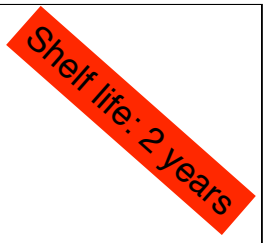


Introduction to Information Retrieval

(SSMS 2006)

C.J. "Keith" van Rijsbergen
(with help from Iadh Ounis &Joemon Jose)
Computing Science
Glasgow University

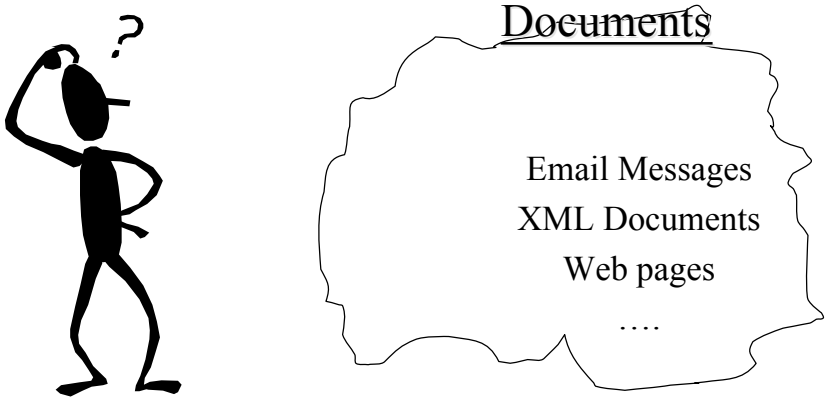
During these lectures I will give an overview of IR. The history of IR is long and fraught. For many years it was unclear whether it was a subject at all, then when it became a subject it was claimed by both Information Science and Computer Science. Although in the early days 50's and 60's this was responsible for a number of frustrations, for example the unwillingness of librarians to accept hard experimental results, it now is also one of its strengths. We interact fruitfully, the IS community guarding us against technological, or system-based excesses, the CS community representing a hard-nosed approach to experimental designs and being forced into taking user-interface issues seriously. A marriage made in heaven!

Information retrieval has in the last few years become a federation of sub disciplines: text mining, image retrieval, web retrieval, information extraction, data mining video retrieval, etc. Each one of these sub-disciplines has it s own mission and would probably claim to be distinct from tradition IR, however their roots in IR are always very apparent.

One of the the outstanding characteristic of IR as a scientific enterprise is that it has a very strong experimental methodology. There is a strong requirement to test and evaluate retrieval performance mostly under laboratory conditions, and sometimes in real situations with real users.

There is always, in pursuing the subject of IR an interesting interplay between theory, experiment, and practice. This has been so pretty well from the beginning since the fifties.

Scenarios & Applications

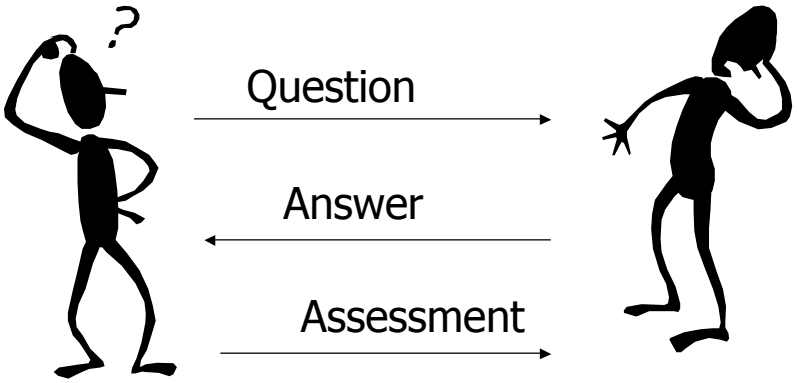


SSMS 2006 © CvR

Initially mechanized document retrieval systems were designed as replacements for traditional manual catalogues giving access to repositories of papers and books. Thus there was a great emphasis on text retrieval and the individual unit of retrieval was a document in the conventional sense of the word. Now things are quite different. With retrieval from different kinds of media and with the retrieval of sub-parts of documents, as in XML retrieval, the word 'document' is used to designate a unit of retrieval which might be an image, a video clip, an audio track, a subsection of a text document, etc. This does not matter, a unit of retrieval is usually represented in such a way that its original nature does not play much of a role during retrieval except at the interface with the user.

Retrieval takes place in different contexts, and as part of a task that a user in is engaged in. These scenarios can influence the choice of retrieval strategy materially. Also, IR is applied in differents domains. Retrieval from a repository of patents may be handled quite differently from the retrieval from a news wire.

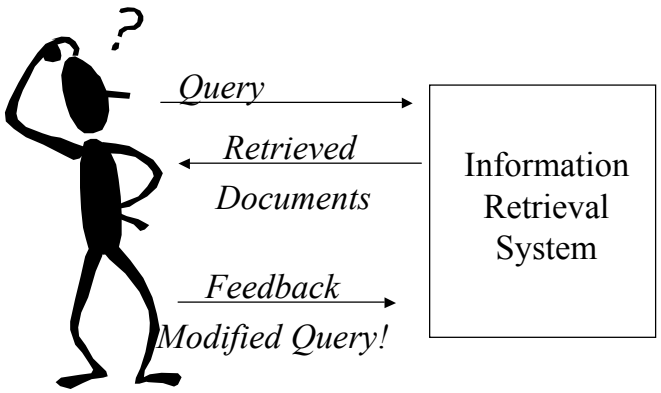
Retrieval – A Question-answer scenario



SSMS 2006 © CvR

A typical set-up for is somewhat similar to a dialogue between a user and an expert, maybe a librarian. A user comes along with an Information Need which he/she attempts to explain to an expert. Usually this is done in the form of a query or request for information. The expert replies with some information, or points at a document, that is expected to satisfy the information need of the user. The user, if satisfied goes away, if not, the user explains to the expert why the information is not adequate and may address a revsied query to the expert; and so they may go round again.

Retrieval Loop



SSMS 2006 © CvR

In the case of a mechanised system the set-up the situation is not very different, except that the expert is now replaced with an IR systems. The user interacts with the system through a highly contrived interface. A user may put in a natural language query, but as in Google, the syntax of the statement is largely ignored. Once a query has been entered, the system will transform it so that it can be handled mechanically by the IR system. The system will return a set of documents, perhaps in the form of a ranking, or organised in some other way, in the expectation that the set will contain mostly relevant documents, but as few non-relevant ones as possible. Ideally, a user will feedback an assessment of each one of the retrieved documents, or some proportion of the retrieved set. The system may then automatically modify the query or may assist the user in generating a modified query. After that a further retrieval run will be initiated and the retrieval loop repeats.

