

Shelf life: 2 years

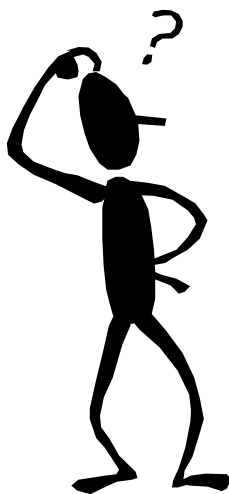
Introduction to some Models in Information Retrieval (EARIA 2006)

C.J. “Keith” van Rijsbergen
(with help from the IR group in Glasgow)
Computing Science
Glasgow University

EARIA 2006

© CvR

Scenarios & Applications



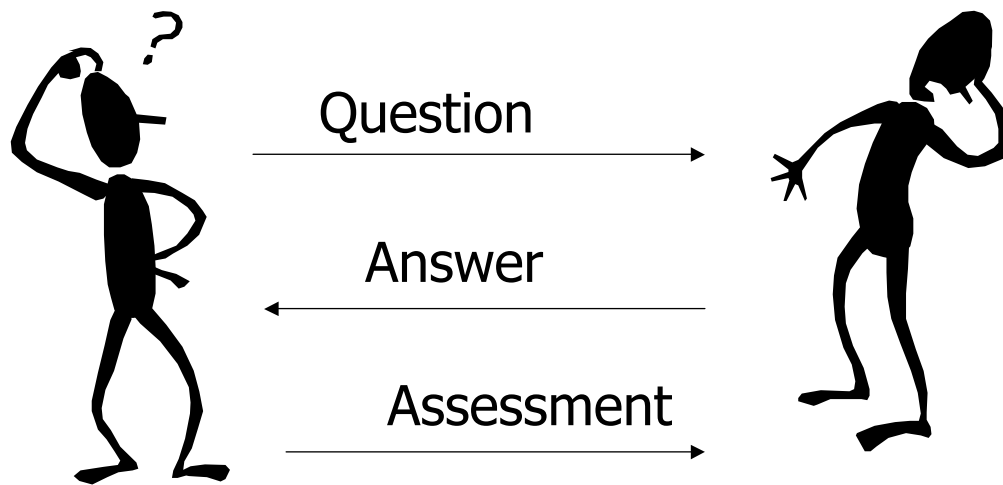
Documents

Email Messages
XML Documents
Web pages
....

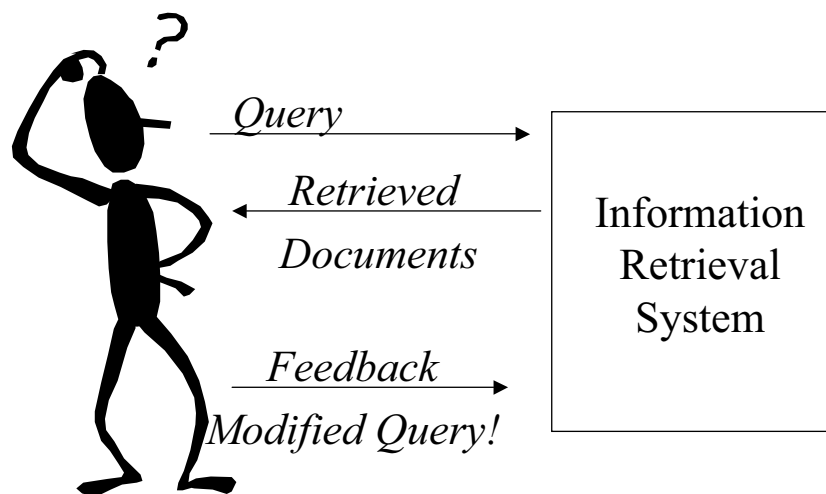
EARIA 2006

© CvR

Retrieval – A Question-answer scenario



Retrieval Loop



What is Information Retrieval? (I)



- Quite effective (at some things)
- Highly visible (mostly)
- Commercially successful (some of them, so far)

- But what goes on behind the scenes?
How do they work?
Is there more to it than the Web?

So, what is IR? (II)

- General definition
 - Retrieval of unstructured data
 - Most often it is
 - Retrieval of text documents
 - Searching newspaper articles
 - Searching on the Web
 - Other types of retrieval
 - Image retrieval
 - Video retrieval
 - Music retrieval

