

The state of information retrieval

a researcher's view

Stephen Robertson

Microsoft Research Cambridge

and City University

ser@microsoft.com

An exceedingly brief history of IR

Guide cards (third millenium BCE)

Classification (first millenium BCE)

Indexing

- Alphabetization (3rd century BCE)
- Printed indexes (15th century CE)

Index languages (1950s)

- Pre- and post-coordination
- Thesauri
- Facet analysis

Punched cards (early 20th century)

- Mechanically sorted
- Hand sorted
- Peek-a-boo

Evaluation (1950s) (more below)

Computers (1950s)

Boolean systems (1960s)

Free text (1960s)

Weighting and ranking (1960s)

The Cranfield tradition

We have been evaluating relevance for over
50 years

We continue to argue over the details...

... but the basic principle remains:

we need humans to tell us how well or badly we
are doing

The Cranfield approach to this:

Documents, requests, relevance judgements

Limitations of Cranfield

Very good for algorithms and core methods

Not so good for user interface, interaction,
cognitive aspects

But this is another talk...

TREC, the Web and Google

(The last 15-20 years)

TREC revitalised the Cranfield tradition
and took it into new domains

The Web made huge quantities of information available
but presented vast problems of discovery

Search engines have provided a general-purpose solution
the web without search engines is almost unimaginable now
... but in the meantime, smaller systems have lagged
behind

From evaluation to optimisation

Cranfield 1 compared four complete systems...

... but ever since, we have been examining components

... and tweaking knobs

Modern ranking algorithms have many knobs
sometimes very many!

Sources of evidence

Sources of evidence for the system

- Local

 - (about the query, document, query-doc match, user)

- Global

 - (about the world)

Sources of evidence for evaluation

 - (about how users react to documents)

 - these may also be evidence for the system

Local sources for the system

(meaning local to particular events: documents, queries)

About the document

- Content (words, sequence, NLP analysis)
- Linkage
- Structure (fields, metadata)
- Other (url, document type, location, ...)

About the query

- Content (words, sequence)
- Other (user, group, location, ...)

Local sources for the system

About the query/document match

- Mix and match the above
 - match any query feature against any document feature
- Potentially large number of features
 - even if the query information is sparse
- ... and potentially complex
 - e.g. phrases matched against body/anchor

Global sources for the system

(meaning sources external to the particular events)

Index languages / taxonomies / ontologies

- sources designed to encode knowledge useful for search

Sources that may be mined for knowledge, e.g.

- wikipedia
 - e.g. person names
- company address book
 - e.g. ditto
- search logs

Sources of evidence for evaluation

Direct

- relevance judgements
- list-based judgements
- task success

Indirect

- clicks
- other user behaviour

[more later on use of this information by the system]

Using the evidence

Richness presents problems

- how do we combine all this evidence?

Note: *all* evidence is noisy, *all* signals are weak!

... which is exactly why statistics are so important

Major lesson of TREC and web search:

- Statistics rules OK!
- guided by models

Models

(an over-used word)

A model encodes an abstract view of the world

- or some part of the world
- to some degree of abstraction

Some classes of models

- Models of knowledge
 - classifications, taxonomies, ontologies, relational models, ...
- Models of language
 - syntax, semantics, morphology, statistical sequence, phrase structure, ...
- Statistical models
 - discriminative, generative, Bayesian, supervised / unsupervised / semi-supervised learning, ...

Models: statistical and other

Broadly statistical models (incl recent “language” models) have been extraordinarily successful

But they do not replace the other kinds of models

- they start where the others leave off

How to integrate other models with statistical ones?

- statistical models seem to encourage simple views of the other areas
- sometimes hard to incorporate insights from other areas into statistical models

Example

Phrases in the Binary (Independence) Model

BIM:

