

Portfolio Theory of Information Retrieval

Jun Wang and Jianhan Zhu

`jun.wang@cs.ucl.ac.uk`

Department of Computer Science
University College London, UK



Outline

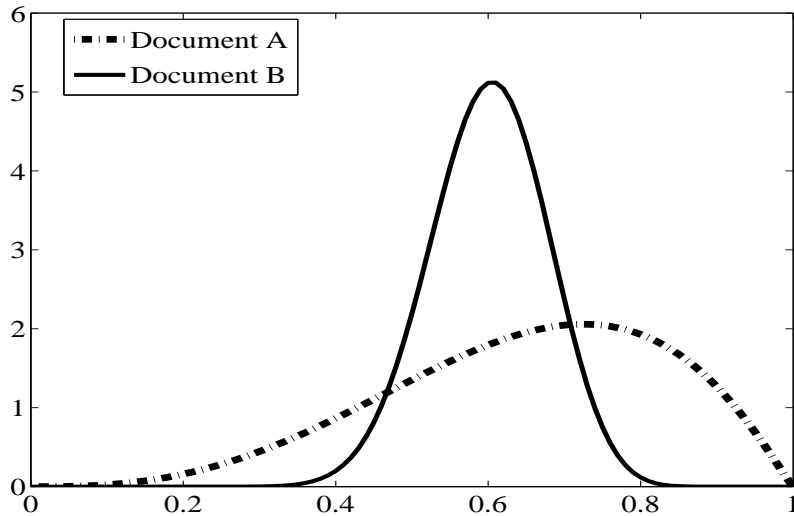
- Research Problem
- An Analogy: Stock Selection In Financial Markets
- Portfolio Theory of Information Retrieval
- Evaluations on ad hoc text retrieval
- Conclusions

Two-stage Process in IR

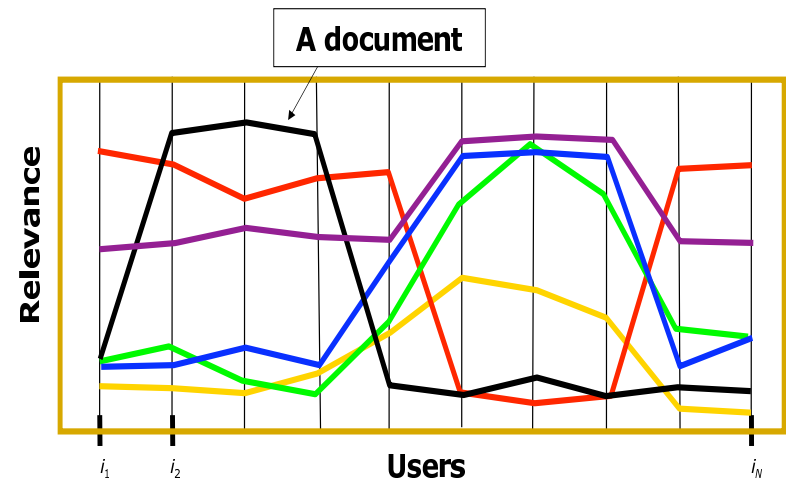
- Stage 1 calculates the relevance between the given user information need (query) and each of the documents
 - Producing a “best guess” at the relevance, e.g. the BM25 and the language modelling approaches
- Stage 2 presents (normally rank) relevant documents
 - The probability ranking principle (PRP) states that “the system should rank documents in order of decreasing probability of relevance” [Cooper(1971)]
 - Under certain assumptions, the overall effectiveness, e.g., expected Precision, is maximized [Robertson(1977)]

Ranking Under Uncertainty

- the PRP ignores [Gordon and Lenk(1991)]:
 - there is uncertainty when we calculate relevance scores, e.g., due to limited sample size, and
 - relevance scores of documents are correlated



Uncertainty of the relevance scores



Correlations of relevance scores

An Example

- Suppose we have query *apple*, and two classes of users U_1 : *Apple_Computers* and U_2 : *Apple_Fruit*; U_1 has twice as many members as U_2
- An IR system retrieved three documents d_1 , d_2 and d_3 ; their probabilities of relevance are as follows:

<i>UserClass</i>	d_1 : <i>Apple_Comp.</i>	d_2 : <i>Apple_Comp.</i>	d_3 : <i>Apple_Fruit</i>
<i>Apple Computers</i>	1	1	0
<i>Apple Fruit</i>	0	0	1
$p(r)$	2/3	2/3	1/3