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?

In the mind of the public IR is virtually synonymous with the popular search engines. For simple retrieva,l such engines can be quite effective, but we have all experienced their inadequacies. The underlying retrieval mechanism may be rather crude, or based on the wrong model. Feedback is rarely implemented. Nevertheless, these engines have been commercially succeessful because of the business model not necessarily because of the quality of retrieval. In some cases attempts are made to show how the retrieval is accomplished, but more often than not the retrieval strategy remain hidden.

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 ….

In the next slide I give some standard definitions of IR. In general one assumes that there is little structure in the data can be used in contrast to databases where SQL queries make heavy use of the structure. Even if there is structure like in NLP or XML its use is extremely varied, and thus is not used in a uniform way.

In IR there is still an emphasis in text retrieval, this is simply because text still dominates information output. Moreover, in the case of other media their retrieval is often enhanced by the sue of text such as annotations.

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

Salton is one of the pioneers of the filed. His 1968 textbook, Automatic Information Organization and Retrieval, McGraw-Hill, is a classic. It continues to be a good source of ideas. In the late sixties and early seventies Salton had a sequence of graduaate students whose groun-breaking work continues to used to this day, e.g. Rocchio.

Needham (in 1961) was possibly the first person to do a PhD thesis in IR in computer science. The definition above was written for a dictionary. Needham had a significant impact on the field through his influence on Sparck Jones. The collaborated extensively in the early days, Needham changed fields subsequently but Sparck Jones continues to so research in IR to this day

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

SSMS 2006

This slide and the next gives a time-line for the development of IR. Not mentioned here are the pre-cursors of the subject like Vannevar Bush, Robert Fairthorne, and Emanuel Goldberg. These are important researchers but only of historical interest now. One can find detailed information about each one on the Web.

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

I have taken the time-line through to 1998. Of course much has happened in the last eight years. For example, most of the search engines for searching the web have established themselves in this period. Progress in theory and experimentation has slowed somewhat. A book by Voorhees and Harman on TREC has just been published which is an excellent retrospective on the TREC intitiative. I myself have published a book on the Geometry of IR which attempts to define a logico-algebraic framework for retrieval models.

Experimental Methodology

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

Here is a role of honour for some significant mile-stones in experimental methodology for IR. The approach adopted in IR was pioneered by Cyril Cleverdon working at Cranfield in the UK. In honour of his work the approach to evaluation is frequently called the Cranfield paradigm. So what is it? It consists of identifying a collection of documents from which will be retrieved. A set of queries is also identified, and for each query it is determined in advance which documents are relevant and which are not. Thus one has three sets:

1. Documents
2. Queries
3. Relevance assessments

Together these three sets make up a *test collection*. For example each year TREC distributes a some test collections; this year it includes a set of blogs. These data are then used to evalaute novel retrieval strategies world-wide.

Evaluation

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

SSMS 2006 © CvR

One of the cornerstones of IR is the set of parameters that are used for evaluating the quality of retrieval performance. The most commonly used parameters (you can find others in my book on IR) are precision and recall.

Precision is a measure of the proportion of relevant documents in the retrieved set, whereas recall is a measure of the proportion of relevant documents retrieved.

A = set of relevant documents

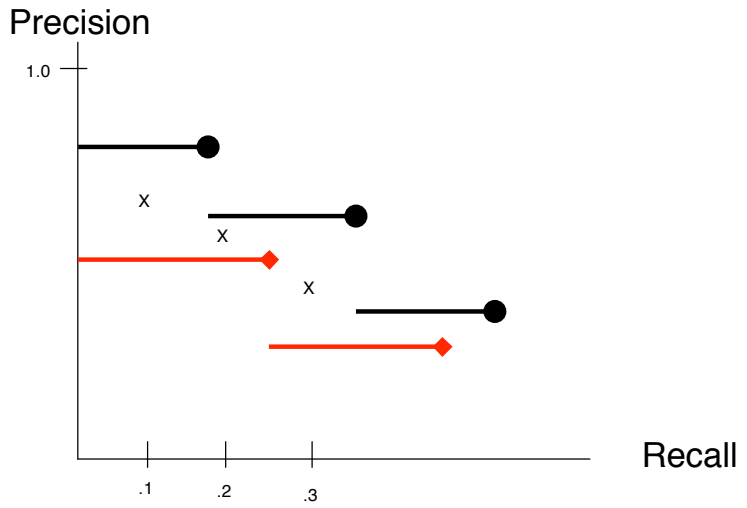
B = set of retrieved documents

$P = |A \cap B| / |B| = P(A|B)$

$R = |A \cap B| / |A| = P(B|A)$

These seemingly innocent parameters disguise a huge set of technical problems in use. One of the characteristics of the sets of results presented in terms of precision/recall is that there is a trade-off between the two. That is, high recall implies low precision and vice versa.

Precision/Recall Graph



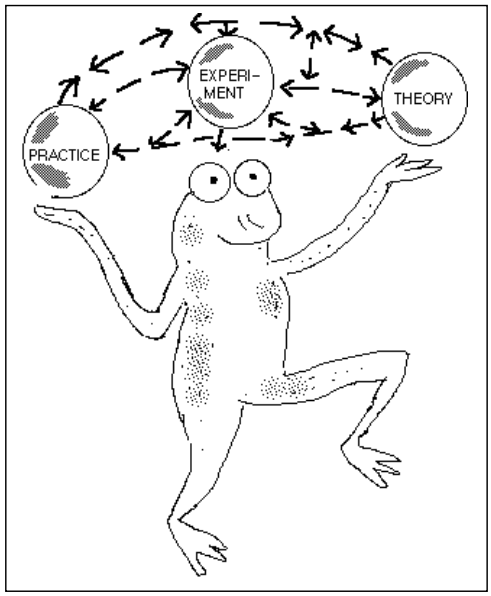
SSMS 2006 © CvR

Each query produces a precision/recall graph. A graph is made up of a set of discrete points, each point represents a precision/recall reading at the point where the number of relevant documents retrieved increases by one. Issues arise over interpolating between points, and averaging bewteen curves. Interpolation is done by fitting a step function which makes macro-evaluation rather straight-forward.

