CIKM 2011 GLASGOW

20TH ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT

Tutorial - AM4

Statistical Information Retrieval Modelling.

Jun Wang, Kevyn Collins-Thompson

CIKM 2011 Tutorial

Statistical Information Retrieval Modelling

From the Probability Ranking Principle to Recent Advances in Diversity, Portfolio Theory and Beyond

Jun Wang University College London

Kevyn Collins-Thompson Microsoft Research

Table of Contents

- Background
  - The need for mathematical IR models
  - Key IR problems and motivation for risk management
- Individualism in IR
  - RSJ model, Language modeling
  - Probability Ranking Principle
- Ranking in context and diversity
  - Loss functions for diverse ranking
  - Less is More, Maximum Marginal Relevance, Diversity Optimization
  - Bayesian Decision Theory
- Portfolio Retrieval
  - Document ranking
  - Risk-reward evaluation methods
  - Query expansion and re-writing
- Future challenges and opportunities

Table of Contents

What is a Model?

- ***Model:***
  - A *simplified representation* of a real object, e.g., a person, thing, a physical system, or a process
  - Used to provide *insight, answers, guidance,* and *predictions*
- *So, a model is* ***a medium between data and understanding***
- Modelling: the construction of physical, conceptual, or mathematical representation (formulations) of a real world object/system

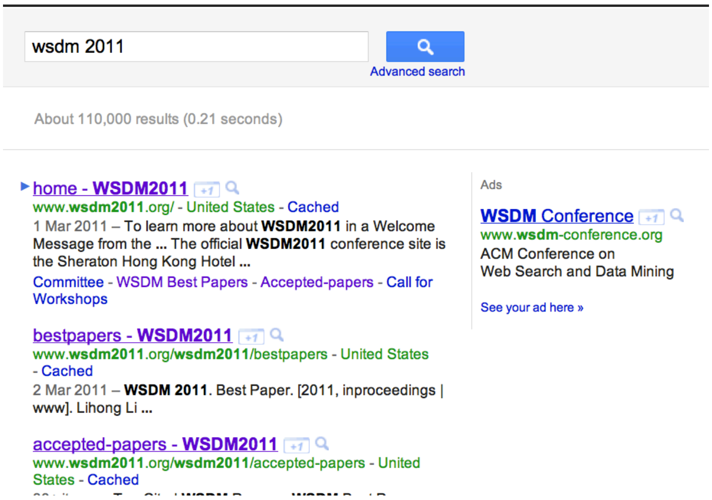
Mathematical Model

- Uses abstract mathematical formulations to describe or represent the real world system or object
- Employs theoretical and numerical analysis to provide *insight, answers, guidance* and predictions

Back to Information Retrieval

- General definition: search unstructured data, mostly text documents. But also include images, videos
- Items include
  - –webpages
  - –product search
  - –enterprise search
  - –desktop/email search

The central problem is to find relevant documents given an information need



What is Relevance?

- Relevance is the "correspondence" between information needs and information items
- But, the exact meaning of relevance depends on applications:
  - = usefulness
  - = aboutness
  - = interestingness
  - = ?
- Predicting relevance is the central goal of IR

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

Retrieval Models

- A retrieval model
  - abstracts away from the real IR world
  - is a mathematical representation of the essential aspects of a retrieval system
  - ***aims at computing relevance and retrieving relevant documents***
  - ***thus, either explicitly or implicitly, defines relevance***

9

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

The history of Probabilistic Retrieval Models

- Probabilistic models

Probabilistic indexing	(1960)
Robertson/Spärck Jones Rel Model	(1976)
Two-Poisson model –> BM25	Okapi
Bayesian inference networks	Indri
Statistical language models	Lemur

- Citation analysis models

Hubs & authorities	Clever (1998)
PageRank	Google

10

Table of Contents

- Background
  - The need for mathematical IR models
  - **Key IR problems and motivation for risk management**
- Individualism in IR
  - RSJ model, Language modeling
  - Probability Ranking Principle
- Ranking in context and diversity
  - Loss functions for diverse ranking
  - Less is More, Maximum Marginal Relevance, Diversity Optimization
  - Bayesian Decision Theory
- Portfolio Retrieval
  - Document ranking
  - Risk-reward evaluation methods
  - Query expansion and re-writing
- Future challenges and opportunities

The Retrieval Problem

- Suppose have *N* documents in a collection
  - *N* is big enough, and nobody is able to go through the entire collection
- A user comes, and specifies an information need by textual query *q*
- A "smart" IR system *should be able to say:*

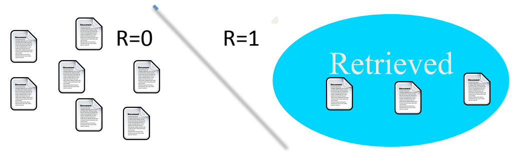
"Based on your request, these are the relevant documents I found for you, and you should read them!"