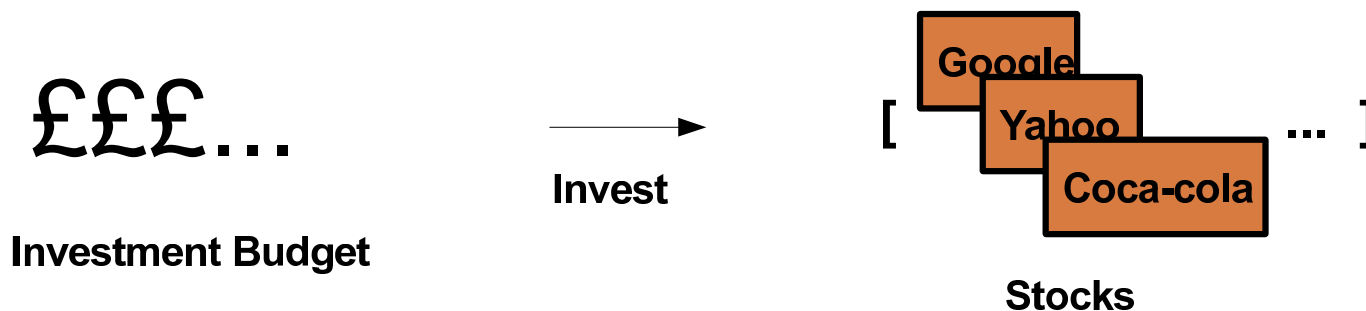
- The PRP is not good ($\{d_1 \text{Apple_Comp.}, d_2 \text{Apple_Comp.}, d_3 \text{Apple_Fruit}\}$) as user group U_2 (*Apple Fruit*) has to reject two documents before reaching the one it wants [Robertson(1977)]

A Little History of Probabilistic Ranking

- 1960s [Maron and Kuhns(1960)] mentioned the two-stage process implicitly
- 1970s [Cooper(1971)] examined the PRP explicitly
 - well discussed in [Robertson(1977)] and Stirling's thesis [Stirling(1977)]
- 1991 [Gordon and Lenk(1991)] studied its limitations
- 1998 [Carbonell and Goldstein(1998)] proposed diversity-based reranking (MMR)
- 2006 The “less is more” model [Chen and Karger(2006)]
 - maximize the probability of finding *a* relevant documents in top-*n* ranked list, where $a < n$

Our View of the Ranking Problem (1)

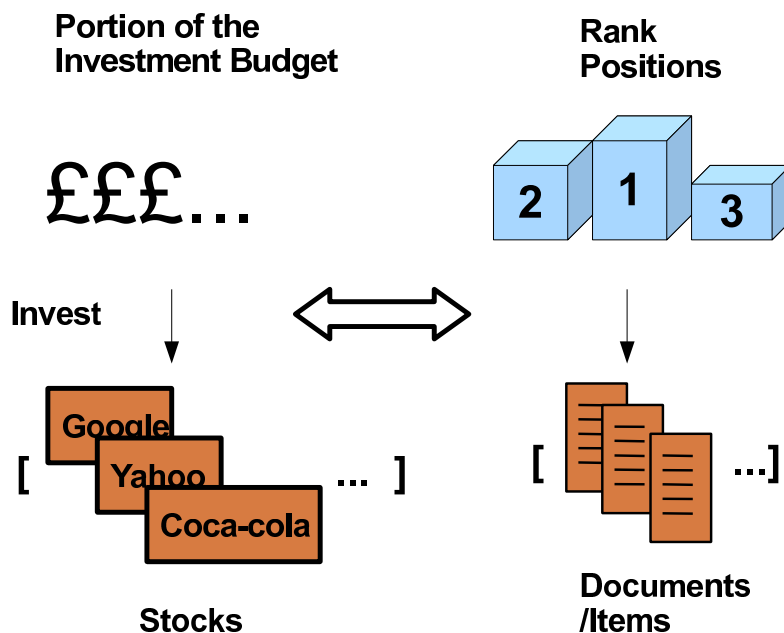
- We argue that *ranking under uncertainty* is not just about picking individual relevant documents, but about **choosing the right combination of relevant document - the Portfolio Effect**
- There is a similar scenario in financial markets:



- Two observations:
 - The future returns of stocks cannot be estimated with absolute certainty
 - The future returns are correlated

Our View of the Ranking Problem (2)

