Shelf life: 3 years

Introduction to Information Retrieval

(ESSIR 2003)

C.J. "Keith" van Rijsbergen
Computing Science
Glasgow University

Definitions of Information Retrieval

(Salton, 1968) – Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.

(Needham, 1977).....the complexity arises from the Impossibility of describing the content of a document, Or the intent of request, precisely, or unambiguously

Time I (highlights for me,biased)

1952 Mooers coins IR
1958 International Conference on Scientific Information
1960 Cranfield I
1960 Maron and Kuhns paper
1961 Towards IR, RAF
1961 (-1965) Smart built
1964 Washington conference on Association Methods
1966 Cranfield II
1968 Salton's first book
197- Cranfield conferences
1975 CvR's book
1975 Ideal test collection
1976 KSJ/SER JASIS paper

Time II

1978 1st SIGIR
1979 1st BCSIRSG
1980 1st joint ACM/BCS conference on IR
1981 KSJ book on IR Experiments
1982 Belkin et al ASK hypothesis
1983 - Okapi started
1985 RIAO-1
1986 CvR logic model
1990 Deerwester et al,LSI paper
1991 CoLIS 1 (in Tampere!)
1991 – Inquiry started
1992 Ingwersen's book
1992 TREC-1
1998 Croft Ponte paper on language models

Experimental Methodology

Cleverdon	Cranfield
Lancaster	Medlars
Keen	Cranfield/Smart
Saracevic	CWRU
Salton	Smart
Sparck Jones	Ideal Test Collection
Blair & Maron	Stairs
Harman	TREC

Evaluation

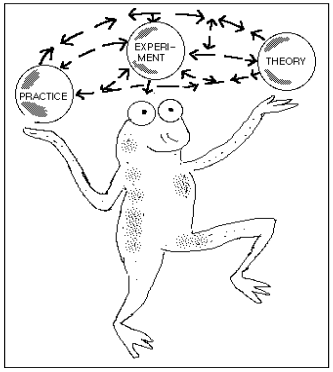
ABNO/OBNA		(Fairthorne)
Precision, Recall	-> trade-off	(Cleverdon)
Probabilistic versions		(Swets)
Measure-theoretic		(Bollman)

Scales

scale	operation	group	statistics
nominal	equality	permutation 1:1	mode
ordinal	greater/less	isotonic monotone	median
interval	equality/diff of intervals	linear x'=ax+b	mean
ratio	equality of ratios	similarity x'=ax	coeff of variation

Some meta thoughts

A posteriori	A priori
OWA	CWA
Adaptive	Non-adaptive
Data driven	Theory driven
Information	Knowledge
Contingency	Necessity
Ostensive	Extensive



Practice: Web
Electronic Publishing
Task-oriented IR
Data Mining
Knowledge Discovery
Distance learning
Video/film asset management

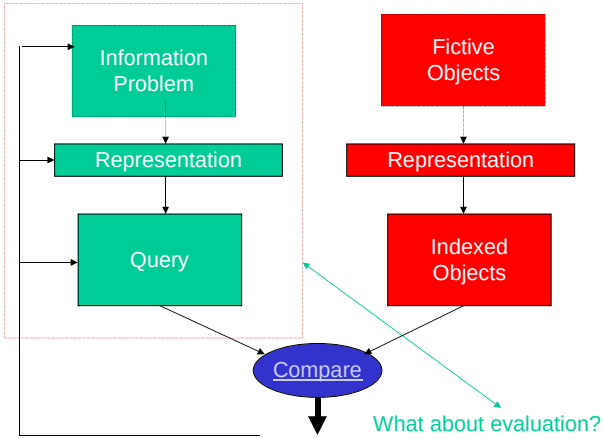
Experiments: TREC
HCI
Visualisation
Work in Context, Cognitive approaches
Cross - lingual
Cross - media
Corpus-based IR (inc. wordnet, etc)
Digital Libraries
CBIR
TDT

Theory

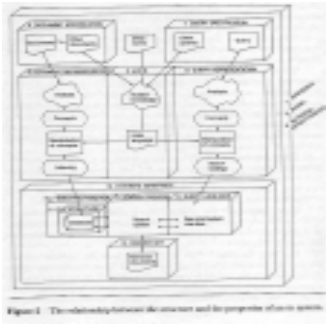
Knob twiddling
Data fusion
Authority/importance models
Logic + Uncertainty models
Filtering/Routing
Language models
Summarisation
Discrimination/Representation
IR + DBMS (inc XML etc)
Clustering the web
Visualising the web
Living with single term queries
Living with no queries

Theory (cont.)