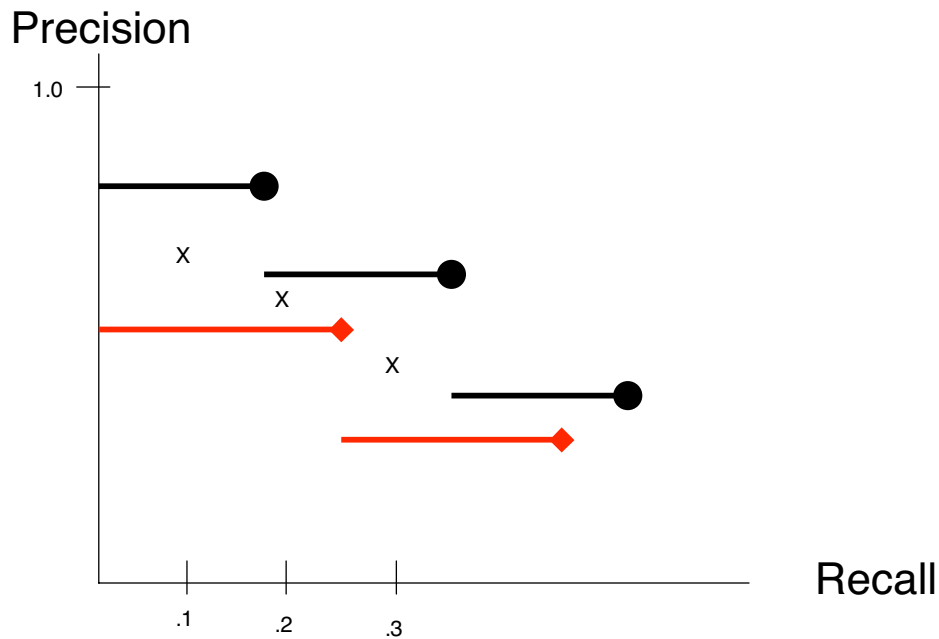
Experimental Methodology

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

Evaluation I

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

Precision/Recall Graph



EARIA 2006

© CvR

Evaluation II

Underlying conjoint structure
mapped to
numerical representation

Krantz, D., et al, Foundations of Measurement: Additive and Polynomial Representations, 1971.

Evaluation III

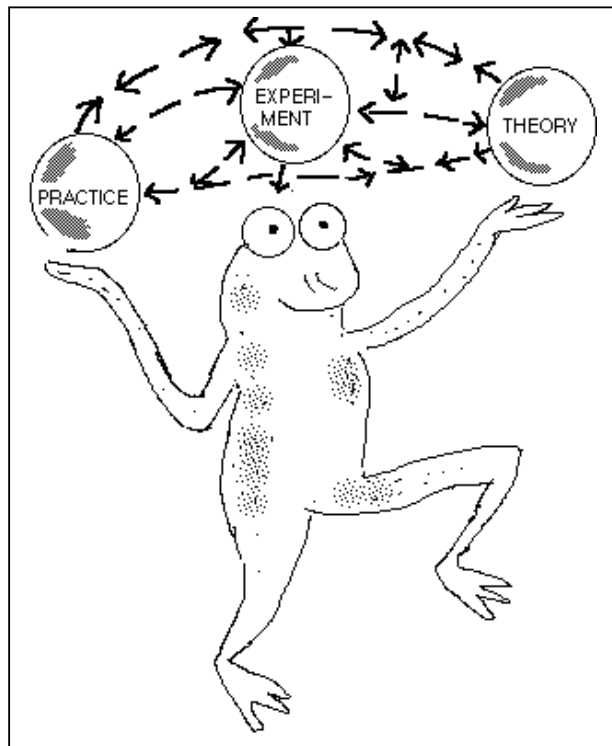
Construction of E & F, why?

Interpolation, why?

Significance testing requires knowledge of
underlying distributions

Swets, example

Swets J. A., Effectiveness of Information Retrieval methods, BBN
Report, 1967.



Theories - Tools - Algorithms

Do we need theories?
Top-down versus bottom up
Principled theories (Einstein)
Science of the Artificial

Jardine, N., Algorithms, methods and models in the simplification of complex data, The Computer Journal, 1970.

Theory

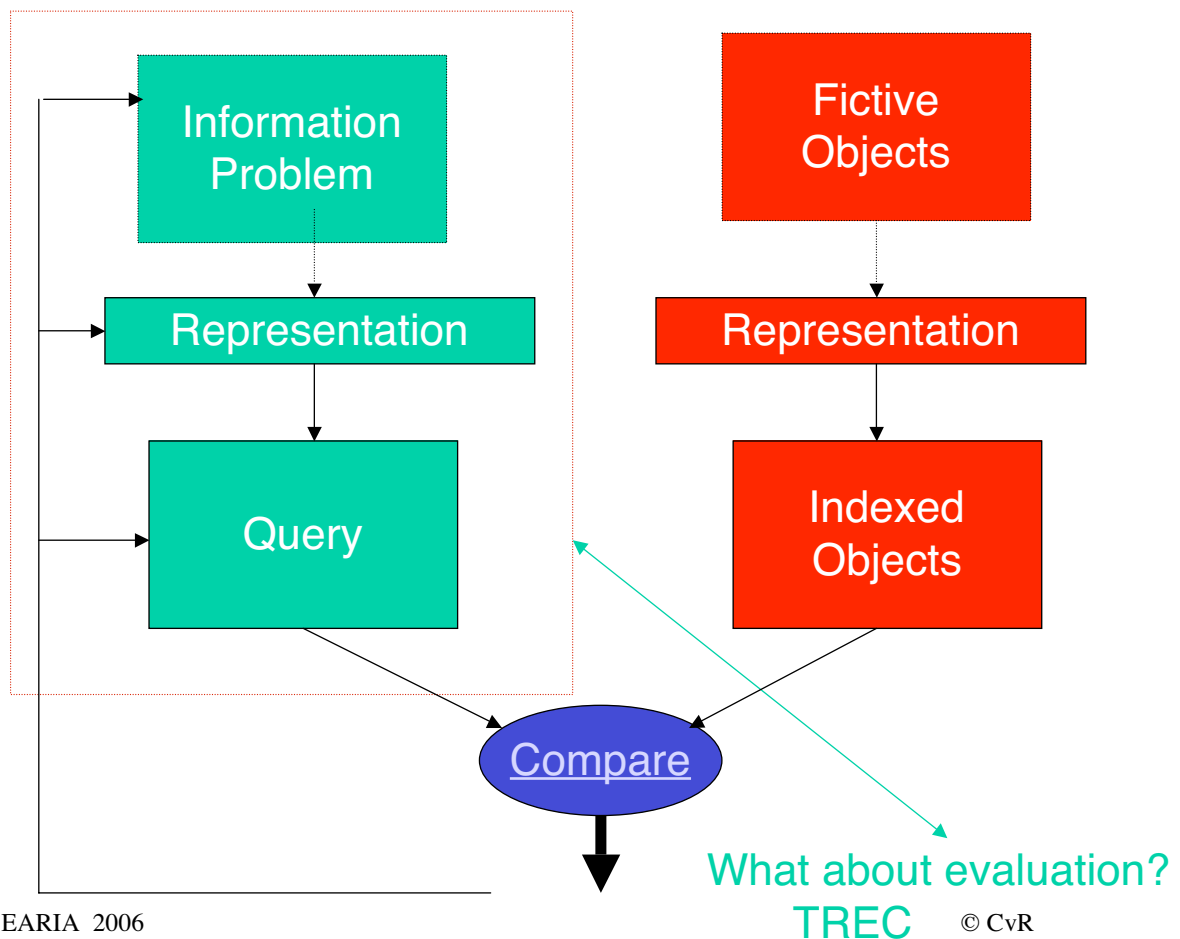
Knob twiddling
Data fusion
Authority/importance models
Logic + Uncertainty models eg QL
Filtering/Routing
Language models
Summarisation
Discrimination/Representation
IR + DBMS (inc XML etc)
Clustering the web
Visualising the web
Living with single term queries
Living with no queries
Context