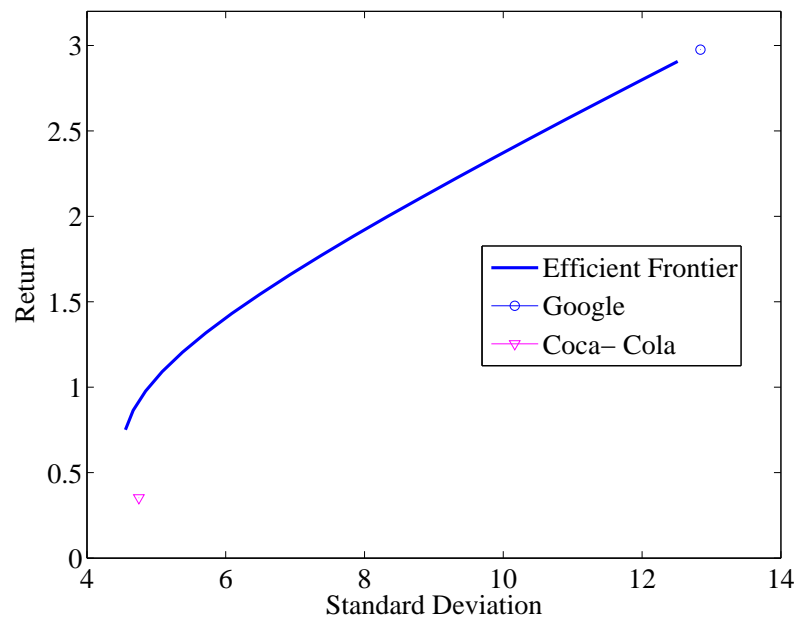
- The analogy:



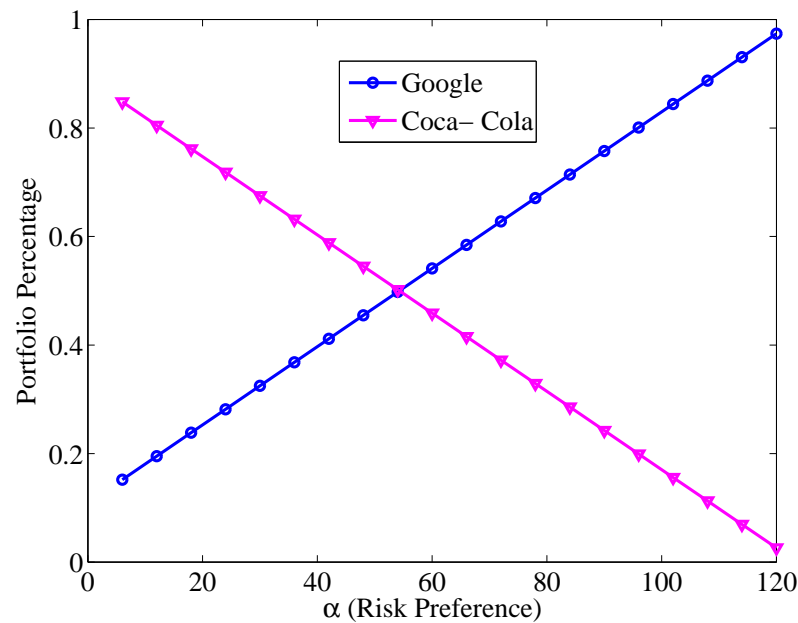
- According to the PRP, one might first rank stocks and then choose the top-n most “profitable” stocks
- Such a principle that essentially maximizes the expected future return was, however, rejected by Markowitz in Modern Portfolio Theory [Markowitz(1952)]

Our View of the Ranking Problems (3)

- Markowitz' approach is based on the analysis of the expected return (*mean*) of a portfolio and its *variance* (or standard deviation) of return. The latter serves as a measure of risk



Efficient Frontier



Percentage in the Portfolio

Portfolio Theory in IR (1)

- Objective: find an optimal ranked list (consisting of n documents from rank 1 to n) that has the maximum *effectiveness* in response to the given information need
- Define effectiveness: consider the weighted average of the relevance scores in the ranked list:

$$R_n \equiv \sum_{i=1}^n w_i r_i$$

where R_n denotes the overall relevance of a ranked list. Variable w_i , where $\sum_{i=1}^n w_i = 1$, differentiates the importance of rank positions. r_i is the relevance score of a document in the list, where $i = \{1, \dots, n\}$, for each of the rank positions

Portfolio Theory in IR (2)