The Retrieval Problem

- In an ideal world, suppose an IR system can estimate the relevance state with **absolute certainty**
- That is to say it is clever enough to know (predict) the ***exact relevance state*** of each doc in the collection

  $r_k \in \{0,1\}$ denotes the relevance state of document k

  – It could then ***just pick up the docs whose relevant state is 1*** and show them to the user



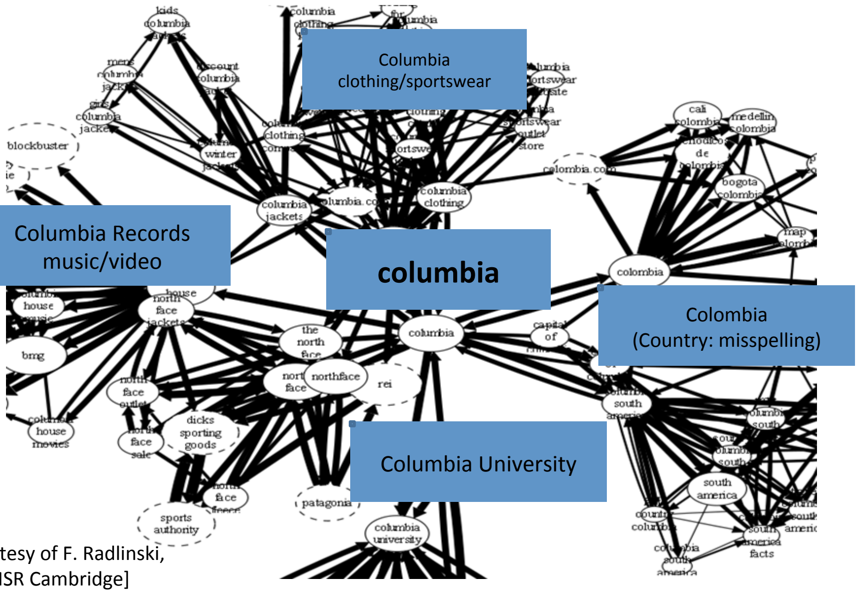
Why IR is so difficult (what are the risks)?

- But, during retrieval, the relevance state is hidden and is difficult to estimate
- Difficulty 1: **unclear about the underlying information needs**
  - still far way from processing the query like

    "*show me the movies I would enjoy this weekend*" or "info helping defining itinerary for a trip to Egypt"
  - thus, queries are usually short -> *ambiguous*

    e.g., issue multiple short queries: "*Egypt*", "*trip Egypt*", "*Egypt hotels*" and examine retrieved docs and gather information

Queries can have ambiguous intents



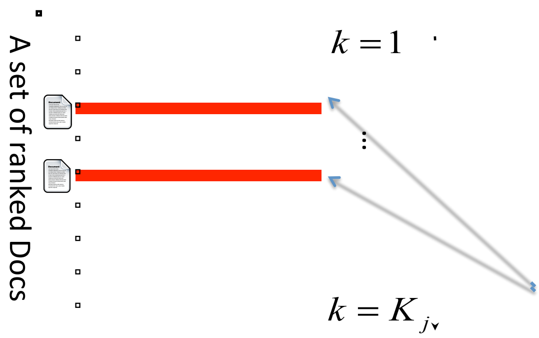
[Courtesy of F. Radlinski, MSR Cambridge]

15

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

Ambiguous queries

- For a given query, we might have different information needs from different users
- By issuing query "*Egypt*", a user might be interested in the recent Egypt uprising and want to know "Egypt" in general