Some meta thoughts

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

I distilled these thoughts after completing the slides and notes for this talk. It seems to me that it is possible to characterise the IR viewpoint in a number of ways. To begin with no a priori assumptions are made about structure or process, unless given by the raw data or external constraints. This is most obvious when it comes to classifications, these are intended to reflect the inherent structure in the data and are not imposed. When it comes to features/attributes, relevance, or aboutness a categorical view is not always taken, that is, a document is not either relevant or not-relevant, a document is not either about X or not about X, etc. Processes in IR are usually adaptive making them user-driven and context dependent, this is particular evident in relevance feedback. The semantics of objects are defined by the data, in other words it is the distribution both within a document and across documents that give the "meaning" of terms. IR on the whole makes no claims about Knowledge we tend to work with notion of Information and as such consider the probability of propositions to be indefinitely revisable in the light of the weight of evidence (this is an issue in the Bayesian context when $P(X) = 1$). Following from this we tend to work with contingent truths rather than necessary truth, and of course this effects the kind of logics we are interested in. Finally, a trend that has emerged in the last few years is that interactions with IR systems is based on ostensive manipulation and definition, that is, systems react to what a user does, or points to, not only to what the user says or writes.



SSMS 2006 © CvR

For years I have advocated the interplay of theory, practice, and experiment. My first serious attempt to talk about this was probably in a seminar presentation I gave in 1977 where I quoted the following from Freud:

"…., I think….that the great problems of the universe and of science have the first claim on our interest. But it is as a rule of very little use to form an express intention of devoting oneself to research into this or that great problem. One is then often at a loss to know the first step to take. It is more promising in scientific work to attack whatever is immediately before one and offers an opportunity for research. If one does so really thoroughly and without prejudice or preconception, and if one has luck, then since everything is related to everything, including small things to great, one may gain access even from such unpretentious work to a study of the great problems"

I still largely agree with this slogan, or motto. Curiously I would claim that considerable progress in IR has been made precisely because IR researchers took seriously the solving of "whatever is immediately before" us. The theoretical models and breakthroughs largely arose out of detailed experimentation and new models sometimes arose out of the failure of existing models to deliver the anticipated experimental performance.

During my 1977 talk, Robert Fairthorne, one the pioneers of IR was in the audience, and clearly taken with my three way balancing act drew this cartoon.

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

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

SSMS 2006 © CvR

Let me say a little more about these three disparate activities in IR. As you can see I list the Web as a major practical example. This is because a huge amount of operational retrieval using the web takes place, and a lot of it is woeful. A major practical challenge for IR is to influence the design of search engines so that retrieval performance goes beyond what you get by just submitting a 2.4 word query. In electronic publishing, as pursued by the large publishers for example, much multimedia data is conveniently made available but unfortunately the search capabilities are mostly inadequate. Commerce seems have discover the knowledge economy and so data mining and knowledge discovery are the flavour of the months. Of course there is a long history in IR using statistical techniques to model significance and dependence. If one thinks about the provision of materials for distance learning whether they be text, image or graphics, then once large repositories of such information becomes available a major issue will be its retrieval.

There is a long and honourable tradition of experimental work in IR. Cyril Cleverdon one of the pioneers, together with Jack Mills and Michael Keen produced a series of reports, initially the Cranfield I (1960) study followed by a more substantial study in 1966, Factors determining the performance of indexing systems. These projects can claim to be responsible for founding the experimental approach that is now know as the “Cranfield Paradigm”, it to do this day continues in the extremely successful series of experiments known as TREC (see trec.nist.gov).

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

SSMS 2006 © CvR

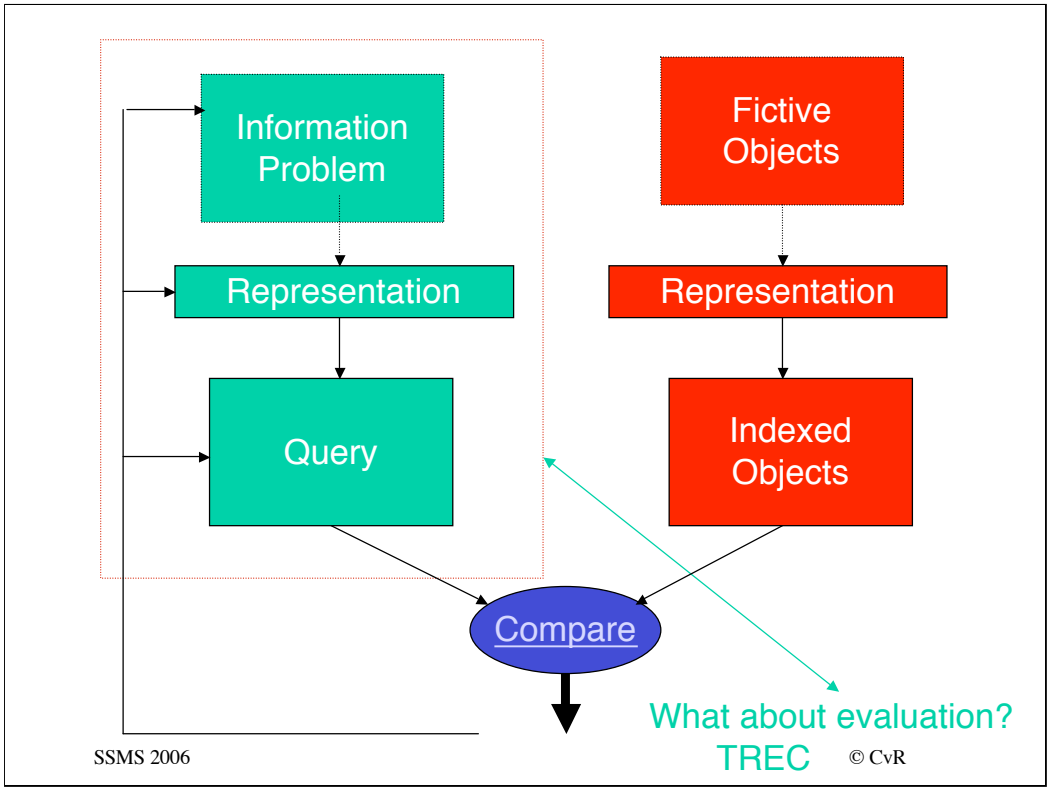
Much theory in IR has come about through "knob twiddling", this generally means adjusting a set of parameters for a give retrieval model and observing the effect on retrieval performance. Of course this can lead to mindless experimentation but it has also led to new variants of statistical models. Dissatisfaction with a given model, often because of poor retrieval, has led to proposals for new models embodying such disparate approaches as Bayesian Inference, Clustering, Non-classical Logic, Dempster Shafer Theory of Evidence, etc.