Theory (cont.)

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

EARIA 2006

© CvR



EARIA 2006

TREC © CvR

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**

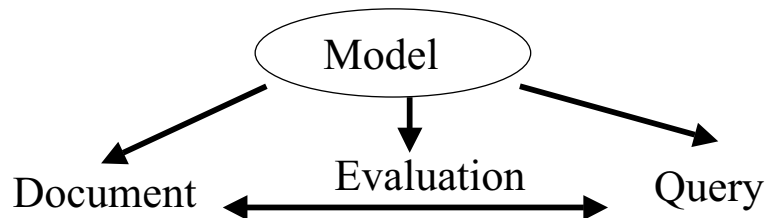
Inference

- Deduction/Induction: $A, A \rightarrow B$ infer B
- Cluster Hypothesis
- Association Hypothesis
- $P(\text{term}_1 | \text{term}_2)$

What is an IR Model ?

- An IR model explains the structure and processes of IR systems, and clarify their **general**, as opposed to *specific*, characteristics
- An IR model furnishes an answer for the **relevance decision mechanism**
- The IR model does not include the *cognitive aspects* of the retrieval aspects, such as query negotiation or output evaluation

IR Models

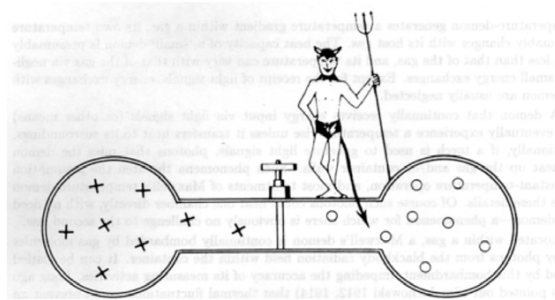
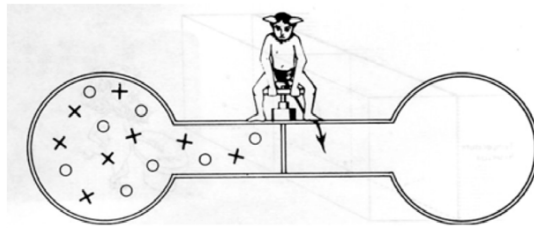


- There are many IR models
 - the relevance decision mechanism can be either **strict** or **flexible**
 - the representation of the data can have a **varying degree of abstraction**

Models

- Boolean
- Vector Space (metrics) - mixture of things
- Probabilistic (3 models)
- Logical (implication) - what kind of logic
- Language models
- Divergence from Randomness (Terrier)
- (Algebraic model): QL; LSI
- Cognitive (users): Context
- Language (distributions) - Bose-Einstein?