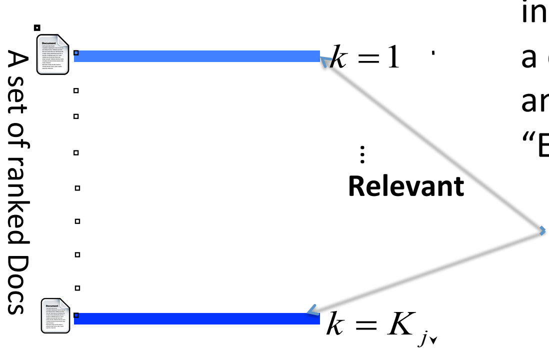


In this case, docs related to politics are perhaps relevant

16

Ambiguous queries

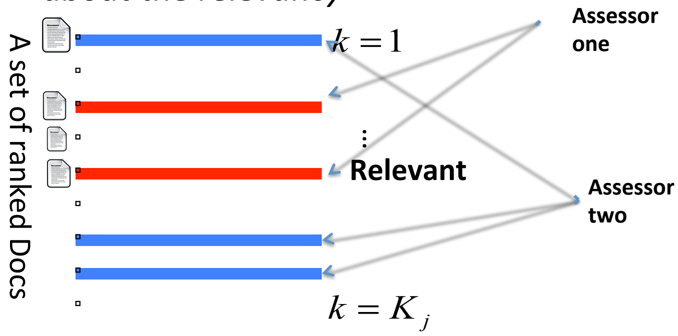
- For a given query, we might have different information needs from different users
- By issuing query "*Egypt*", **another user** might be interested in planning a diving trip in red sea and want to know "Egypt" in general



Perhaps docs related to weather or travel are relevant

Why IR is so difficult?

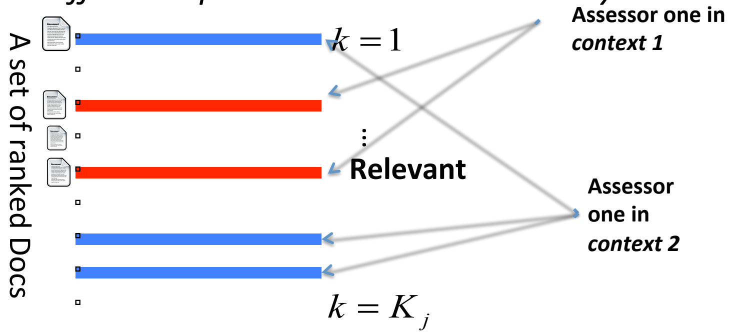
- <u>Difficulty 2 **the uncertainty nature of relevance**</u>
  - Even if the information need is identified exactly, *different users might still have different opinions about the relevancy*



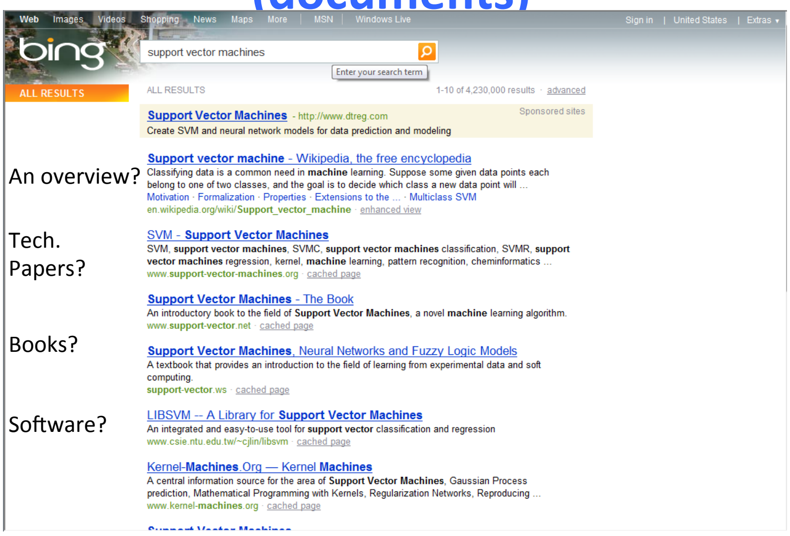
Why IR Models → Individualism → Rank Context → Portfolio Retrieval

Why IR is so difficult?

- Difficulty 2 **the uncertainty nature of relevance**
  - Even if the information need is identified exactly, *same users in different contexts might still have different opinions about the relevancy*



Ambiguity by Result Type (documents)



Why IR is so difficult?

- Difficulty 3 **documents are correlated**
  - *Redundancy:* Some docs are similar to each other
  - *Doc != answers*: have to gather answers from multiple docs

  - *Novelty*: don't want to retrieval something the user already know or retrieved

Difficulties in IR Modelling: Summary

Difficulty 1: Underlying information needs are unclear
Difficulty 2: The uncertain nature of relevance
Difficulty 3: Documents are correlated

Let us first start with Difficulty 2 and try to estimate the relevance as accurately as possible.