Considerable theoretical work has also gone into the design of evaluation measures, that is, ways to mathematically, represent retrieval effectiveness, to average it, and to establish statistical significance. Ever since the time of Cleverdon Precision and Recall have been favoured. Unfortunately recall is not always readily available, think of retrieval from the Web, nor is precision always appropriate in dynamic task-oriented environments. Nevertheless future experimenters should take note of the approach to experimentation in IR. A classic summary of the IR approach can be found in the collection of papers edited by Sparck Jones, "Information retrieval experiment", Butterworth, 1981.

Theory (cont.)

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

A few words about scale free networks. This is an area that burst on the scene through the work of several people working on 'small worlds', 'seven degrees of seperation'. These networks are such that an average distance between nodes does not make much sense. One can view the web as a scale free network. There is much theoretical and popular work available fort hese structures see the book by Barabasi, *Linked*, and the book by Watts, *Small Worlds*. Exploitation of this kind of theory for web retrieval is only in its infancy.



This slide represents the traditional view of IR

But consider what happens if 'green' side is replaced with by a form interaction that does not usue language, for example the usual simply points at objects displayed at the interface. In particular how does one construct a test collection when there are no queries?

Comparing IR to Databases

	Databases	IR
Data	**Structured**	**Unstructured**
Fields	**Clear semantics** (SSN, age)	**No fields** (other than text)
Queries	**Defined** (relational algebra, SQL)	**Free text** ("natural language"), Boolean
Recoverability	**Critical** (concurrency control, recovery, atomic operations)	**Downplayed**, though still an issue
Matching	**Exact** (results are always correct)	**Imprecise** (need to measure effectiveness)

Before examining in more detail characteristics of IR research, it may be worth spending a little time contrasting IR with databases. In the slide a contrast in made in terms of five dimensions. Clearly the comparison is explicitly extreme. Many DB systems would claim to have IR features, and vice versa.

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

SSMS 2006 © CvR

I originally (1979) designed this table as a way of comparing databases with information retrieval, however over time this comparison has become more generic. The differences between DB and IR have become less marked. I now view this table as a way of focussing attention on a number of salient dimensions that span research in in areas such as IR, DB, data-mining, knowledge discovery etc. It enables me discuss IR research in a limited and constrained way without taking on the whole subject. In what follows I will address each one of these dimensions and describe where we are with research in that area. For a more recent discussion of this table in terms of data and document retrieval I recommend David Blair's book, Language and Representation in Information Retrieval, Elsevier, 1990.

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

SSMS 2006 © CvR

Fundamental to any retrieval operation is the notion of matching. One can track the progress in IR in terms of the increased sophistication of the matching function. Typically these functions are the consequence of a model of retrieval. For example the Boolean matching, and the logical uncertainty principle both presuppose and elementary model and proof theory from formal logic. In the case of the LUP an assumption is made about how to measure partial entailment. There are three major IR models, vector-space, probabilistic, and logical. Each has its corresponding matching function. Optimality criteria come into play in deriving these functions, sometimes related to performance (PRP), sometimes related to minimal change (LUP). These functions do not necessarily presuppose a representation mechanism, so objects may be represented by absence/presence of index terms, or indeed may involve frequency data related to index term distributions within a document or over the entire collection.

One of the difficulties in extending this type of matching to web data is that frequency data may not be available - we operate in an open world rather than a closed one. Another problem is associated with extending these functions to image matching. Right now it is fashionable to invent ontologies for representing and describing web documents, it is difficult to see how to combine the results of "ontology matching" or inference with the standard IR forms of matching. Of course different search engines use different matching functions and combining the results of those is a problem in its own right.

Inference

- Deduction/Induction: A, A→B infer B
- Cluster Hypothesis
- Association Hypothesis
- P(term$_1$|term$_2$)

SSMS 2006 © CvR

The major kind of inference that is used in IR is inductive (and sometimes abductive), that is a weight of evidence calculation is done to support a hypothesis or its alternative. One could characterise this by saying that what is important here is to be able to execute intelligent guessing. So for example, modus ponens is usually subject to degrees of uncertainty, where A and A -> B is known only with a probability and we "guess" at the probability of B. The kind of inference common in ontology based reasoning does not allow for these uncertainty. This raises the spectre of the debate about controlled versus uncontrolled vocabularies that took place in IR many years ago.

Some of the inductive inferences in IR are based on assumptions about the associations between descriptors/attributes used to represent objects. A frequent assumption is that attributes are probabilistically independent. This means that one attribute does not contain any information about another. The reverse assumption can be made, that is, that knowing something about one attribute e.g. it features prominently in relevant documents, that a closely associated attribute will also be a good indication of relevance. IR has found a number of ways of exploiting this. In the past it has been difficult to do so because of the small data sets available to estimate importance of attributes. Within the context of the web this is not a problem. This approach could be extremely useful if users only type very short queries and techniques are needed to (automatically) extend such queries.

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

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

SSMS 2006 © CvR

A longstanding inductive hypothesis is the Cluster Hypothesis. This was originally formulated to justify the use of automatic classification, or clustering, of documents. I originally formulated this hypothesis as, *closely associated documents tend to be relevant to the same requests*. Much experimentation his since gone into establishing empirical evidence for and against it (see the extensive work by Voorhees). One way to look at this hypothesis is that it is a reflection of the operation of a "hidden variable", relevance with is the common cause for the behaviour of X and Y. For example, under the Reichenbach formulation of common cause we get that P(X,Y|rel) = P(X|rel)P(Y|rel) and P(X,Y|nonrel) = P(X|nonrel)P(Y|nonrel), etc. A hidden variable like that will generate a dependence between X and Y through the common cause rel/nonrel. Reichenbach insists on a further two conditions namely, P(X|rel) > P(X|nonrel) and P(Y|rel) > P(Y|nonrel). Having found rel/nonrel as the common cause one can use Bayes Theorem to invert the probabilities to compute the P(rel| .). I am merely illustrating the use of induction at the document level here. At an operational level one would need to find an algorithm to detect the common cause (hidden) variable. The answering here may be clustering. There is a long tradition of clustering in IR, some of the earliest work was done by Salton's students (especially Murray and Rocchio), Van Rijsbergen and Croft. This work lay unexploited until recently when the development in the web caused people to rethink the sue of data reduction and representation techniques (e.g Harper, Hearst, Pedersen).