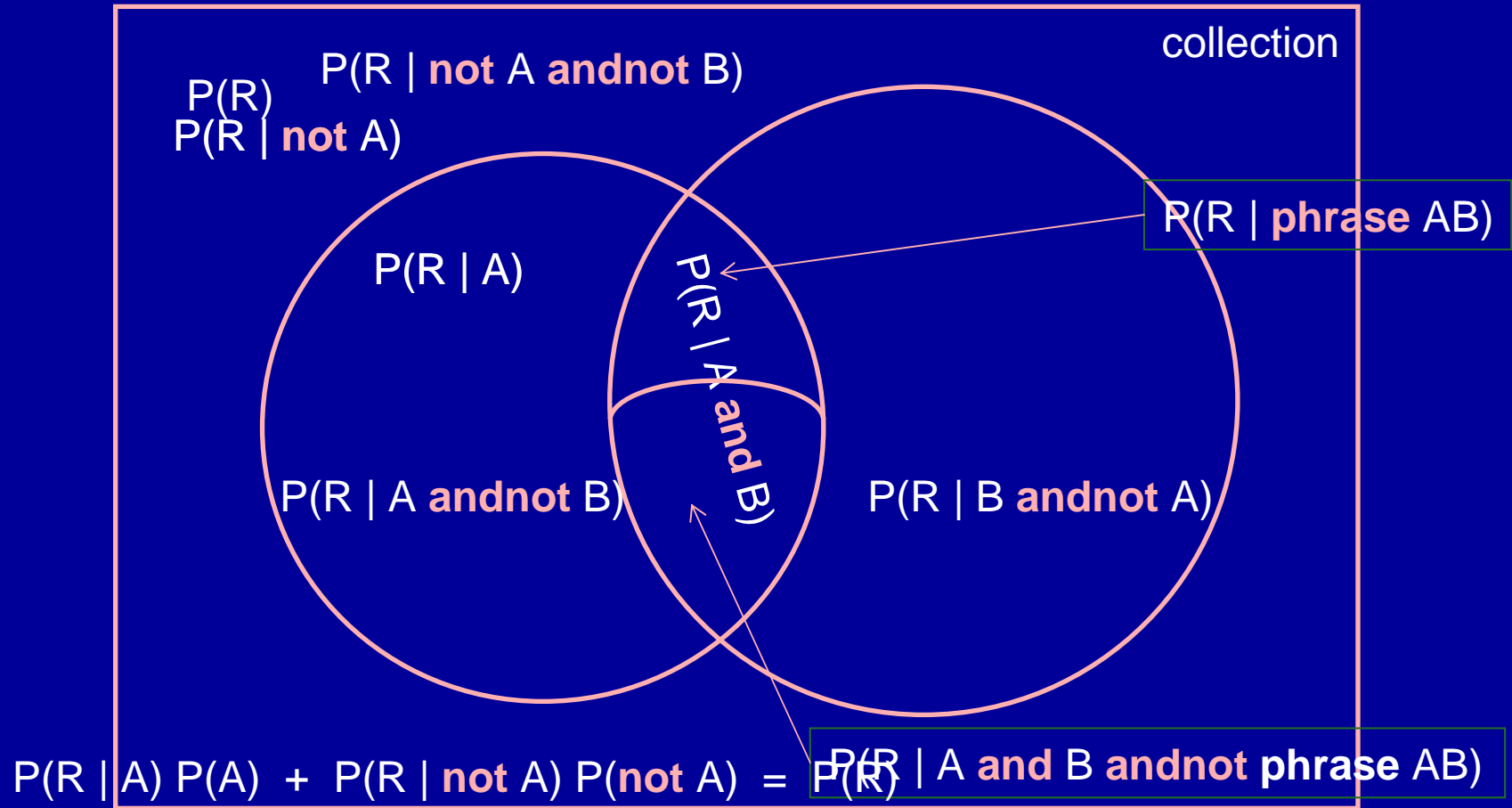
- Terms occur independently *given relevance*
- Predict probability of relevance given combination of terms

BIM provides a consistent model of the distribution of relevance probability

... but of course terms are not independent

One source of dependence: phrases

Example



Another example

Soft query expansion

- add related terms to a query term, e.g.
 - words with same stem
 - synonyms from WordNet
 - subordinate terms from a taxonomy...

Simple (hard) version: Boolean **or**

treat (A **or** A' **or** A'') as if it were a single term

Better:

treat A' or A'' as providing some evidence

... but probably less than A itself

Machine learning

Machine learning:

- statistical model with free parameters
 - maybe many free parameters
- set these parameters by learning from examples
 - maybe many examples

... and of course we have the examples

- Cranfield-style queries and relevance judgements
- and/or logs with clicks

In principle ML should be able to help

Machine learning

Typically the process of learning involves

- Choosing a metric ('objective function') which we want to optimise
- Associating an inference/optimisation method with the parameters of the model
- Running this method on some set of training data

(many more issues and details!)

One issue:

- metric = IR effectiveness measure??

Machine learning

What we need to do

- Devise models which encode our prior knowledge and understanding
- ... but with all that might be learnt embedded in the free parameters
- Understand the metric issue
- Test, test, test

Some of this is happening now

- but we are only at the beginning!

Different uses of evidence

Traditional: relevance feedback

- use relevance judgements to improve *this* search
- or to improve the indexing of *this* document
(for other searchers)
- or (somehow) to do both (the *unified model*)
(variants such as PRF)

The new source: clicks

Uses of clicks

Clicks are noisy evidence

(but then so are relevance judgements)

May be enhanced by additional evidence

e.g. dwell time

Same usage issue

this search / *this* document / both together

... but with an additional wrinkle

– identical queries

Final thoughts

Search systems are seeking to use an ever richer range of sources of information to enhance indexing and search

Statistical models are pervasive – we can't do without them

- The larger the system, the more important they are

But we need to connect our statistical viewpoint with all the other types of knowledge that we have

We also need to deal better with smaller systems where extensive ML is not feasible

- e.g. desktop, within-site, specialist