(forget about Difficulties 1 and 3, assuming we know the underlying information need exactly, and documents are NOT correlated)

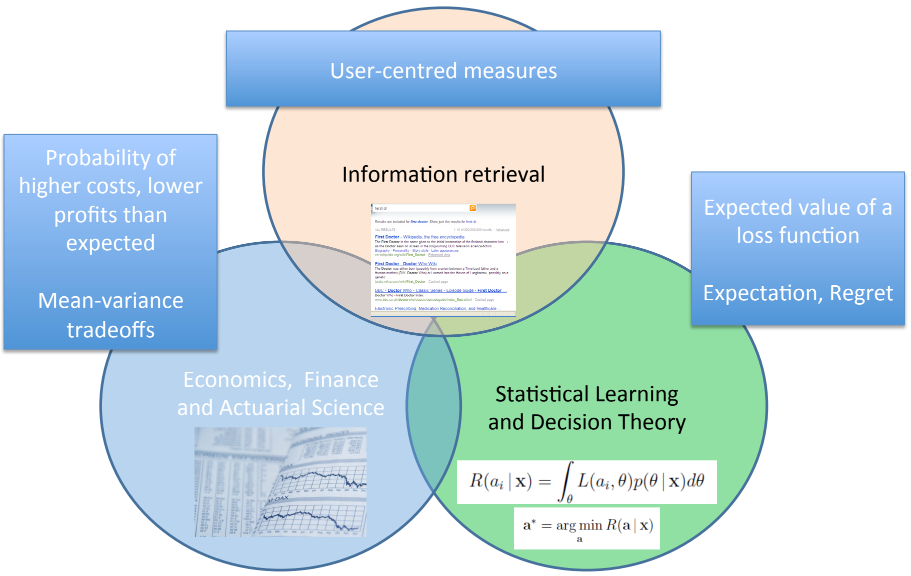
The methodology: we call it individualism in this tutorial

Unified view: motivation

- Why a statistical approach?
- Uncertainty is everywhere in IR
- Uncertainty gives rise to random variables having distributions
  - Can compute mean, variance of distribution
  - Especially interested in distribution over outcomes
- Traditional focus has been on maximizing expectation
  - E.g. average NDCG across queries
- But variance can be critical
  - Especially worst-case: people remember bad errors
  - Risk examples
- Information retrieval: Ranking and query expansion both deal with uncertainty
  - Portfolio theory provides one unified method to help address these problems

23

Risk: statistics of undesirable outcomes



The broad applications of risk management in CIKM fields

- Databases
  - Probabilistic databases
    - Represent correlations between variables or tuples
  - Predicting average and worst-case resource requirements
    - Memory, query execution time, Top-k keyword ranking (large datasets)
- Knowledge management
  - Allocation problems: managing a portfolio of resources
  - Reducing the cost of critical failures
    - Knowledge loss
    - Problem-solving failures
  - Better quality decision models
- Machine learning
  - Variance reduction: reduce training needed; reduce risk of choosing bad model
- Information retrieval (*<u>This tutorial</u>*)
  - Query processing
  - Ranking reliability and diversity

Risk, bias and variance in machine learning

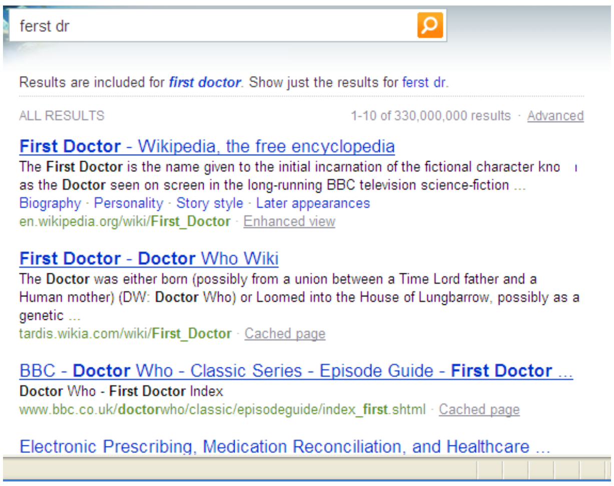
- Across different possible training sets of given size:
  - <u>Bias</u>: how well average prediction of the learning algorithm matches optimal prediction (Bayes rate)
  - <u>Variance</u>: how much the algorithm's prediction fluctuates
  - <u>Squared error</u> is affected by both bias and variance
- Why is variance bad?
  - Increases variance term in bias/variance decomposition so expected accuracy is hurt
  - Increases # of experiments needed for parameter tuning
    - E.g. 50% variance reduction means 1/1.41
  - Creates <u>risk</u> when selecting final model
- All things being equal, lowest-variance model preferred

High risk hurts perceived system quality: User-centric evaluation of recommender systems
[Knijnenburg et al. 2011]

- Maximizing average accuracy is not enough
  - Too little variation is bad
    - Results too similar to each other, with high choice difficulty
  - Some variation is good
    - Diversity, serendipity, lower choice difficulty
  - Too much variation is bad
    - Increased chance of including bad results
- Risky recommender systems result in lower perceived system quality for users
  - Screwing up a lot isn't worth it..
  - Even if the system frequently does very well

27

The risk of making (multiple) errors in Web search



Users remember the one spectacular failure,
not the 200 previous successful searches!