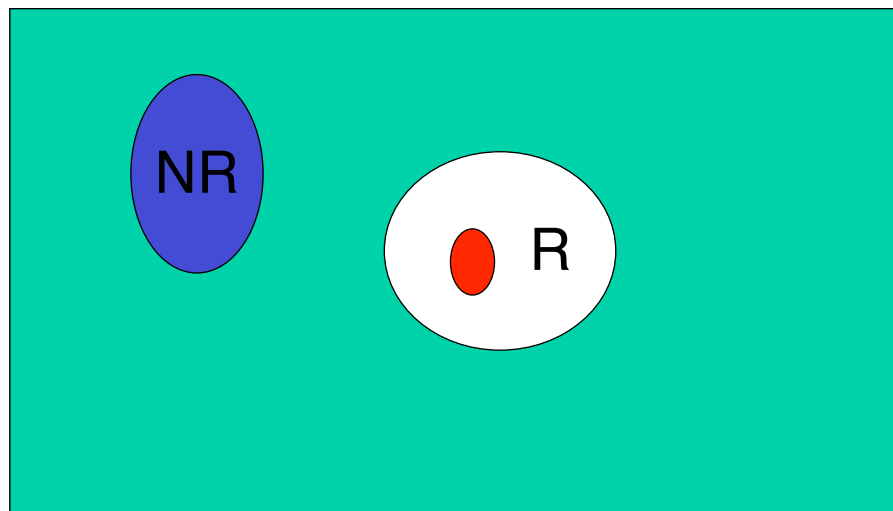
IR demon



EARIA 2006

© CvR

Partial Models



EARIA 2006

© CvR

Probabilistic Retrieval - Decision Theory

No queries

Optimality

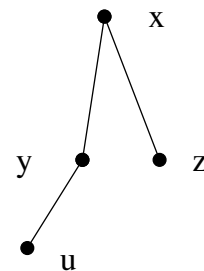
$(g(x) > 0 \rightarrow \text{ret, not ret})$

Nilsson, N.J., Learning Machines: Foundations of Trainable Pattern-Classifying Systems, 1965.

Independence - Dependence

$$P(A,B) = P(A)P(B)$$

$$p(x,y,z, u) \sim p(x)p(y|x)p(z|x)p(uly)$$



Optimise using EMIM

Chow, C.K., and Liu, C.N., Approximating discrete probability distributions with dependence trees, IEEE Trans on IT, 1968.

Yu, C.T, et al, A generalised term dependence model in information retrieval, IT: R&D., 1983.

2-Minute Probabilistic Model

R: Relevance

x_i : index terms $\underline{x} = (x_1 \dots x_k)$

$$P(Rx_1 \dots x_k) = P(R)P(x_1|R) \dots P(x_k | Rx_1 \dots x_{k-1})$$

$$P(x_1 \dots x_k R) = P(x_1)P(x_2|x_1) \dots P(R|x_1 \dots x_k)$$

$$P(R\underline{x})/P(\underline{x}R) = 1 = \frac{P(R)}{P(R|\underline{x})} \frac{P(x_1|R)}{P(x_1)} \times \dots \times \frac{P(x_k | Rx_1 \dots x_{k-1})}{P(x_k | x_1 \dots x_{k-1})}$$

$$P(x_i|x_j) = P(x_i) \quad P(x_i | Rx_j) = P(x_i|R)$$

$$P(x_i|x_j \dots) = P(x_i) \quad P(x_i | R \dots) = P(x_i|R)$$

$$P(R|\underline{x}) = P(R) \prod_i \{P(x_i|R) / P(x_i)\}$$

$$\log P(R|\underline{x}) = \sum_i \delta_{is} \log \{P(x_i|R) / P(x_i)\} + \log P(R)$$

Probabilistic Information Retrieval

$$\frac{P(R|x)}{P(\bar{R}|x)} = \frac{P(x|R)P(R)}{P(x|\bar{R})P(\bar{R})}$$

à

$$\log \frac{P(R|x)}{P(\bar{R}|x)} = \log \frac{p_1(x_1)}{q_1(x_1)} + \log \frac{p_2(x_2)}{q_2(x_2)} + \dots + \frac{p_n(x_n)}{q_n(x_n)} + \frac{P(R)}{P(\bar{R})}$$

$$P(x|R) = p_1(x_1)p_2(x_2)\dots p_n(x_n)$$

$$P(x|\bar{R}) = q_1(x_1)q_2(x_2)\dots q_n(x_n)$$

BAYES' DECISION RULE

$$[P(w_1|x) > P(w_2|x) \rightarrow \mathbf{x} \text{ is REL, } \mathbf{x} \text{ is NREL}]$$

$$P(\text{error}|\mathbf{x}) = \begin{cases} P(w_1|\mathbf{x}) & \text{if we decide } w_2 \\ P(w_2|\mathbf{x}) & \text{if we decide } w_1 \end{cases}$$

$$\begin{aligned} P(\text{error}) &= \sum_{\mathbf{x}} P(\text{error}|\mathbf{x}) P(\mathbf{x}) \\ &= \sum_{\mathbf{x}} \min [P(\mathbf{x}|w_1) P(w_1), P(\mathbf{x}|w_2) P(w_2)] \end{aligned}$$

PROBABILISTIC RETRIEVAL (Simple)

$$P(w_1|\mathbf{x}) > P(w_2|\mathbf{x}) \rightarrow \mathbf{x} \text{ is REL, } \mathbf{x} \text{ is NREL}$$

$$P(\mathbf{x}|w_1) P(w_1) > P(\mathbf{x}|w_2) P(w_2) \rightarrow \mathbf{x} \text{ is REL, } \mathbf{x} \text{ is NREL}$$

INDEPENDENCE ASSUMPTIONS

$$P(\mathbf{x}|w_1) = P(x_1|w_1) P(x_2|w_1) \dots \dots \dots P(x_n|w_1)$$

$$P(\mathbf{x}|w_2) = P(x_1|w_2) P(x_2|w_2) \dots \dots \dots P(x_n|w_2)$$

$$\text{Let } p_i = \text{Prob}(x_i = 1|w_1)$$

$$q_i = \text{Prob}(x_i = 1|w_2)$$

COSTS

l_{ij} = cost of deciding i when j is the case.