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.

SSMS 2006 © CvR

Another inductive hypothesis is concerned with index terms. This seems to be a possible flip-side of the Cluster Hypothesis. In this hypothesis one is looking for dependence between attributes. It is a well known fact that the set of documents and the set of index terms can be viewed as dual spaces of each others. Thus in principle given enough information about the the index term space and the document space one can model retrieval in one or the other. A quantitative development of this is the discrimination gain hypothesis which I will discuss later. One comment I would make is that given the tendency of users searching the web to generate short queries, often a single term, one can see how an hypothesis like this might be exploited to help improve the query through a form of query expansion.

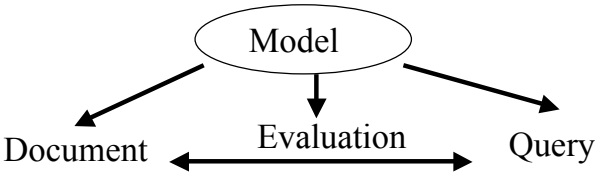
What is an IR Model ?

- 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

SSMS 2006 © CvR

The use of model in mathematics and physics is somewhat ambiguous. In science we have theories that account for phenomena, the observable structures and processes, which may postulate processes and structures not directly accessible to observation. Models are then commonly thought of as interpretations of such theories; this is the logical view. Another view is that models are a kind of a picture of the processes and structures under study. Most IR models fit the second kind of view, however, it remains controversial what the model is a picture *of*.

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**

The issue as to whether a relevance decision is strict or flexible is intimately connected with the kind of probabilistic model one can build for such a decision process. In general it is considered a binary decision and probability is attached to this binary event. Whether it can be modelled as an event is subject to debate. One can also consider relevance as a property. The level of abstraction usually depends on the media. For example, text is subject to less abstraction than image.

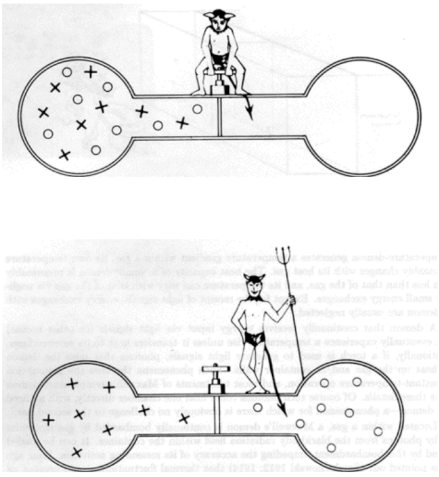
Models

- 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?

SSMS 2006

One of the interesting aspects of current search technology for the WWW is that it is almost model free, although one could claim that many search engines approximate some of the IR models more or less. This is not necessarily a bad thing to ensure that these engines work reliably and scaleably. Unfortunately to improve the effectiveness of such searches one will need to pay more attention to models of the process so that one can reason about it and make predictions. In IR there has been a steady development of such models. The first four I have already alluded to when describing different matching functions. The VS model is very dependent on the choice of inter document similarity/dissimilarity, that is it takes its structure from the 'metric' on the document space. The probabilistic model comes in various flavours, one is determined by the probability with which a term occurs in the relevant document, a second is based on an estimate of the probability with which a user would use a term to ask for a particular document,and then of course one could combine these (see Maron). It is curious that the difference between Objective and |Subjective probability is reflected here. Until recently statistical information about the occurrence of terms (tokens) in a document or over a collection of documents has been largely used in a heuristic manner. In the last few years elaborated stochastic process models have been proposed to represent the tokens within a document: retrieval is then determine by the probability with which a query is generated. In my view this is a development of the logical framework which attempts to give a semantics for P(d -> q).

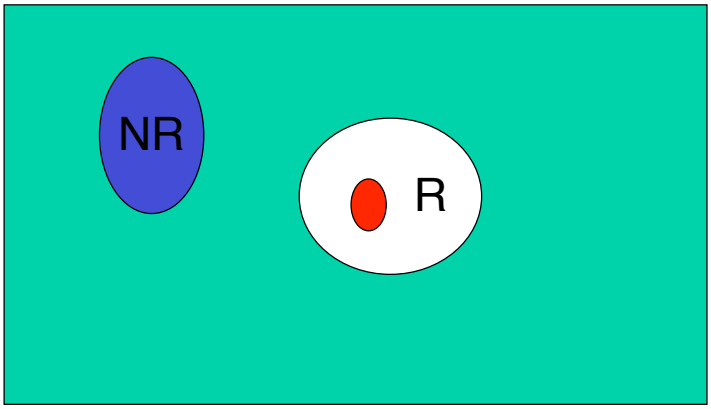
IR demon



SSMS 2006 © CvR

Think of Maxwell's demon in physics and you will see quickly what the IR demon does. The '+' and 'o''s represents two kinds of documents, relevant and non-relevant. The demon is an intelligent retrieval agent that retrieves with 100% precision and 100% recall. Retrieval can be seen as building a model for thsi demon.

Partial Models



SSMS 2006 © CvR

All this talk of models and modelling often leaves lost in levels of abstraction. Let me try and describe in a general way how models arise in IR. We make the assumption that at any moments in time, there are relevant documents (white) to found amongst the set of documents (green). Let us assume that by some means one can identify some of the relevant ones (red) and non-relevant ones (blue). This means that one has devised a way, maybe a decision function, to separate at least partially the relevant from the non-relevant ones. Most of the retrieval models are able to make this initial separation. Also, mostly this initial separation is not good enough. The grand challenge is to use the 'sample' information to adapt the way of separation to reflect the user's orientation so that the remaining relevant documents can be found. To this end the full strength of all the modelling: metrics, logics, stochastic processes, inference, etc come into play. This is similarly the case when the green set is the entire web.