Scale free networks
Trading media (text helps images!)
Temporal dimensions (topics,events)
Evaluation (Time to dump 'P and R'?)
NLP in IR



Architecture (Brenda Gerrie, 1983)



Matching	Exact Match	Partial (best) Match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query Language	Artificial	Natural
Query Definition	Complete	Incomplete
Query Dependence	Yes	No
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive
Logic	Classical	Non-classical
Representation	A priori	A posteriori
Language Models	Logical	Statistical

Matching

- exact/partial match e.g SQL/Dice
- Boolean matching (Fairthorne, 50)
- co-ordination level matching (Cleverdon,60)
- cosine correlation (Salton, 70) VS
- probabilistic (ranking principle) (SER,80) PRP
- logical uncertainty principle (CvR, 90) LUP
- plausible inference (Croft,90) NET

Inference

- Deduction/Induction: A, A→B infer B
- Cluster Hypothesis
- Association Hypothesis
- $P(term_1|term_2)$

Cluster Hypothesis

If document X is closely associated with Y, then over the population of potential queries the probability of relevance for X will be approximately the same
as the probability of relevance for Y, or in symbols

$$P(\text{relevance}|X) \sim P(\text{relevance}|Y)$$

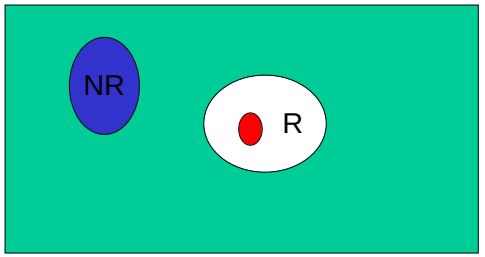
Association Hypothesis

If one index term is good at discriminating relevant from non-relevant documents, then any closely associated index term is also likely to be good at this.

Models

- Boolean
- Vector Space (metrics) - mixture of things
- Probabilistic (3 models)
- Logical (implication) - what kind of logic
- Language models
- (Algebraic model)
- Cognitive (users)
- Language (distributions) - Bose-Einstein?

Partial Models



Classification

* Studied early in IR (1960s, 1970s). Lost favour in 80s

* Returned in 90s for different applications (e.g. browsing)

* Van Rijsbergen did early work on applying more formal techniques , e.g. single-link hierarchies - followed by....

* Sparck Jones did early work on term clustering

* Salton's group did many experiments with different clustering techniques

* Roger Needham did a thesis on clustering (!)

* Bruce Croft did his thesis on clustering

Celestial Emporium of Benevolent Knowledge

"On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included into this classification
- those that tremble as if they were mad,
(j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance."

Borges

Query Language

- Artificial/Natural (web)
- multilingual/cross-lingual
- images
- none at all!

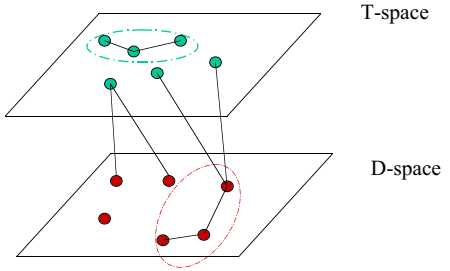
Query Definition

- Complete/Incomplete
- Independence/Dependence
- Weighted/Unweighted (tf $\times$ idf)
- Query expansion/one shot (feedback, web)
- Sense disambiguation
- Cross-lingual

Query Dependence

- Ostensive retrieval
- hyperlinks
- citation links
- filtering
- collaborative filtering
- authority/importance

Navigation - Browsing



Items Wanted

- Matching/Relevant or Correct/Useful
- The function of a document retrieval system cannot be to retrieve all and only the relevant documents....but to *guide* the patron in his search for information (Maron)
- Topical/tasks
- Meaning/content

Some difficulties with 'relevance'

Goffman, 1969: '..that the relevance of the information from one document depends upon what is already known about the subject, and in turn affects the relevance of other documents subsequently examined.'

Maron, : 'Just because a document is about the subject sought by a patron, that fact does not imply that he would judge it relevant.'

Maron's theory of indexing

…..in the case where the query consists of single term, call it B, the probability that a given document will be judged relevant by a patron submitting B is simply the ratio of the number of patrons who submit B as their query and judge that document as relevant, to the number of patrons, who submit B as their search query

'That is the relevance or irrelevance of a given retrieved document may affect the user's current state of knowledge resulting in a change of the user's information need, which may lead to a change of the user's perception/ interpretation of the subsequent retrieved documents….' Borlund, 2000

Error Response

- Precision: error where an irrelevant is retrieved
- Recall: error where a relevant document is not retrieved
- Trade-off
- How to cope with lack of recall
- Cranfield →Ideal test collection →TREC →????