- Weight w_i is similar to the discount factors that have been applied to IR evaluation in order to penalize late-retrieved relevant documents [Järvelin and Kekäläinen(2002)]
- It can be easily shown that when $w_1 > w_2 \dots > w_n$, the maximum value of R_n gives the ranking order $r_1 > r_2 \dots > r_n$
- This follows immediately that maximizing R – by which the document with highest relevance score is retrieved first, the document with next highest is retrieved second, etc. – is equivalent to the PRP

Portfolio Theory in IR (3)

- During retrieval, the overall relevance R_n cannot be calculated with certainty
- Quantify a ranked list based on its expectation (*mean* $E[R_n]$) and its *variance* ($Var(R_n)$):

$$E[R_n] = \sum_{i=1}^n w_i E[r_i]$$

$$Var(R_n) = \sum_{i=1}^n \sum_{j=1}^n w_i w_j c_{i,j}$$

where $c_{i,j}$ is the (co)variance of the relevance scores between the two documents at position i and j . $E[r_i]$ is the expected relevance score, determined by a point estimate from the specific retrieval model

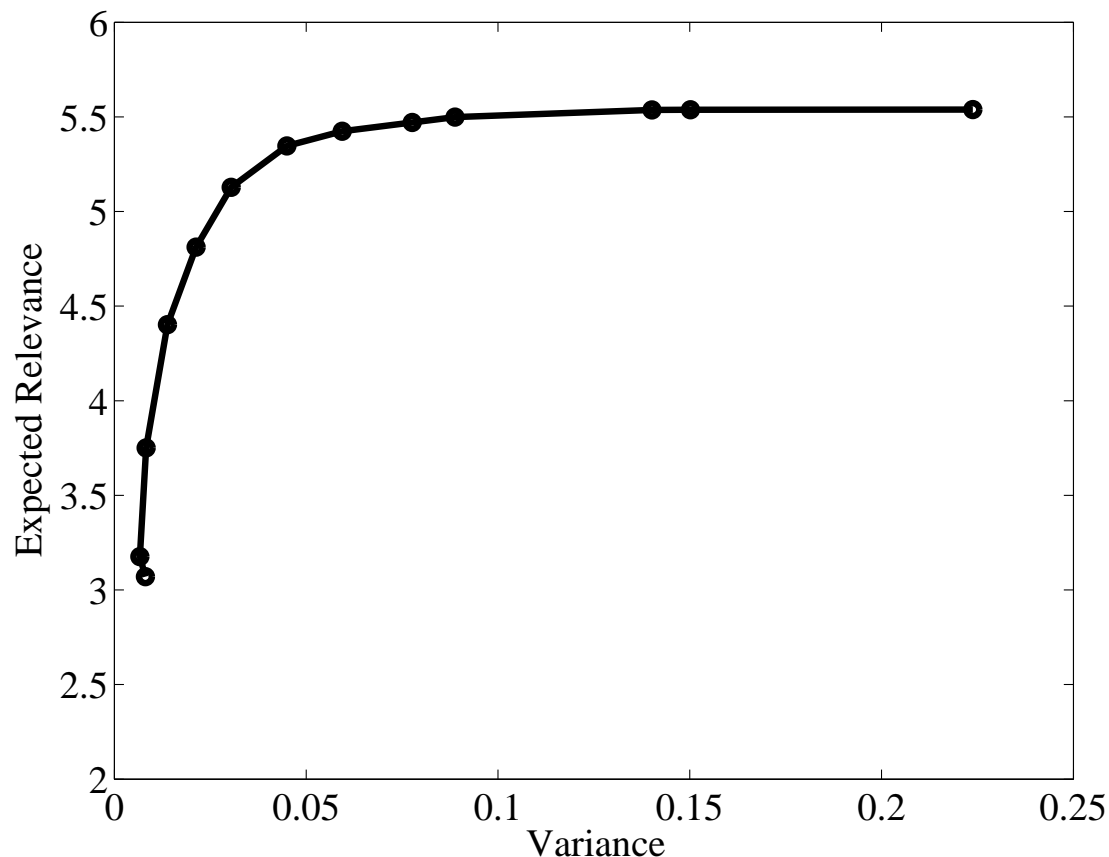
Portfolio Theory in IR (4)

● What to be optimized?