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

SSMS 2006

At this point it may be of interest to give a little history about the use and development of automatic classification techniques for IR, especially since recently these techniques have found favour again for supporting browsing and for the generation of thesaurus classes. One of the earliest people to use clustering in IR, this may come as a surprise, was Roger Needham in Cambridge. Both he an Karen Sparck Jones worked in the Cambridge Language Research Unit, and of course Sparck Jones continued to work in classification publishing a book and a number of papers on the subject. Salton and his students did extensive work on document clustering and so that, the last piece of extensive work in the late seventies was done by Bruce Croft. Much of this early was influenced by the theoretical work carried out in Numerical Taxonomy (see Sneath and Sokal). The recent return of interest in clustering does seem to have picked up on the extensive theoretical work that was done in the sixties and eighties. This is particularly noticeable in the work that is proposing the Kullback-Leibler information as an asymmetric measure of similarity. For example, Jardine and Sibson and contains an extensive account of how to construct dissimilarity measures based on information-theoretic considerations. Similarly many of the theoretical properties of clustering methods, such as order independence, continuity, go unnoticed, which from a scientific point of view: reproducibility and reliability of experiments, are important.

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

In IR classification has always been seen as "classification for a purpose". The idea that one could define classification independent of other considerations has never been attractive. The Borges quote shows wonderfully how extreme these purposes might be. In the world of Ontologies there is an inclination to perceive classification as defined in absolute terms, for example one might write down the necessary and sufficient conditions for class membership to arrive at "natural kinds" (Hardegree), this would be an example of a monothetic classification, a polythetic approach would be less strict about class membership, in the latter case membership might depend on the number of shared attributes.

Query Language

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

SSMS 2006 © CvR

One of the main thrust of IR research has been to concentrate on natural language queries. This is in contrast to the work in databases which has been mostly concerned with artificial query construction such as SQL and QBE, although that has changed recently. This concentration on NL has led to a reasonable amount of work being devoted to generating processes that can take pieces of text and normalise them so that they can entire into a computational comparison/matching calculation leading a score which indicates degree of relevance. Furthermore, this quantitative approach scales and has worked effectively in IR (e.g. Porter stemmer). More recently this naïve approach to the semantics of NL has made it relatively easy to address multi- and cross- lingual approaches to IR, especially the problem of retrieving from a foreign language collection by means of the formulation of a query in a different language, that is, retrieving from French documents by putting, say, an English query. The "mathematico-statistical semantics" for text has to some extent transferred with obvious limitations to the retrieval of images, although there are no visual keywords (yet). The feedback loop kicked of by an initial query, transfer quite happily to the retrieval of images. What is especially interesting is the way one medium (text) can assist another (image), or vice versa, in retrieval.. (see Dunlop). There is also an approach called the ostensive approach which dispenses with queries altogether (see Campbell). This latter approach should be of great interest for browsing the web.

Query Definition

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

The assumption made in IR is that a query is always an incomplete specification of an information need, moreover, it is also assumed that at any one stage in a search a user's information need has only partially emerged. So although the very precise mathematical approach to representing a query leads one to think that information needs are mapped down onto mathematical structures once and for all, this is not so. The difficulty of eliciting information needs has led to a number of ways for overcoming it. One of the basic IR models assumes that the index terms are distributed independently, this is obviously not so, models have been created that attempt to capture arbitrary dependence between terms thereby representing information needs more accurately e.g. money/bank, money/travel. These techniques for capturing the implied relationships between index terms have been exploited in a number of contexts. At another level statistical counting is used to increase the precision with which query describes the information need. Although I do not know with any detail how the various search engines process queries, it is my impression that they do it very coarsely.

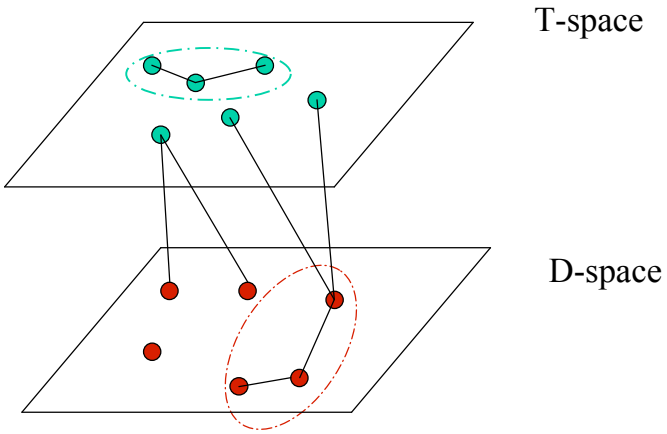
Query Dependence

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

SSMS 2006 © CvR

Although most of the search engines would have one think that retrieval is a matter of formulating a query and then doing a search looking for matching documents, there is a welter of other techniques that do not depend on a query except perhaps for starting things off. A prime example is the ostensive approach investigated by Campbell in myself, here the retrieval process and visualisation is entirely driven by pointing and user actions. More obvious ones come about through the linking of objects which is well known to you all. The use of citation links has a venerable history in IR, for example, in 1980, Belver Griffiths published a collection of "Key Papers in Information Science"which emphasised the importance of citation linkage. Some of the early IR models were based on decision theory and did not presuppose a query but took as their starting point that the objects to be retrieved were separable in at least two classes and went on to use, what are now called, machine learning or vector support machines to generate a decision function separating the classes. More recently the design of recommender systems has given rise to filters that are not based on content at all but use the actions of a user and his or her friends to construct appropriate filters. Some of the above techniques are of course used in the current Google implementation but to the best of my knowledge very little probabilistic or frequency information is used: there is scope for generalising these techniques incorporating some of the probabilistic approaches to IR. As always it is easier to model things either deterministically or stochastically but mixing the models is hard.

Navigation - Browsing



SSMS 2006 © CvR

This slide represents how one might combine Navigation and Browsing in a document space and a term/attribute space. These two spaces are dual of each other. This means that in general operations in the ones space are reflected in the other because of of the interpretation links running in both directions. This means that each object in the D-space has a number of attributes linked to it, and conversely each object in T-space is linked to a number of objects in D-space. This relationship can be formalised algebraicly and indeed is done so in my Geometry of IR.

