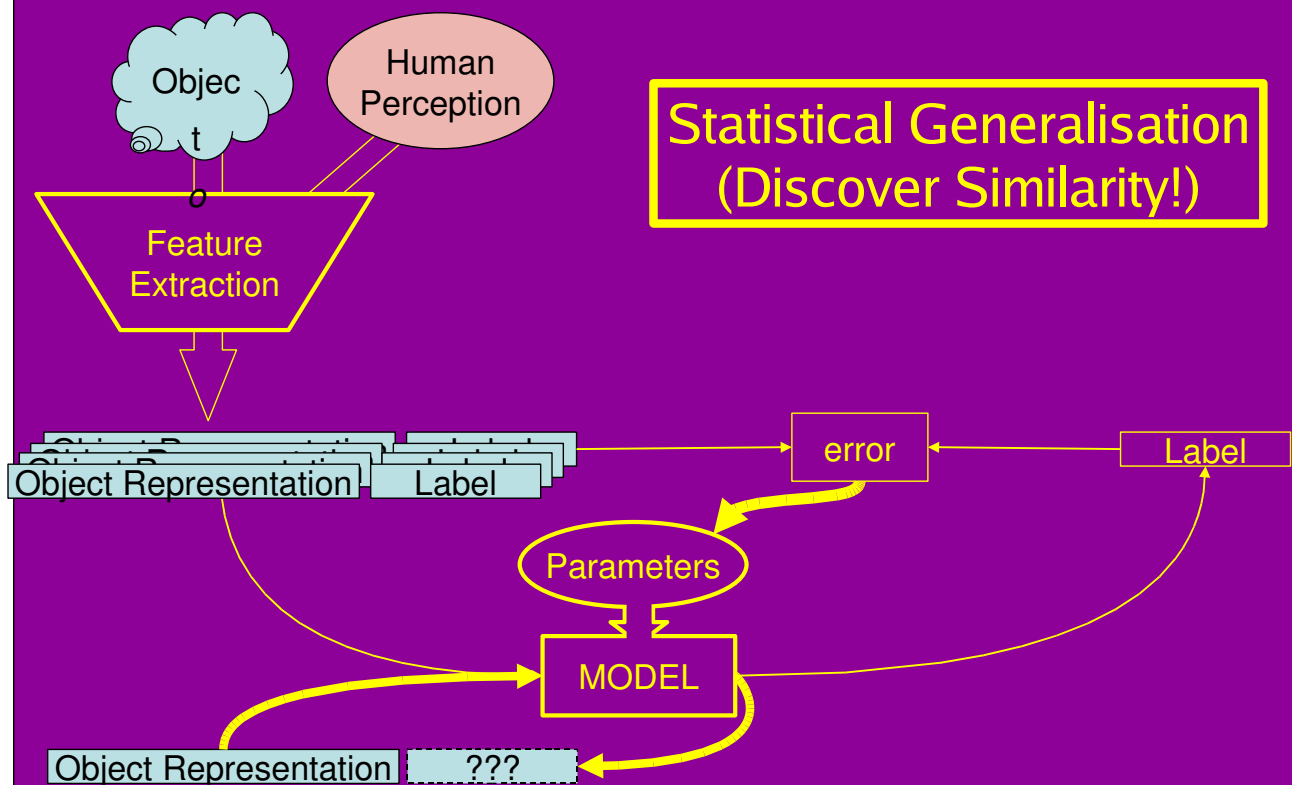


"Tuning": Error Optimisation in Ad-Hoc Retrieval

Hugo Zaragoza,
Yahoo! Research Barcelona.

(This work was completed at Microsoft Research Cambridge,
in collaboration with:
Ralf Herbrich, Stephen Robertson, Michael Taylor, Nick Craswell & Chris Burges.)

Machine Learning intro.



Machine Learning intro.

- Label type:

- Binary $\{A, \neg A\}$
- Discrete $\{A, B, C\}$
(Discrete but unknown)
- Preferences $\{A > B > C\}$
- Continuous $\{R\}$

Task:

- Clasification
- Multiclass Class.
Clustering
- Ordinal
Regression
- Regression

Machine Learning intro.

- The promise...

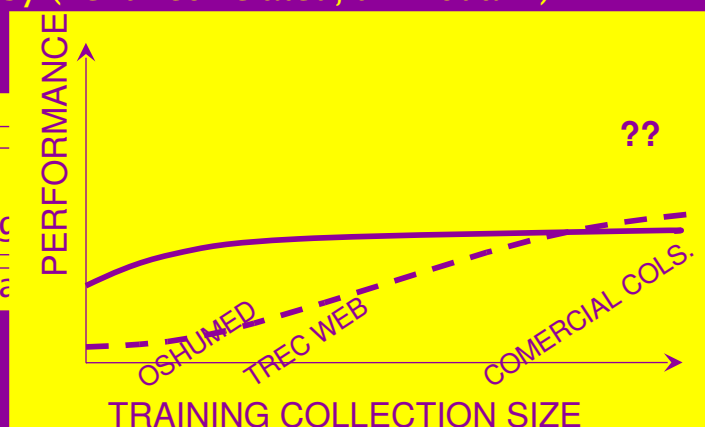
- Features + labels + ML = Good Model.
- Concentrate on innovation, tasks, forget the details...