1. *Maximize the mean $E[R_n]$ regardless of its variance*
2. *Minimize the variance $Var(R_n)$ regardless of its mean*
3. *Minimize the variance for a specified mean t (parameter): $\min Var(R_n)$, subject to $E[R_n] = t$*
4. *Maximize the mean for a specified variance h (parameter): $\max E[R_n]$, subject to $Var(R_n) = h$*
5. *Maximize the mean and minimize the variance by using a specified risk preference parameter b :
 $\max O_n = E[R_n] - bVar(R_n)$*

Portfolio Theory in IR (5)

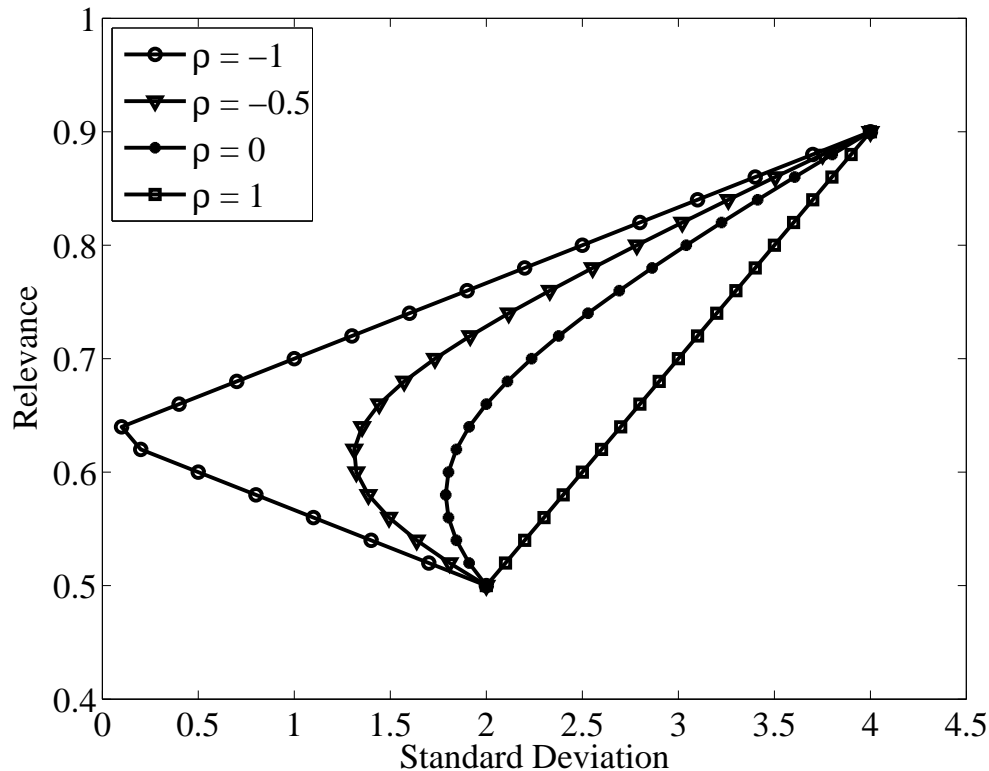
- The Efficient Frontier:



- Objective function: $O_n = E[R_n] - bVar(R_n)$ where b is a parameter adjusting the risk level

Portfolio Theory in IR (6)

- Our solution provides a mathematical model of rank diversification
- Suppose we have two documents. Their relevance scores are 0.5 and 0.9 respectively