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

SSMS 2006 © CvR

The nature of what is wanted by a user is a matter for debate. In IR the approach is to assume that a user has an information need which will reveal itself through interaction with a system, this may involve query formulation and reformulation. It is not enough to say that what is wanted a matching item, matching items may be irrelevant or useless. Thus the specification of what is wanted may be left to unfold through interaction and the "passage of experience". Indeed it may be the case that it is not possible to come up with a propositonal form of what is wanted, of course, SQL -like systems assume that it always is! Furthermore, in the end users seek information which may or may not be contained in what are apparently relevant documents. It is a convenience to conflate relevance with aboutness, but now especially in the context of web searching it may be necessary to begin separate these. Also, increasingly searches are done within a context of performing a task, the nature of the task could have a significant effect on what is worth retrieving. To date IR has concentrated on modelling content to support retrieval, but increasingly it is other factors that play a significant role, some of these may only appear as a consequence of iterating a search. Take for example the average query that is put to a search engine which will contain 2.5 query terms, it cannot be assumed that 2.5 terms is a good representation of a user's information need, so what to do? IR offers obvious techniques like relevance feedback, query expansion and a host of other techniques for going beyong a simple query.

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.’

Relevance is usually assumed to be static, but several authors in the apst, epeciall Goffman and maron have pointed out that it is more natural to assume that it is dynamic. More about this later.

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

Maron's theory of indexing makes the assignment of an index term dependent on the user population. This makes it a stochastic event. It cannot be decided simply by looking at the content of a document whether it is relevant to a query term appearing in the document.

‘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

Another expression of the fact that relevance is a dynamic notion.

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 →????

SSMS 2006 © CvR

The evaluation methodoly in IR has been extremely strong, and I would say that the continuing success of the subject as a discipline owes much to that strength. It is is also a good example of something that research concerned with the web as a source of information for utilisation and discovery would do well to look at. Much IR research is subject to extensive testing and experimentation which has led to very modest claims being made about the success of IR. On the other hand such claims generally have stood the test of time. The basis of much evaluation has been the two well known parameters precision and recall used in conjunction with each other. Their use has been backed by extensive statistical analysis and indeed a theory of measurement. The approach arising out of the Cranfield Paradigm via the ideal collection (Sparck Jones and Van Rijsbergen) culminating in TREC has been to design data for experimentation so that the evaluation a parameters make sens, thus the implied trade-off between the two parameters is taken seriously, quoting one without the other makes little sense. Unfortunately the data available on the web does not fall within this paradigm although the retrieval performance is still subject to the trade-off. Hence it would seem important to extend the IR evaluation approach to web data, but to do this problems will have to be solved, for example, how to deal with the lack of recall.

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)

SSMS 2006 © CvR

There are two conflicting ways of looking at the problem of characterising documents for retrieval. One is to characterise a document through a representation of its contents, regardless of the way in which other documents may be described, this might be called *representation without discrimination*. The other way is to insist that in characterising a document one is discriminating it from all, or potentially all, other documents in the collection, this we might call *discrimination without representation*. Naturally neither of these extreme positions is assumed in practice, although identifying the two is useful when thinking about the problem of characterisation. In reality there is a trade-off between the two. Traditionally this is described as the trade-off exhaustivity and specificity of indexing. To the best of my knowledge in IR this has been inescapable, in fact the balance between within document term frequency (tf) and inverse document frequency (idf) can be seen as an attempt to control this balance. Clearly one can adopt either a representation orientation which would emphasise the modelling of documents, for example through a language model. Or one could adopt a discrimination orientation which would emphasise the query leading to query expansion techniques. But which ever one emphasises it is generally at the loss of the other. The implication of these considerations (and others) is that *perfect retrieval is impossible,* this is by way of The Second Law of Retrieval (Van Rijsbergen, 1979). This is a statistical statement, namely, it applies for sets of queries and documents, clearly if there was only one document then perfect retrieval would be easy.

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

SSMS 2006 © CvR

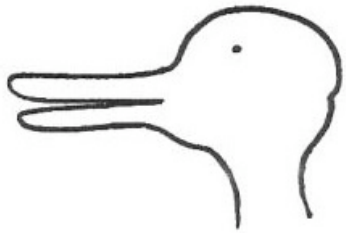
If we contrast retrieval from text with retrieval from images, some of the notions that seem intuitively acceptable for text come less acceptable for images. Most significantly is the importance of tf/idf weighting for text, there is no obvious equivalent for images. What would one count in images corresponding to term frequency in text documents? What an image is about is very context dependent, see my duck/rabbit example later. Similarly relevance may be a function of what has been seen before in images. When providing feedback for images, the nature of the feedback may depend very much on what part of the images the user concentrates on. The interaction between relevance decisions and aboutness may be much more extreme for images than for text.

Text	Images
• keywords	• ?
• frequency	• ?
• meaning	• object recognition
• grammar	• geometry
• salience?	• eyetracking/EEG
• relevance	• path dependent
• query expansion	• how?

SSMS 2006 © CvR

Let us see if we can set up a correspondence between critical aspects of text retrieval and image retrieval.

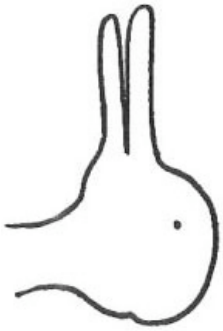
DUCK



SSMS 2006 © CvR

If you are not told what this is a picture of what would you guess. Now look at the next picture.

RABBIT



SSMS 2006 © CvR

The picture is the same but turned through ninety degrees. Now guess what it is a picture of. Now ask yourself the question, does the property of it being an X or Y belong to the image? or does it emerge from the interaction with you the observer?

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

An early sales pitch for the use of non-classical logic in IR. I am about to illustrate the use of such non-classical logics. For details you should look at book by Crestani, et al I cite at the end of the slides. The Logical Uncertainty Principle is quoted extensively in Crestani, et al:

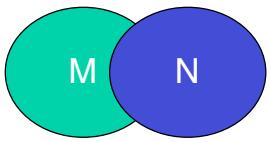
Given any two sentences x and y: a measure of the uncertainty of y -> x relative to a given data set is determined by the minimal extent to which we have to add information to our data set, to establish the truth of y -> x.