28

Some Key Research Questions

- How can we detect risky IR situations? What are effective risk estimation methods and measures?
- How can search engines effectively "hedge" their bets in risky situations?
- When should IR algorithms attempt to find an optimal set of objects instead of scoring objects individually?
- How should we evaluate risk-reward tradeoffs achievable by systems?

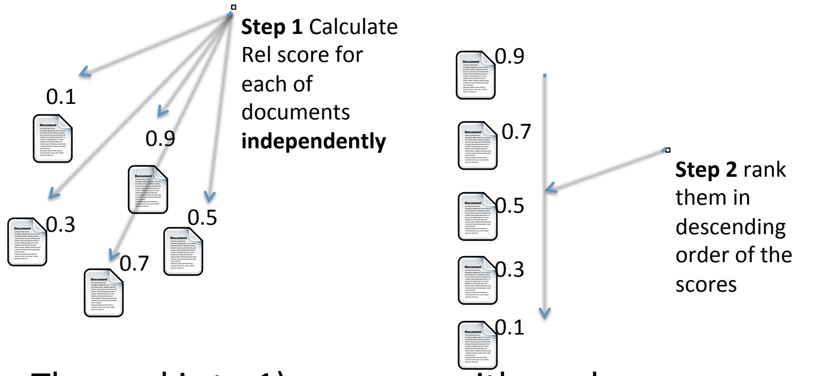
29

Table of Contents

30

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

Methodology: Individualism



- The goal is to 1) come up with a relevance score for each of the documents independently, and 2) to rank them with respect to those scores
- Three models: Relevant Model (RSJ model and BM25), Language Models, and PageRank

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

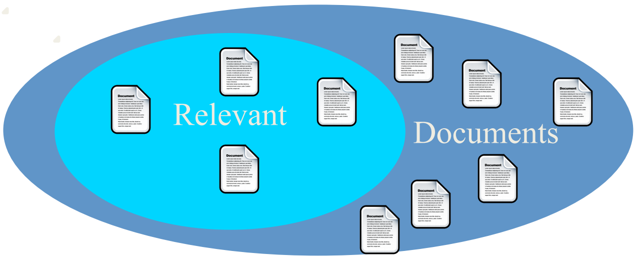
Look back at history: Probability Indexing

- (Maron and Kuhns 1960) "Allows a computing machine, given a request for information, to make a statistical inference and derive a number (called the "relevance number") for each document, which is *a measure of the probability that the document will satisfy the given request*" -> **Calculating Probability of Relevance**
- "The result of a search is an ordered list of those documents which satisfy the request ranked according to their probable relevance".

-> **rank documents based on the scores**

How to calculate Probability of Relevance?

- It depends on the available information
- Suppose given a query, we observed that, out of *N* documents, there are *R* number of relevant documents



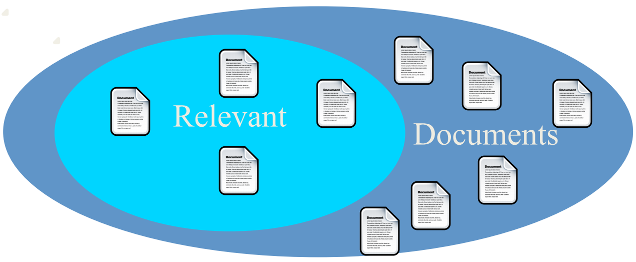
Question: what is the probability a document is relevant (if we randomly pick it up it from the collection)?

$$P(r=1)=\frac{R}{N}$$

Robertson and Spärck-Jones (RSJ) Model

- *logit* of the probability of relevance is commonly used to score a document

$$\log\frac{P(r=1)}{p(r=0)}=\log\frac{R/N}{(N-R)/N}=\log\frac{R}{N-R}$$



Suppose given an information need, you observe:
Num. Docs: *N*
Num. Rel. Docs: *R*