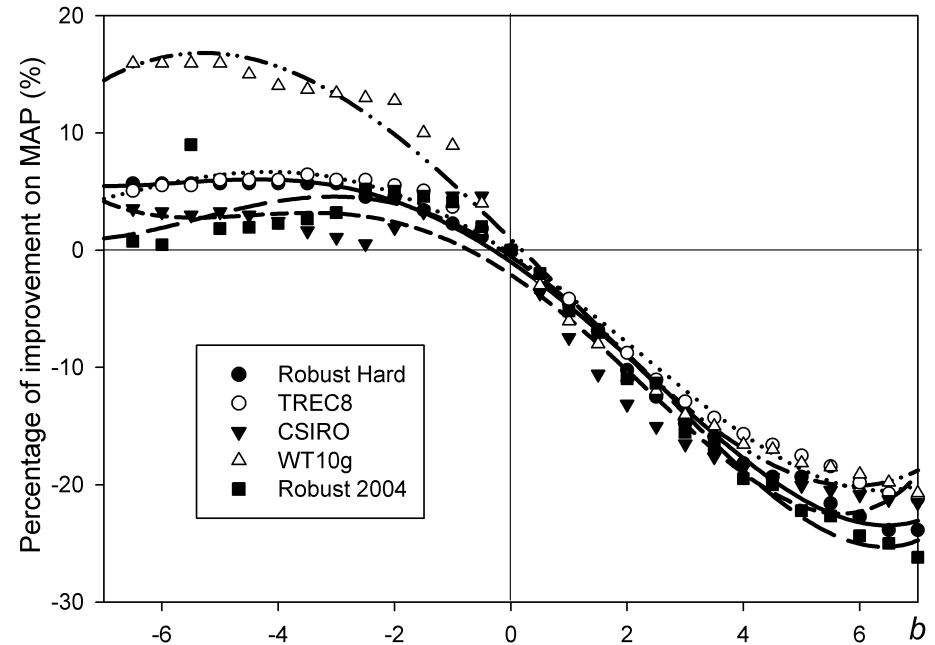
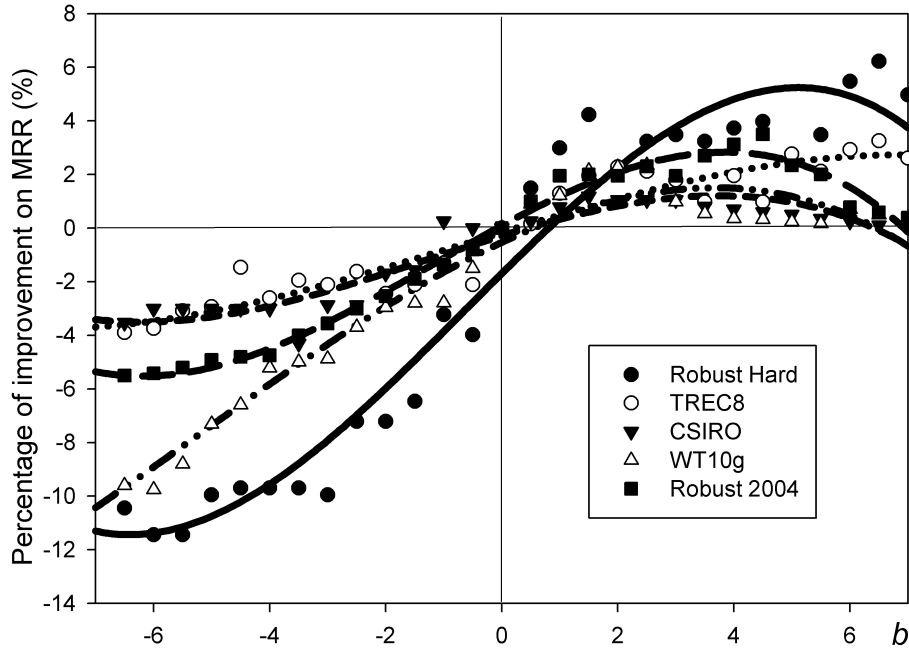


Evaluations on Ad hoc Text Retrieval (1)

- We experimented on Ad hoc and sub topic retrieval
- Calculation of the Mean and Variance:
 - *Mean*: posterior mean of the chosen text retrieval model
 - *Covariance matrix*:
 - largely missing in IR modelling
 - formally, should be determined by the second moment of the relevance scores (model parameters), e.g., applying the Bayesian paradigm
 - can be approximated by the covariance with respect to their term occurrences

Evaluations on Ad hoc Text Retrieval (2)

- Impact of parameter b on different evaluation metrics



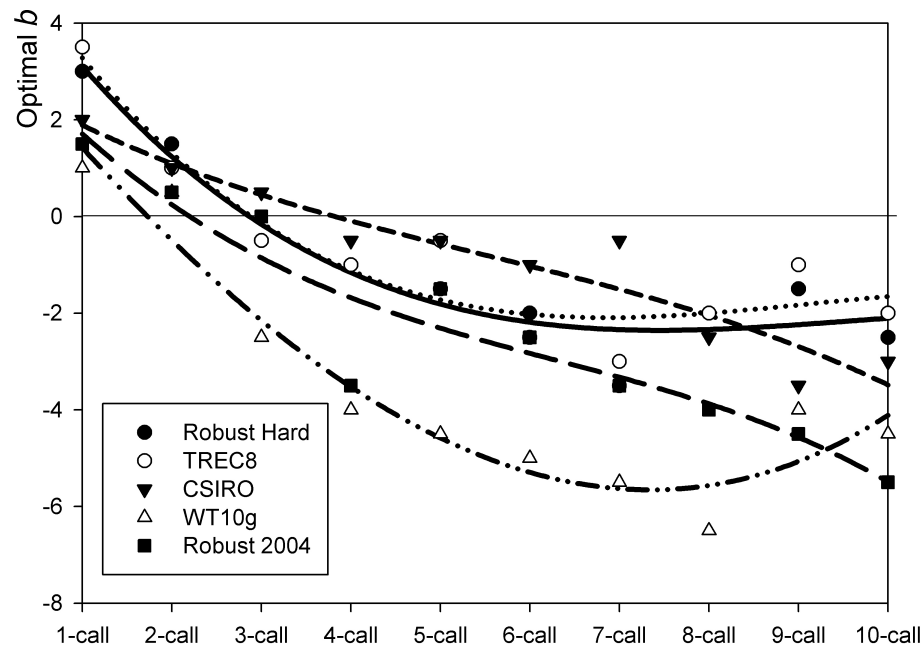
(a) Mean Reciprocal Rank (MRR)

