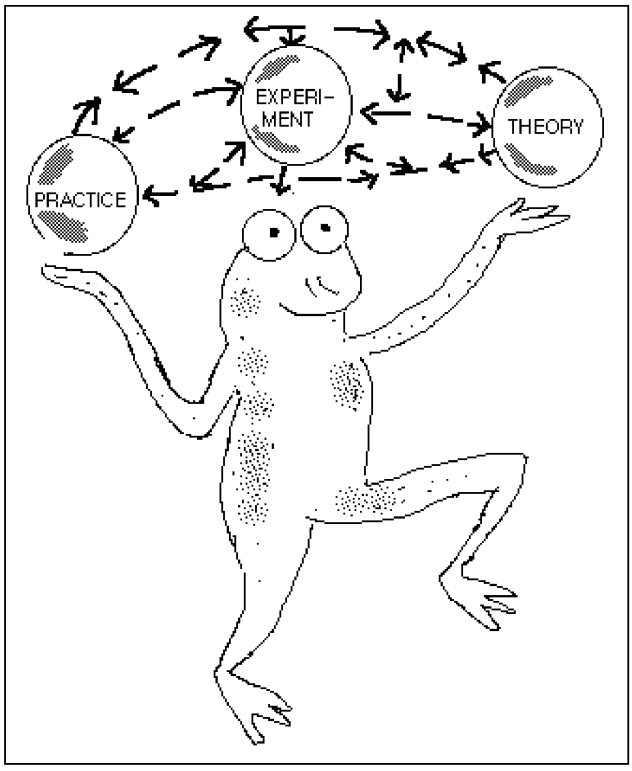
Introduction to Information Retrieval

C.J. “Keith” van Rijsbergen

Computing Science

Glasgow University



Some meta thoughts

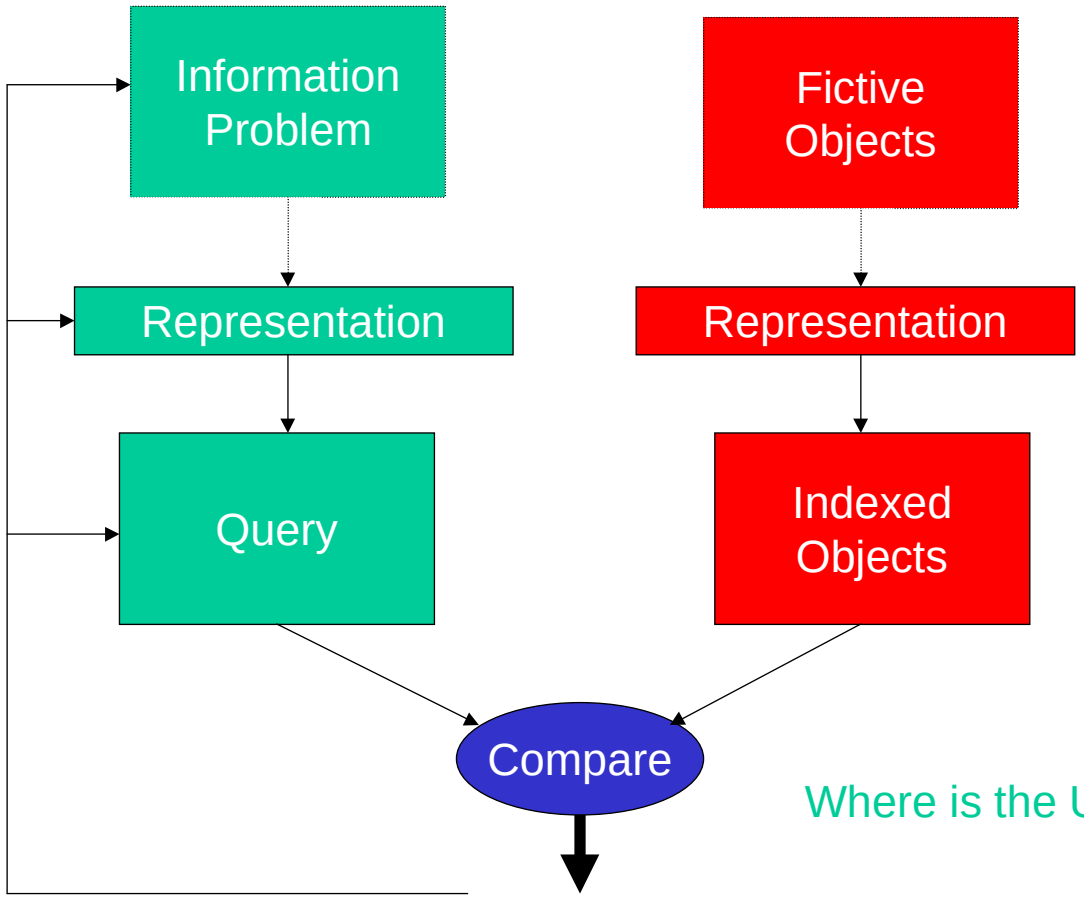
A priori	A posteriori
CWA	OWA
Adaptive	Non-adaptive
Data driven	Theory driven
Information	Knowledge
Contingency	Necessity
Ostensive	Extensive