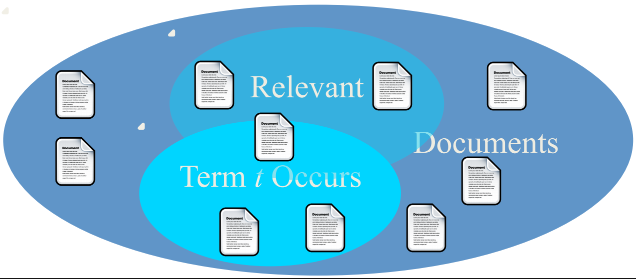
Robertson, S.E.; Sparck Jones, K. (1977), Relevance weighting of search terms, Journal of the American Society for Information Science 27

RSJ Model (joint prob.)

- Now we have some additional observation about individual documents

  i.e., given the query, we also observe that there are n_t documents containing term t (t from a vocabulary)

- $P(t = 1, r = 1|q)$ means the probability that a document is relevant and term t also occurs



Observations:
Num. Docs: N
Num. Docs a term t occurs: n_t
Num. Rel. Docs: R

RSJ Model (scoring function)

Contingency table:

	Relevant	Non-relevant	
Term t Occur	r_t	n_t-r_t	n_t
Term t Not Occur	R-r_t	N-R-n_t+r_t	N-n_t
	R	N-R	N

The *logit* of the probability of relevance:

$$score(d) = \log \frac{P(r = 1 | q, d)}{p(r = 0 | q, d)}$$

In the end, we get

$$score(d) = \sum_{t \in d \cap q} \log \frac{(r_t + 0.5)(N - R - n_t + r_t + 0.5)}{(R - r_t + 0.5)(n_t - r_t + 0.5)}$$

For the detailed derivation, refer to Appendix A

Inverse Document Frequency (IDF)

- In many cases, we have no relevance information
- The collection normally consists of a very large extent of non-relevant documents
- Assuming the all documents are non-relevant

  $R = r_t = 0$ gives $score(d) = \sum_{t \in d \cap q} \log \frac{N - n_t + 0.5}{n_t + 0.5}$

- As n_t is much smaller than N, the above is equivalent to IDF

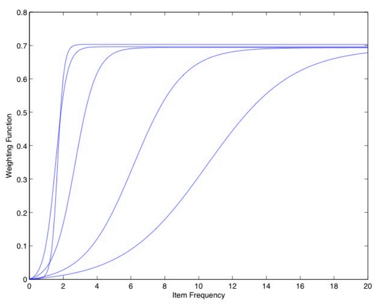
Why IR Models → Individualism → Rank Context → Portfolio Retrieval

BM25: dealing with Term Frequency

- RSJ model does not consider term frequency
- Saturation Function of Term Frequency

$s(tf) = \frac{s_{MAX} \cdot tf}{tf + K}$, tf : term freq in doc

s_{MAX} : max score, K controls the slop



Michaelis–Menten Eq. in Biochemistry

- The BM (Best Math)25 formula

$score(d) = \sum_{t \in d \cap q} \frac{tf}{tf + K} \log \frac{N - n_t + 0.5}{n_t + 0.5}$

K= $k_1((1 - \lambda) + \lambda L_d)$, L_d is the normalized doc length

(i.e. the length of this doc d divided by the avg. len. of docs).

$\lambda \in [0, 1]$ and k_1 are constant.

Relevance Model (RSJ Model): Summary

Relevance model: estimate the relevance between a user need and a document

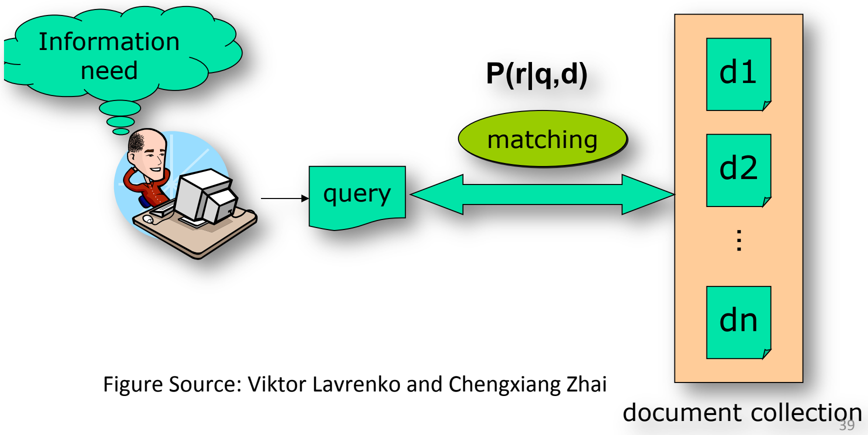


Figure Source: Viktor Lavrenko and Chengxiang Zhai

Language Model

Language Model: construct a document *model* and see how likely a query can be generated from it