(b) Mean Average Precision (MAP)

- (a) positive b : “invest” into different docs. increases the chance of early returning the first rel. docs
- (b) negative b : “invest” in “similar” docs (big variance) might hurt the MRR but on average increases the performance of the entire ranked list

Evaluations on Ad hoc Text Retrieval (3)

- The impact of the parameter b on a risk-sensitive metric, k -call [Chen and Karger(2006)]
 - 10-call: *ambitious*, return 10 rel. docs.
 - 1-call: *conservative*, return at least one rel. doc.



- Positive b when k is small (1 and 2). diversifying reduces the risk of not returning any rel docs.
- Negative b as k increases. Taking risk increases the chance of finding more rel. docs.

Evaluations on Ad hoc Text Retrieval (4)

● Comparison with the PRP (Linear smoothing LM)

Measures	CSIRO	WT10g	Robust	Robust hard	TREC8
MRR	0.869	0.558	0.592	0.393	0.589
	0.843	0.492	0.549	0.352	0.472
	+3.08%	+13.41%*	+7.83%*	+11.65%*	+24.79%*
MAP	0.41	0.182	0.204	0.084	0.212
	0.347	0.157	0.185	0.078	0.198
	+18.16%*	+15.92%*	+10.27%*	+7.69%*	+7.07%*
NDCG	0.633	0.433	0.421	0.271	0.452
	0.587	0.398	0.396	0.252	0.422
	+7.88%*	+8.82%*	+6.25%*	+7.55%*	+7.05%*
NDCG@10	0.185	0.157	0.175	0.081	0.149
	0.170	0.141	0.169	0.078	0.140
	+8.96%*	+11.23%*	+3.80%	+3.90%	+6.36%*
NDCG@100	0.377	0.286	0.314	0.169	0.305
	0.355	0.262	0.292	0.159	0.287
	+6.25%*	+9.27%*	+7.55%*	+6.58%*	+6.34%*

Evaluations on Ad hoc Text Retrieval (5)