Expected Costs

$$R(w_1|\mathbf{x}) = l_{11} P(w_1|\mathbf{x}) + l_{12} P(w_2|\mathbf{x})$$

$$R(w_2|\mathbf{x}) = l_{21} P(w_1|\mathbf{x}) + l_{22} P(w_2|\mathbf{x})$$

Minimum Risk Decision

$$[R(w_1|\mathbf{x}) < R(w_2|\mathbf{x}) \rightarrow \mathbf{x} \text{ is REL, } \mathbf{x} \text{ is NREL}]$$

$$\begin{array}{l} \text{Typically } l_{11} = l_{22} = 0 \\ l_{12} = a \\ l_{21} = b \qquad \qquad a \neq b \end{array}$$

Overall Risk

$$R = \sum_{\mathbf{x}} \min(R(w_1|\mathbf{x}), R(w_2|\mathbf{x})) P(\mathbf{x})$$

$$P(\mathbf{x}|w_1) = \prod_{i=1}^n p_i^{x_i} (1 - p_i)^{1-x_i}$$

$$P(\mathbf{x}|w_2) = \prod_{i=1}^n q_i^{x_i} (1 - q_i)^{1-x_i}$$

$$P_{w_1}(0,1,1,0,0,1) = (1 - p_1)p_2 p_3 (1 - p_4)(1 - p_5)p_6$$

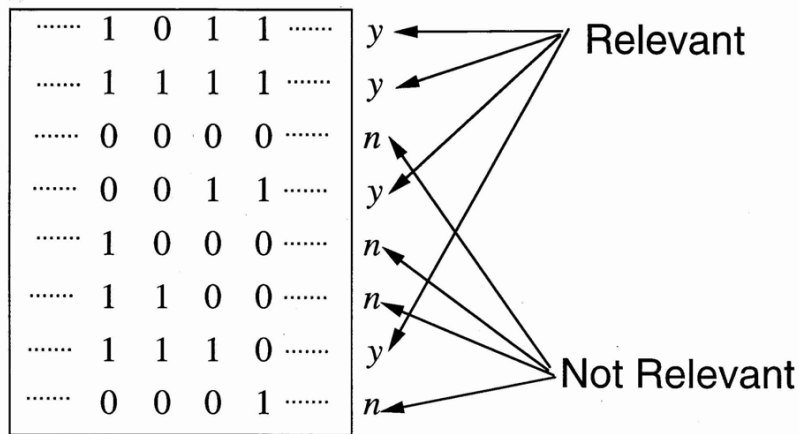
$$P(\mathbf{x}|w_1)P(w_1) \geq P(\mathbf{x}|w_2)P(w_2)$$

$$\frac{P(\mathbf{x}|w_1)P(w_1)}{P(\mathbf{x}|w_2)P(w_2)} \geq 1 \quad \text{Take logs}$$

$$\log \frac{\prod_{i=1}^n p_i^{x_i} (1 - p_i)^{1-x_i}}{\prod_{i=1}^n q_i^{x_i} (1 - q_i)^{1-x_i}} + \log \frac{P(w_1)}{P(w_2)} \geq 0$$

$$\sum_{i=1}^n x_i \log \frac{p_i}{q_i} + (1 - x_i) \sum_{i=1}^n \log \frac{1 - p_i}{1 - q_i} + \log \frac{P(w_1)}{P(w_2)}$$

$$\sum_{i=1}^n x_i \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)} + \sum_{i=1}^n \log \frac{1 - p_i}{1 - q_i} + \log \frac{P(w_1)}{P(w_2)}$$



$$P(x_i = 1|R) = p_i \quad \hat{a}$$

$$P(x_i = 1|\bar{R}) = q_i$$

$$p_1 = \frac{2}{4} \quad p_2 = \frac{2}{4} \quad p_3 = \frac{4}{4} \quad p_4 = \frac{3}{4}$$

$$q_1 = \frac{2}{4} \quad q_2 = \frac{1}{4} \quad q_3 = \frac{0}{4} \quad q_4 = \frac{1}{4}$$



$$\log \frac{1 - \frac{3}{4}}{1 - \frac{2}{4}} + \log \frac{\frac{2}{4}}{\frac{1}{4}} + \log \frac{\frac{4}{4}}{\frac{0}{4}} + \log \frac{\frac{3}{4}}{\frac{1}{4}}$$

EARIA

© CvR

PROBABILITY RANKING PRINCIPLE

If a reference retrieval system's response to each request is a ranking of the documents in order of decreasing probability of relevance than that is optimal. **WHY?**

Consider the set B for any cut-off

$E = \sum P(A \mathbf{x})$	\mathbf{x} is summed over the set B
------------------------------	--

a *maximum* compared with any other set B obtained differently.

Hence $P(A | B)$, $P(B | A)$ will be maximised at any cut-off.