- Until today...

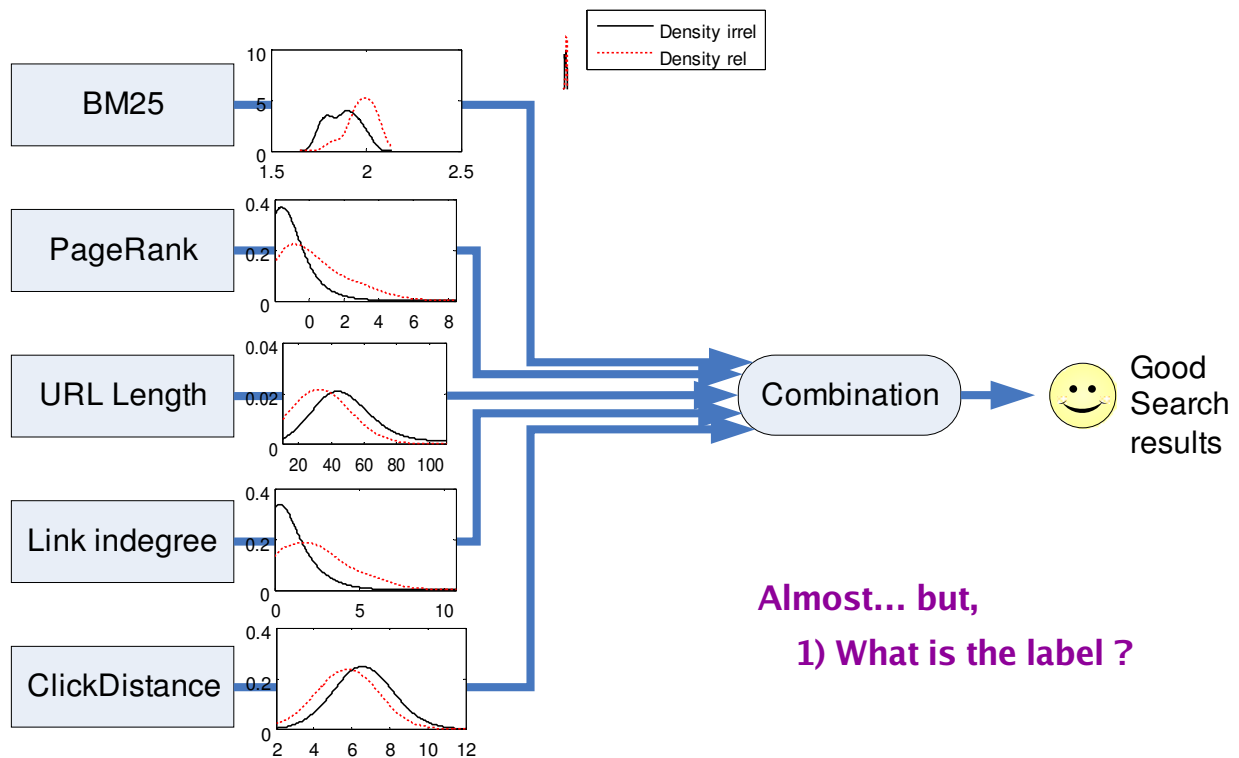
- *Good* features.
- *Lots* of labels, good (#labels / #feature) ratio.
- *Not too noisy* & nicely noisy (i.e. uncorrelated, unimodal...)
- *Element-wise* error.

- To

- i.e.
- i.e.
- $tf, df \rightarrow tf \cdot \log$
- HARD: not enough la



Learning Web Ranking Functions



“Ranking” issues

- Rank-dependant evaluation

The *goodness* of one score depends on all the others.

score	rank	relevant?
17.9	1	<input checked="" type="checkbox"/>
17.6	2	<input checked="" type="checkbox"/>
12.5	3	<input checked="" type="checkbox"/>
8.5	4	<input checked="" type="checkbox"/>
5.3	5	<input checked="" type="checkbox"/>
...

Precision@5 = 2/5

- Query-dependant scale

The relevance scale changes across queries.

query1 = “Britney”
 doc. A () score = 5
 doc. B () score = 0

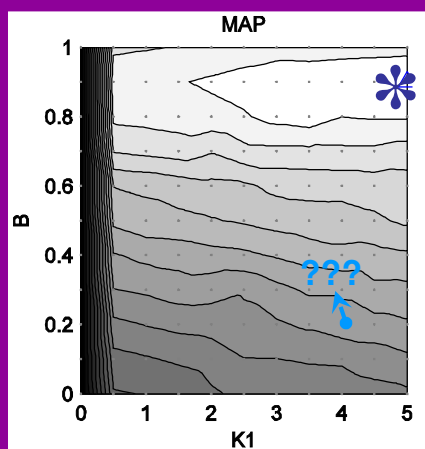
query2 = “Britney Spears”
 doc. A () score = 15
 doc. B () score = 7