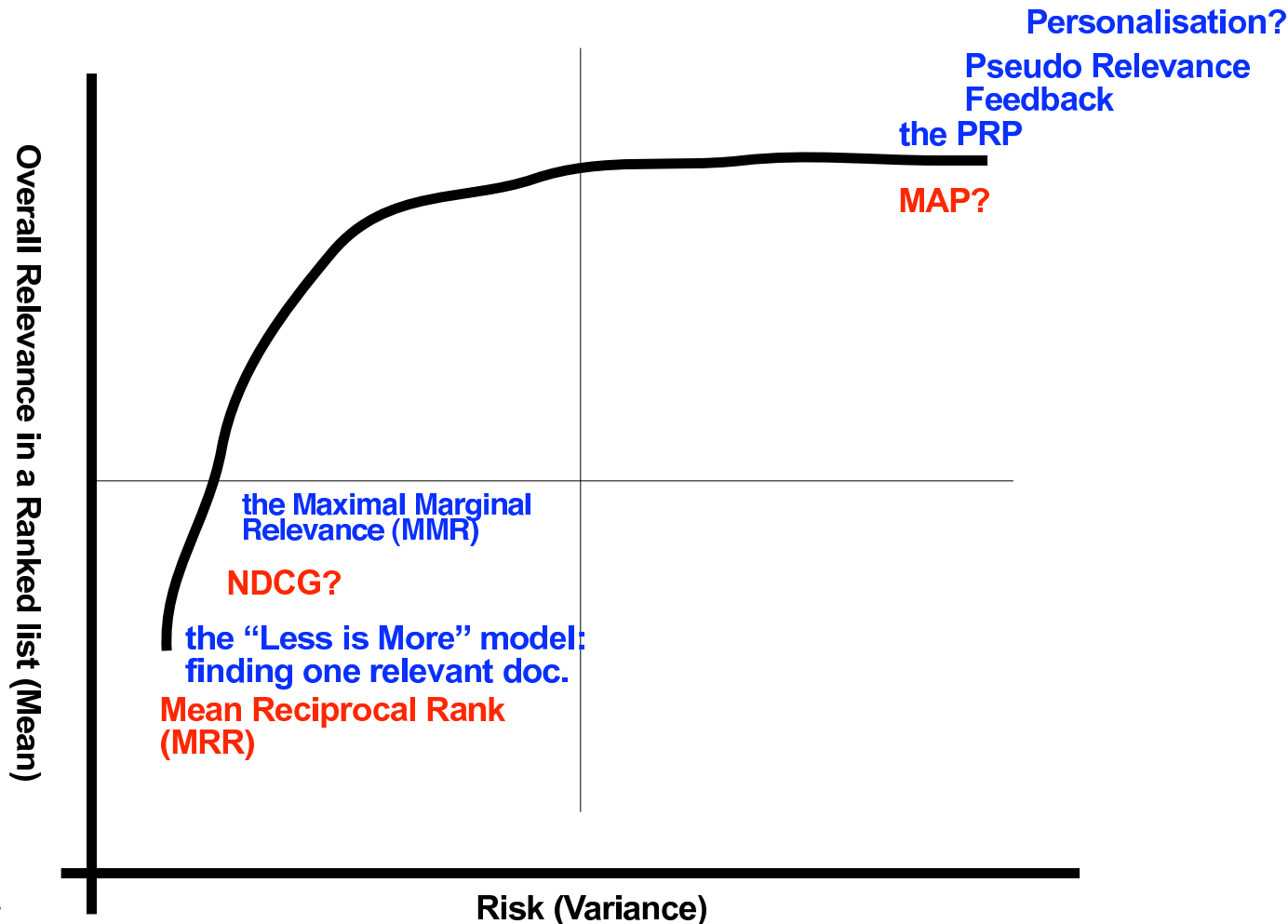
- Comparison with diversity-based reranking, the MMR [Carbonell and Goldstein(1998)]

In each cell, the first line shows the performance of our approach, and the second line shows the performance of the MMR method and gain of our method over the MMR method.

Models	Dirichlet	Jelinek-Mercer	BM25
sub-MRR	0.014 0.012 (+16.67%*)	0.011 0.009 (+22.22%*)	0.009 0.007 (+28.57%*)
sub-Recall@5	0.324 0.304 (+6.58%*)	0.255 0.234 (+8.97%*)	0.275 0.27 (+1.85%)
sub-Recall@10	0.381 0.362 (+5.25%)	0.366 0.351 (+4.27%)	0.352 0.344 (+2.33%)
sub-Recall@20	0.472 0.455 (+3.74%)	0.458 0.41 (+11.71%*)	0.464 0.446 (+4.04%)
sub-Recall@100	0.563 0.558 (+0.90%)	0.582 0.55 (+5.82%*)	0.577 0.558 (+3.41%)

Conclusions

- We have presented a new ranking theory for IR
- The benefit of diversification is well quantified
- Is it a unified theory to explain risk and reward in IR?



References

- [Cooper(1971)] William S. Cooper. The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. *University of California, Berkeley*, 1971.
- [Robertson(1977)] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, pages 294–304, 1977.
- [Gordon and Lenk(1991)] Michael D. Gordon and Peter Lenk. A utility theoretic examination of the probability ranking principle in information retrieval. *JASIS*, 42(10):703–714, 1991.
- [Maron and Kuhns(1960)] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3), 1960.
- [Stirling(1977)] Keith H. Stirling. *The Effect of Document Ranking on Retrieval System Performance: A Search for an Optimal Ranking Rule*. PhD thesis, UC, Berkeley, 1977.
- [Carbonell and Goldstein(1998)] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.
- [Chen and Karger(2006)] Harr Chen and David R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR*, 2006.
- [Markowitz(1952)] H Markowitz. Portfolio selection. *Journal of Finance*, 1952.
- [Järvelin and Kekäläinen(2002)] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 2002.