A : relevant documents B : retrieved documents

Precision : $P(A | B)$ Recall : $P(B | A)$

Expected number of relevant documents in a set B :

$$\sum P(A | \mathbf{x}) = E \Rightarrow \text{Precision} = E / |B|$$

$$\text{Recall} = E / |A|$$

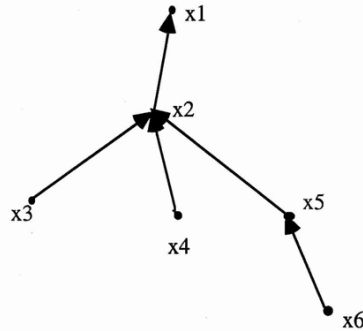
EARIA

vR

Dependence

$$P(\underline{x}) = P(x_1)P(x_2|x_1) \prod_{i=2}^n P(x_i|x_1x_2\cdots x_{i-1})$$

$$P_i(\underline{x}) = \prod_{m=1}^n P(x_m | x_{m_{(i)}}) \quad 0 \leq j(i) \leq i$$



EARL $P(\underline{x}) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2)P(x_5|x_2)P(x_6|x_5)$ © CvR

$$I(P, P_a) = \sum_{\underline{x}} P(\underline{x}) \log \frac{P(\underline{x})}{P_a(\underline{x})}$$

$$I(x_i, x_j) = \sum P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

$$\sum_{i=1}^n I(x_i, x_{j(i)}) \quad \text{find tree that maximises this!}$$

This MST gives P_a

Discrimination Gain Hypothesis

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.

$$P(x_i, x_j | w_1) = P(x_i | w_1)P(x_j | w_1)$$

$$P(x_i, x_j | w_2) = P(x_i | w_2)P(x_j | w_2)$$

$$I(x_i, x_j) = \sum P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

$$I(x_i, w) = \sum P(x_i, w) \log \frac{P(x_i, w)}{P(x_i)P(w)}$$

EARIA 2006

$$I(x_i, x_j) \leq I(x_i, w) \text{ and } I(x_i, x_j) \leq I(x_j, w)$$

© CvR

Clustering

(Dis)similarity measurement

Combination of scales

Mapping DC -> UDC

Measurement of fit

Dynamic clustering

Sneath, P.H.A, and Sokal, R.R., Numerical Taxonomy, 1973.

Cormack, R.M., A review of classification, The Jnl Roy Stats Soc, 1971.

Cluster Hypothesis

‘Cluster-based retrieval has as its foundation a hypothesis, the *cluster hypothesis*, which states that closely associated documents tend to be relevant to the same requests’ CvR, 1971

Voorhees, E.M., The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval, PhD thesis, 1985.

Dissimilarity/Similarity

$$d(x,y) \geq 0 \text{ for all } x,y$$

$$d(x,x) = 0 \text{ for all } x$$

$$d(x,y) = d(y,x)$$

$$d(x,y) \leq d(x,z) + d(z,y)$$

$$\{d(x,y) \leq \max [d(x,z), d(z,y)]\}$$

Information-theoretic approach I

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1|x)}{P(H_2|x)} - \log \frac{P(H_1)}{P(H_2)}$$

$$I(1:2) = \int f_1(x) \frac{f_1(x)}{f_2(x)} dx$$

$$J(1,2) = I(1:2) + I(2:1)$$

Information-theoretic approach II

$$\begin{aligned} & K(1,2|w_1, w_2) \\ &= \frac{1}{w_1 + w_2} \int w_1 f_1 \log \frac{f_1(w_1 + w_2)}{w_1 f_1 + w_2 f_2} + w_2 f_2 \log \frac{f_2(w_1 + w_2)}{w_1 f_1 + w_2 f_2} dx \end{aligned}$$

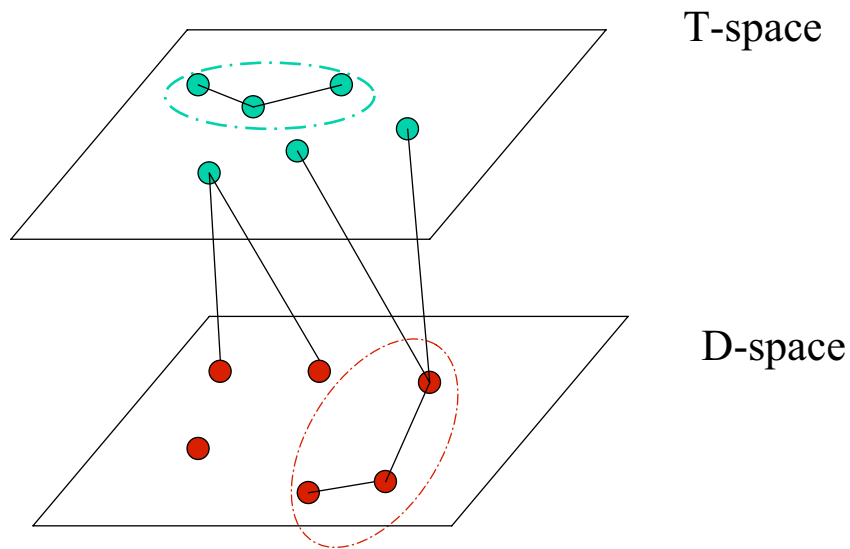
$$K(1,2) = 0 \text{ if } w_1 = 0 \text{ or } w_2 = 0$$

$$K(1,2) \geq 0$$

$$K(1,2) = K(2,1)$$

$$K(1,2|w_1, w_2) = K(1,2|cw_1, cw_2)$$

Navigation - Browsing



Duality is the key.

