

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

LM 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. OBSERVATIONS

statistical operations on text surfaces work
because they capture actual language use

they fit a fundamental feature of language :
individual words are ambiguous
word combinations are unambiguous

ie redundancy counteracts noise

so push statistics as far as you can :

- **have bulk language data to hand**
- **have general processing methods to hand**
  - **(pattern matching, classification, learning)**
  - **for 'finding like things'**

statistical approaches are easy :

- **good for some tasks eg retrieval**
- **sufficient for simple forms of some tasks**
  - **eg rough summarising, question answering**
- **supports for complex tasks involving NLP**
  - **in modules, via resources**

generality, simplicity, encourage multi-tasking

in basic 'can do something' mode -

eg basic retrieval AND query-oriented summary
Google, AltaVista ...

categorisation AND filtering

within 'can do a good deal' mode -

===>

multi-task integrated systems :

example - Mitre's MiTAP

informing about biological emergencies

capture documents
- transcribe spoken
- translate foreign
extract 'named entities'
- people, places, times, diseases ...
summarise texts
- single, grouped
route to newsgroups
retrieve from file

operational prototype

- **real time, large data streams, many users**
- **high-class user interface**

applies available tools, subsystems

exploits statistics at many points

- **eg tf * idf for retrieval**
- **eg tf * idf for summarising**

THE SHAPE OF THINGS TO COME