Representation of Information

- Discrimination without Representation (specificity)
- Representation with Discrimination (exhaustivity)

...defining a concept of 'information',....[that] once this notion is properly explicated a document can be represented by the 'information' it contains (CvR, 1979)

Images not Text: how might that make a difference?

no visual keywords (yet)
- tf/idf issue

aboutness revisable (eg Maron)
relevance revisable (eg Goffman)
feedback requires salience
aboutness -> relevance -> aboutness

Text

- keywords
- frequency
- meaning
- grammar
- salience?
- relevance
- query expansion

Images

- ?
- ?
- object recognition
- geometry
- salience
- path dependent
- how?

DUCK

RABBIT

Inference

It is a common fallacy, underwritten at this date by the investment of several million dollars in a variety of retrieval hardware, that the algebra of Boole (1847) is the appropriate formalism for retrieval design…..The 'logic' of Brouwer, as invoked by Fairthorne, is one such weakening of the postulate system,……
Mooers, 1961

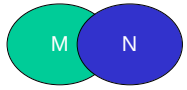
Another one:
Logical Uncertainty Principle
CvR, 1986

Logic

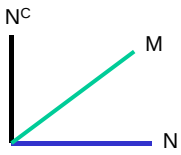
If Mark were to loose his job, he would work less
If Mark were to work less, he would be less tense

If Mark were to loose his job, he would be less tense

$A \rightarrow B$, $B \rightarrow C$ infer $A \rightarrow C$



$M \cap (N^C \cup N) = M$
$(M \cap N^C) \cup (M \cap N) = M$



$M \otimes (N^C \oplus N) = M$
$(M \otimes N^C) \oplus (M \otimes N) = \Phi \neq M$

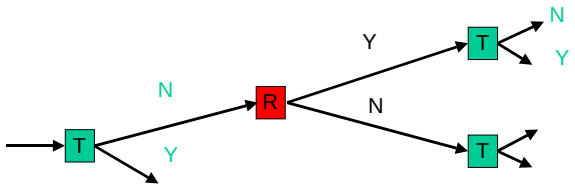
Interaction (Aboutness)

Objects: documents, queries ⟶ Relevance

Model

Observable(States) ⟶ ??

Relevance/Aboutness
is
Interaction/User dependent



Where are we now in IR?

- Landmarks
- Hypotheses/Principles
- Postulates of Impotence
- Long-term challenges
- Areas of research

Landmarks

Luhn's tf weighting
Architecture
Relevance Feedback
Stemming
Poisson Model -> BM25
Statistical weighting tf*idf
Various models

Hypotheses/Principles

Items may be associated without apparent meaning but exploiting their association may help retrieval

P & R trade-off – ABNO/OBNA
Exhaustivity/Specificity
Cluster Hypothesis
Association Hypothesis
Probability Ranking Principle
Logical Uncertainty Principle
ASK
Polyrepresentation

Postulates of Impotence

(according to Swanson, 1988)

- An information need cannot be expressed independent of context
- It is impossible to instruct a machine to translate a request into adequate search terms
- A document's relevance depends on other seen documents
- It is never possible to verify whether all relevant documents have been found
- Machines cannot recognise meaning -> can't beat human indexing etc

….more postulates

- Word-occurrence statistics can neither represent meaning nor substitute for it
- The ability of an IR system to support an iterative process cannot be evaluated in terms of single-iteration human relevance judgment
- You can have either subtle relevance judgments or highly effective mechanised procedures, but not both
- Thus, consistently effective fully automatic indexing and retrieval is not possible

Long-term Challenges – workshop Umass. 9/2002

Global information access: Satisfy human information needs through natural efficient interaction with an automated system that leverages world-wide structured and unstructured data in any language.

Contextual Retrieval: Combine search technologies and knowledge about query and user context into a single framework in order to provide the most "appropriate" answer for a user's information need.

Areas of Research

- How does the brain do it? (neuroscience)
- How do we see to retrieve? (computer vision)
- How do we map IR onto Quantum Computation? (QM)
- How do we reduce dimensionality in dynamic fashion? (Statistics)
- What is a good logic for IR? (mathematical logic)
- What is a good theory of uncertainty? (frequency/geometry)
- How do we model context? (HCI)
- How do we formally capture interaction?
- How do we capture implicit/tacit information?
- Is there a theory of information for IR?

Useful References

Readings in Information Retrieval, Morgan Kaufman, Edited by Sparck Jones and Willett

Advances in Information Retrieval: Recent Research from CIIR, Edited by Bruce Croft.

Information Retrieval: Uncertainty and Logics,Advanced Models for the Representation and Retrieval of Information, Edited by Crestani, Lalmas, Van Rijsbergen.

Finding out about, Richard Belew.

<new book>, Ingwersen and Jarvelin.