EARIA 2006

Class definition!
© CvR

Static Clustering

1. dependence on rank-ordering of dissimilarity
2. insensitive to small errors in DC
3. preservation of well marked clusters
4. stable under growth
5. labelling independence
6. invariance of ultrametric
7. subject to 3 minimises distortion

B	.1				B	.1					
C	.4	.2			C	.3	.3				
D	.3	.3	.3			D	.3	.3	.2		
E	.4	.4	.4	.1			E	.3	.3	.2	.2
A	B	C	D			A	B	C	D	E	

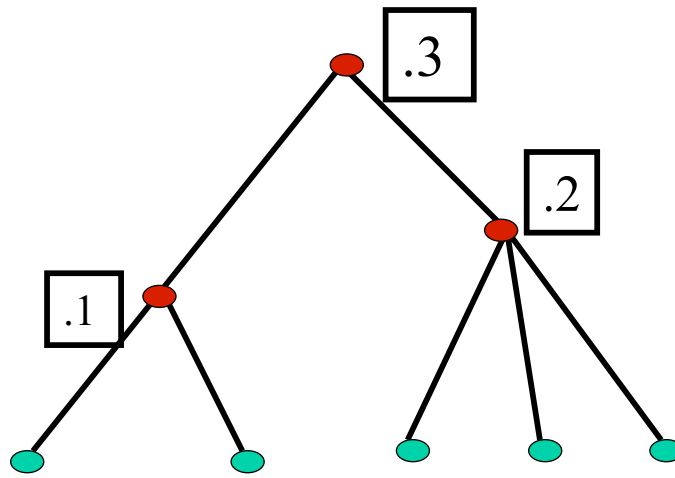
T

→

EARIA 2006

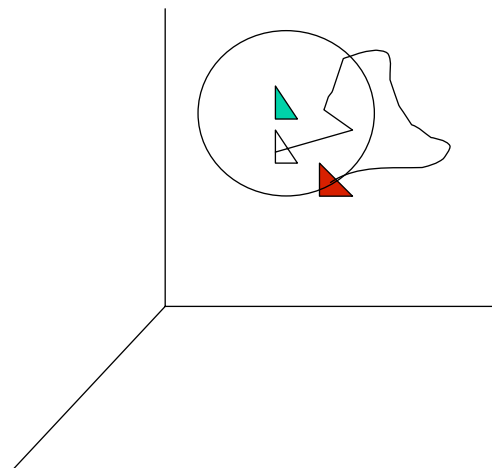
© CvR

Dendrogram



Spanning tree?

Dynamic Clustering



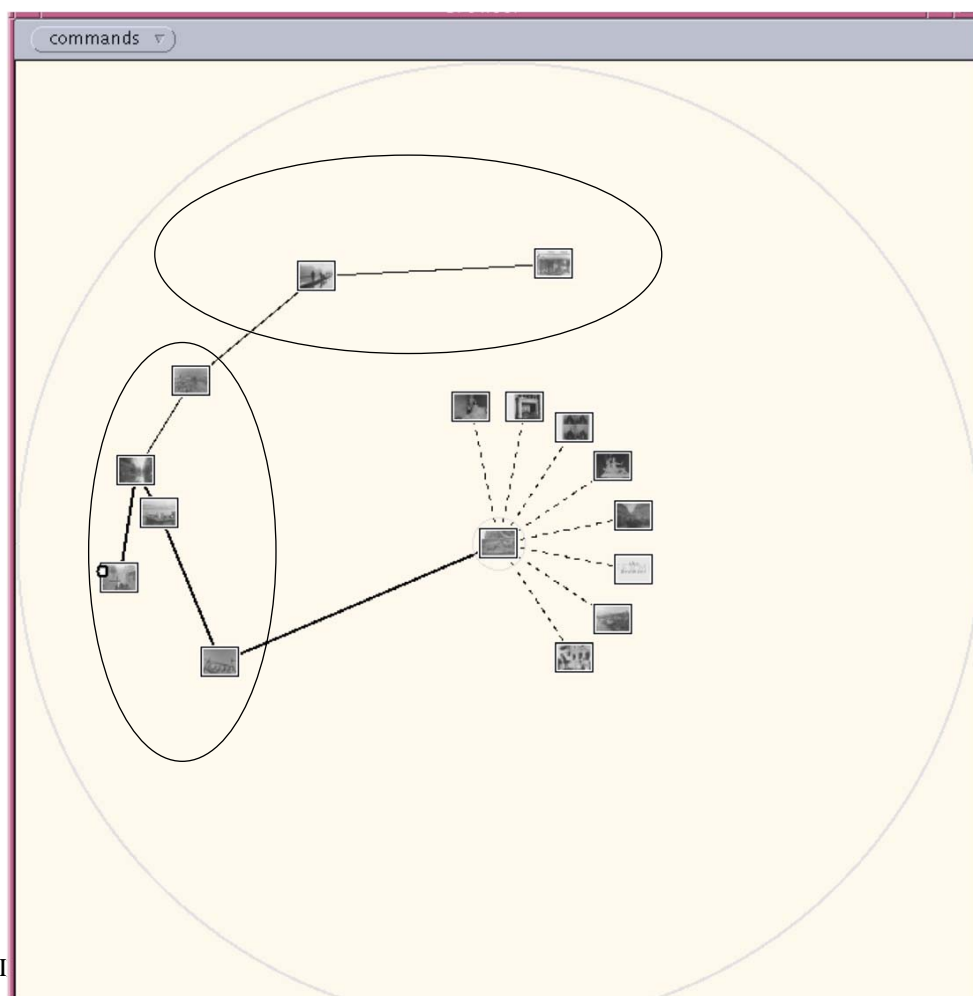
Hilbert-Schmidt: $(A,B) = \text{trace}(A'B)$