“Ranking” problem

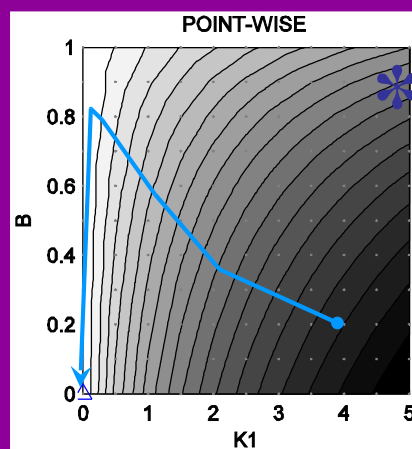
1. We are only interested in **top ranked** objects.
 2. Labels are only **relative** to other labels.
 - i.e. *Doc324 is best, Doc311 second best, etc.*
- *e.g.*
 - Find the parameters that minimise L2 distance on the highest 20 target values
 - Precision at 5
 - Average Precision, NDCG, Reciprocal mean, ...
 - *Difficult, but ML is starting to propose practical solutions to this problem!*

Example (Topic Distillation, TREC'04)

Avg. Precision Surface



Point-Wise (Cross-Entropy)

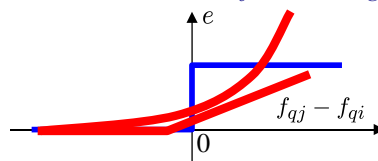


Pair-wise error measures

Possible relevance pairs

(i, j)	error
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	: 0
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	: 0
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	: $e(f_{qi}, f_{qj})$
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	: $1 - e(f_{qi}, f_{qj})$

j scores lower \leftarrow \rightarrow j scores higher



$$e(f_{qi}, f_{qj}) := \text{sign}(f_{qj} - f_{qi})$$

Smooth-Step,
Hinge Lose,
Exponential...
+ standard
ML machinery

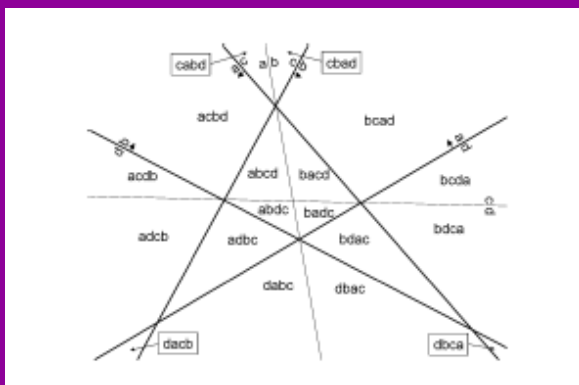
Number of irrelevant above relevant doc i :

$$\text{mistakes}(i) = \sum_{j \in \mathcal{T}_q} e(f_{qi}, f_{qj})$$

Precision at 5:

$$\text{Precision@K} = \frac{1}{Z} \sum_q \max_{j | \text{rank}(j) \leq K} \{ \text{mistakes}(j) \}$$

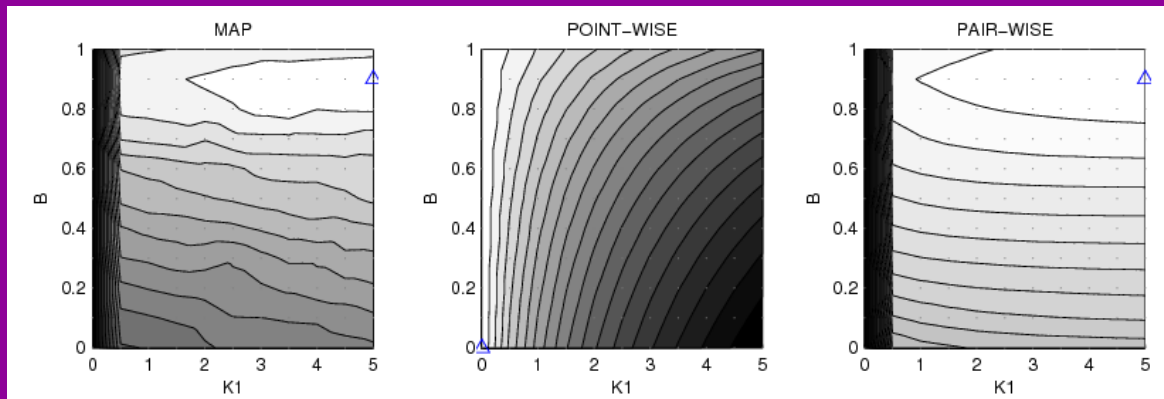
Pair-wise for NDCG: should it work?



[Robertson S., Zaragoza H., MSR-TR-2006-61]

- At a maxima, we cannot decrease pair-wise errors without decreasing NDCG or Avg. Prec.
- Therefore, these measures must share local maxima!

Pair-wise for NDCG: does it work?



average precision

cross-entropy

hinge loss

Pair-wise for NDCG: does it work?