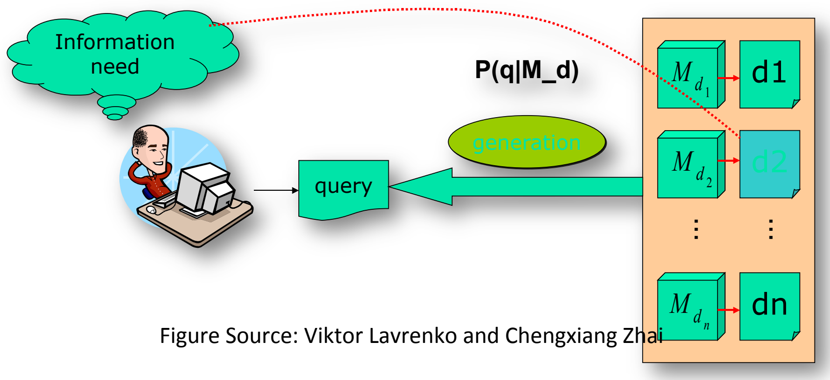
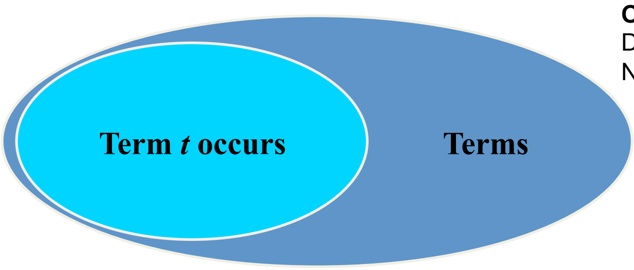


Figure Source: Viktor Lavrenko and Chengxiang Zhai

Language Models (Conditional Independence)

$$score_q(d) = \log p(q \mid d) = \log \prod_{t \in q} p(t \mid d) = \sum_{t \in q} \log \frac{tf_t}{L_d}$$

Give the document *d*:



Observations:
Doc Length: L_d
Num. Occurrences: tf_t

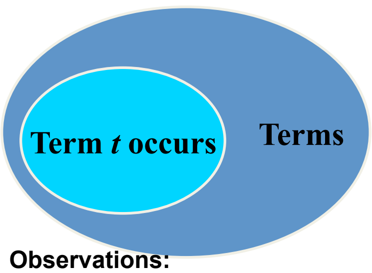
Language Models (para. esti.)

- Linear Smoothing LM models

$$score_q(d) = \log p(q \mid d) = \log \prod_{t \in q} p_{LS}(t \mid d) = \sum_{t \in q} \log \frac{tf_t}{L_d}$$

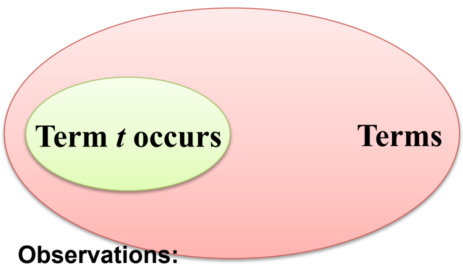
$$= \sum_{t \in q} \log(\lambda \frac{tf_t}{L_d} + (1-\lambda)\frac{tf_{t,c}}{L_c}),\ \lambda \in [0,1] \text{ is constant}$$

Give the document *d*:



Observations:
Doc Length: L_d; Num. Occurrences: tf_t

Give the collection *c*



Observations:
Doc Length: L_c ; Num. Occurrences: $tf_{t,c}$

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

Language Models (para. esti.)

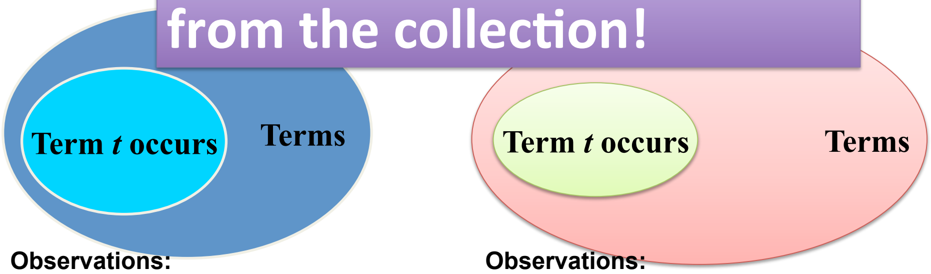
- Linear Smoothing LM models

$$score_q(d) = \log p(q \mid d) = \log \prod_{t \in q} p_{LS}(t \mid d) = \sum_{t \in q} \log \frac{tf_t}{L_d}$$

$$= \sum_{t \in q} \ldots$$

Give the

Smoothing the estimation from the collection!