Applications

- Image Retrieval
- Web Retrieval

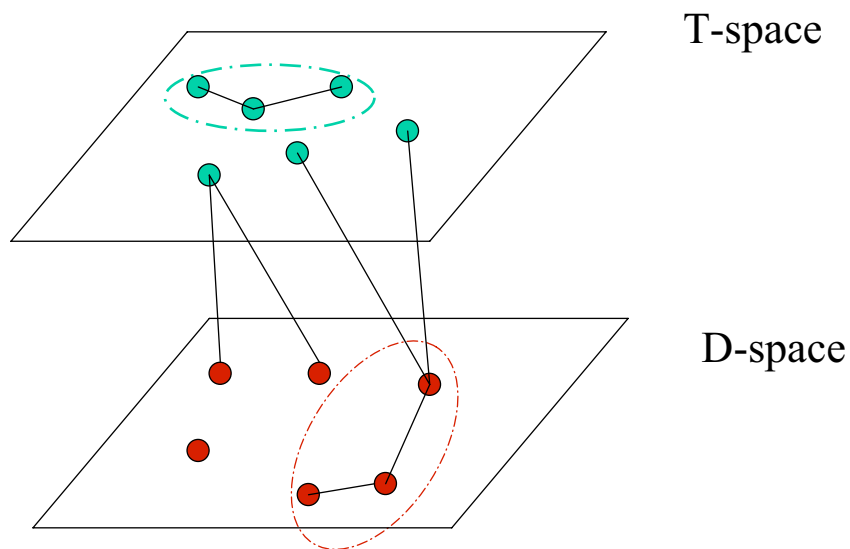
EARIA 2006

© CvR



EARI

Navigation - Browsing



EARIA 2006

© CvR

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
- SIS

EARIA 2006

© CvR

‘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

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: semantic gap
- 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

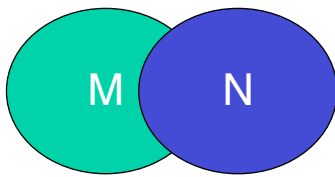
- ?
- ?
- object recognition
- geometry
- eyetracking/EEG
- path dependent
- how?

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

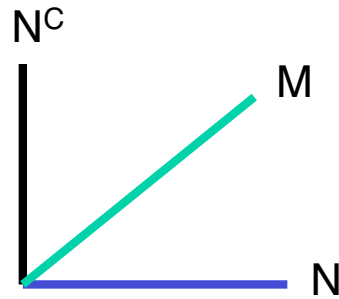
$A \rightarrow B, B \rightarrow C \text{ infer } A \rightarrow C$



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

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

EARIA 2006



$$M \otimes (N^c \oplus N) = M$$

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

© CvR

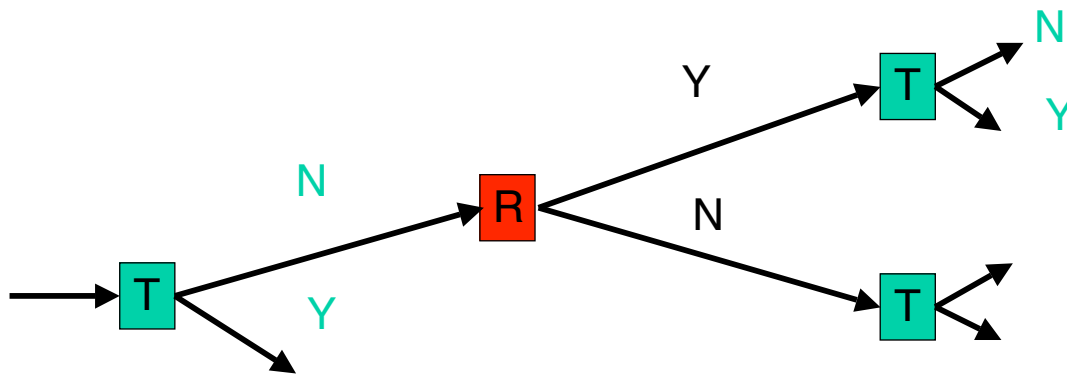
Interaction (Aboutness)

Objects: documents, queries \longrightarrow Relevance

Model

Observable(States) \longrightarrow ??

Relevance/Aboutness
is
Interaction/User dependent



EARIA 2006

© CvR

Where are we now in IR?

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

EARIA 2006

© CvR

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

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.

The Turn, Ingwersen and Jarvelin.

IF THERE'S ONE THING MORE BORING THAN
YOUR HOLIDAY VIDEOS NEVILLE IT'S YOU
BANGING ON ABOUT YOUR INDEXING SYSTEM