Practice:
- Web
- Electronic Publishing
- Task-oriented IR
- Data Mining
- Knowledge Discovery
- Distance learning

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

Theory:

Knob twiddling
Data fusion
Authority/importance models
Logic + Uncertainty models
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
Trading media (text helps images!)
Temporal dimensions (topics, events)
Evaluation (Time to dump ‘P and R’?)
NLP in IR
.....



Where is the User?

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(relevance|X) ~ P(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.

Discrimination Gain Hypothesis

(hidden variables)

Under the hypothesis of conditional independence the statistical information contained in one index term about another is less than the information contained in either index term about relevance.

$$P(X,Y|W) = P(X|W) * P(Y|W)$$

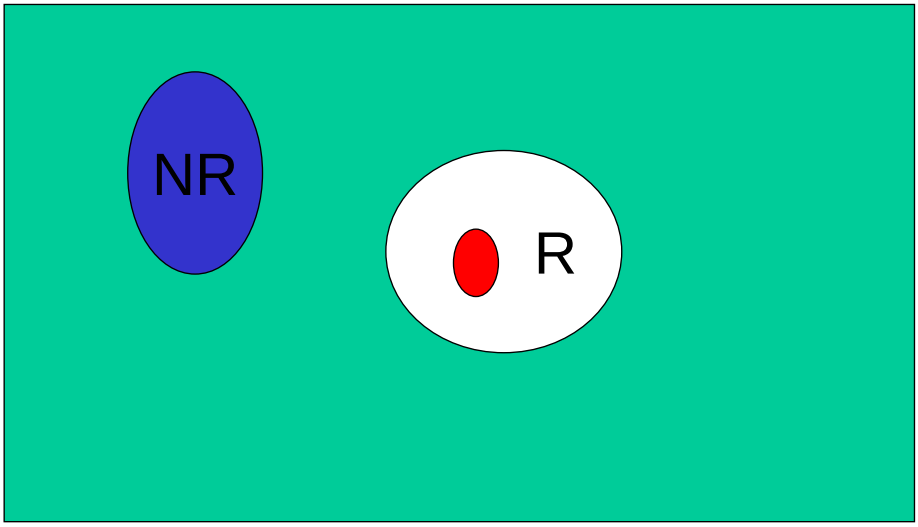
$$I(X,Y) < I(X,W) \text{ or } I(Y,W)$$

Are there other explanatory variables?
Does W exist as a variable?

Models

- Boolean
- Vector Space (metrics) - mixture of things
- Probabilistic (3 models)
- Logical (implication) - what kind of logic
- (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 and 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 (i) 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

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

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)

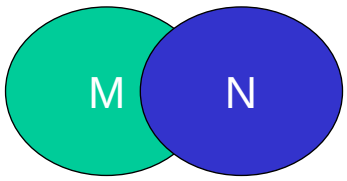
...defining a concept of 'information',....[that] once this notion is properly explicated a document can be represented by the 'information' it contains (CvR, 1979)

Logic

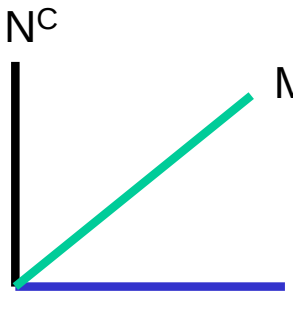
If Mark were to loose his job, he would work less
If Mark were to work less, he would be less tense

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