Logic

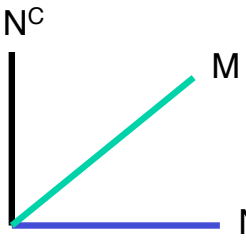
If Mark were to loose his job, he would work less
If Mark were to work less, he would be less tense

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

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



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



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

SSMS 2006

© CvR

One of the active areas of research in IR is the search for appropriate logics to support the reasoning about objects. What has become increasingly clear is that classical Boolean logic is not appropriate in IR, and it is my guess, that the same is true for the use of ontologies. Much of this boils down to how to represent the aggregation of objects into subsets or subspaces, and what the relationship between an object and a subset might be. For example in Boolean Logic when the aggregation is simply subset formation and the relationships are given by inclusion, union, intersection etc, things are relatively straight forward. But in IR we have more structure than just the naming of objects, we have a notion of similarity/dissimilarity on the information space, and aggregate objects algebraically through something akin to subspace formation. The logic that comes with the increased space structure is typically non-classical, for example it fails to meet the distribution law. More than that, we require the generation of an appropriate probability measure on the space, this in itself is non-trivial. My understanding is that in the world of ontology construction similar problems are starting to emerge and it would seem that there is scope for collaboration.

Interaction (Aboutness)

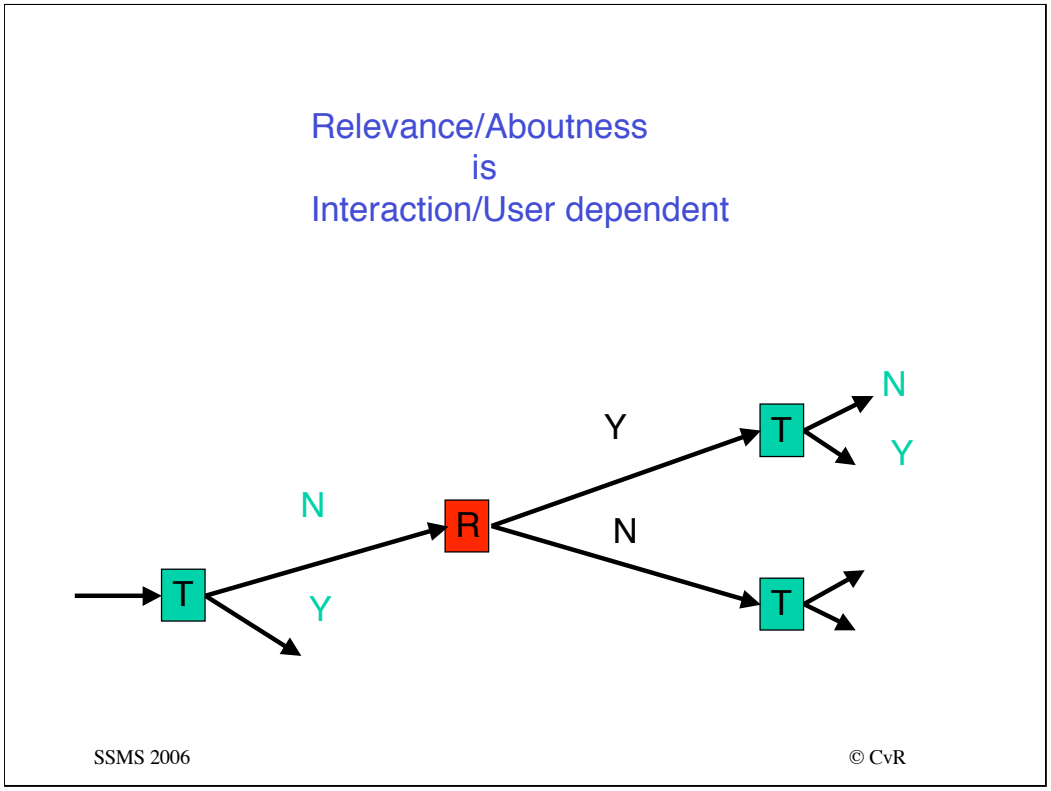
Objects: documents, queries → Relevance

Model

Observable(States) → ??

SSMS 2006 © CvR

One of the hallmarks of IR to date is that within the interaction between the user and a document, the document is seen as a passive object. I would like to suggest that perhaps we should consider a document is being active. The model I have in mind for that is somewhat akin to the the "expectation catalogue" idea Schrodinger had for the state, or wave function in Quantum Mechanics. According to this view a document is stochastic object and it is only through interaction with it that we uncover its meaning. The result of an interaction, or the application of an observable, is a measurement which is inherently uncertain. Thus relevance and aboutness are seen as observables which are represented by Operators (queries), the documents are state functions. To apply an operator is to elicit a measurement with a certain probability. For example, a document is seen to be about ducks or rabbits, but what it is actually about will depend on who is looking with what probability. Looking at things this way opens a duality between the document and the query space. documents can be seen as on the space of queries (operators). Some of you will see the direct parallel between this model and the Von Neumann model for QM, this is not accidental. In fact some of the Quatum Logics are the same as some of the non-classical logics for IR. I would suggest that if we pursue the development of a framework such as this the ontological approach will integrate nicely with the statistical, or probabilistic, approach of IR.



Given that we model queries as operators on state space, then it becomes possible to model the dependence between observables. It has always struck me as absurd that in classical IR models relevance and aboutness do not interact. For example, observing that a document is not about banks, followed by an observation that it is, say, relevant should affect a subsequent observation about it "bankness"; in current IR it does not. I am sure you can think of a host of examples where such an interaction should be expected and not blocked. Thinking of documents as dynamic objects, and modelling them in relation to operators in the way I have described should open the door to such dependence between attributes.

Where are we now in IR?

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

What follows next is a summary view of the IR landscape.

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

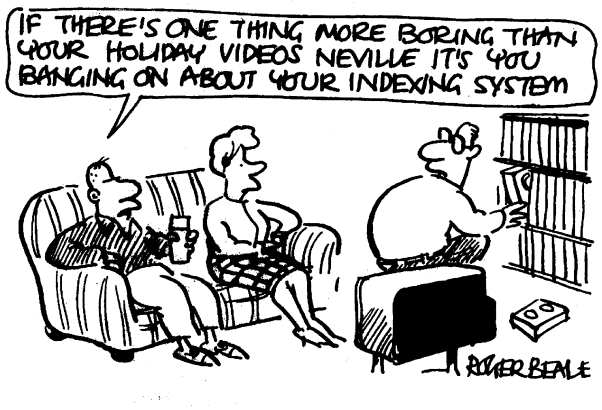
Readings in Information Retrieval, Morgan Kaufman, Edited by Sparck Jones and Willett

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.

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
ROGER BEALE