Observations:
Doc Length: L_d; Num. Occurrences: tf_t

Observations:
Doc Length: L_c ; Num. Occurrences: $tf_{t,c}$

43

Table of Contents

- Background
  - The need for mathematical IR models
  - Key IR problems and motivation for risk management
- **Individualism in IR**
  - RSJ model, Language modeling
  - **Probability Ranking Principle**
- Ranking in context and diversity
  - Loss functions for diverse ranking
  - Less is More, Maximum Marginal Relevance, Diversity Optimization
  - Bayesian Decision Theory
- Portfolio Retrieval
  - Document ranking
  - Risk-reward evaluation methods
  - Query expansion and re-writing
- Future challenges and opportunities

44

Understanding individualism

- **Probability Ranking Principle**: "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of ***decreasing probability of usefulness to the user***... then the ***overall effectiveness of the system to its users*** will be the best obtainable on the basis of that data"

William S. Cooper. The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. *University of California, Berkeley*, 1971.

S. E. Robertson. The probability ranking principle in IR. *Readings in information retrieval*, pages 281–286, 1997.

Assumptions in PRP

- Document relevancy (or usefulness) is binary
- Probability of relevance can be obtained with certainty -> extension: Probability of Probability (a Bayesian viewpoint)
- The relevance of a document is independent of other documents in the collection

We will show (in next few slides) that under the assumptions PRP maximizes expected Precision or minimizes expected search length

Maximizing Expected Precision

- Given a request (query), suppose we retrieve *n* documents $\{r_1,...,r_j,...r_n\}$, where $r_j \in \{0,1\}$ is binary relevance
- Precision: $P = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Retrieved}|} = \frac{n_r}{n} = \frac{\sum_j r_j}{n}$
- Expected Precision@n:

$$E[P] = \frac{[n_r]}{n} = \frac{\sum_j [r_j]}{n} = \frac{\sum_j p(r_j = 1)}{n}, \text{ where } p(r_j = 1) \text{ Prob of rel at rank } j$$

Recall we assume the rel. of doc. is independent with each other

- Therefore, the optimal strategy is to retrieve the *n* documents which have the largest probabilities of relevance $p(r_j = 1)$

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

Minimizing Expected Search Length

- **Search Length**: how many non-relevant docs encountered before seeing the first relevant
- **Expected Search Length** is the summation of all possible search lengths weighted by their respective probabilities:

$$E[L] = \sum_j ((j-1)p(r_j = 1, r_1 = 0, ..., r_{j-1} = 0))$$

$$= \sum_j ((j-1)p(r_j = 1)\prod_{i=1}^{j-1} p(r_i = 0)) \quad \text{<= independent assumption}$$

$$= 0p(r_1 = 1) + 1p(r_2 = 1)p(r_1 = 0)...$$

Minimizing Expected Search Length

- **Search Length**: how many non-relevant docs encountered before seeing the first relevant
- **Expected Search Length** is the summation of all possible search lengths weighted by their respective probabilities:

$$E[L]=\sum_j ((j-1)p(r_j=1,r_1=0,\ldots,r_{j-1}=0))$$

$$=\sum ((j-1)p(r_j=1)\prod^{j-1} p(r_i=0)) \quad \text{<= independent assumption}$$

Again, the optimal ranking strategy is to place the documents having larger probabilities of relevance in the lower rank

But why if we are not sure about Information Need (difficulty 1)

- Suppose we have query "egypt", and two classes of users: *U1: Egypt_politics* and *U2: Egypt_travel*; U1 has twice as many members as U2
- An IR system retrieved three docs d1,d2 and d3 and their probs of relevance are as follows:

UserClass	D1:*Egypt_politics*	D2:*Egypt_politics*	D3:*Egypt_travel*
Egypt_politics	1	1	0
Egypt_travel	0	0	1
P(r)	2/3	2/3	1/3

But why if we are not sure about Information Need (difficulty 1)

UserClass	D1:*Egypt_politics*	D2:*Egypt_politics*	D3:*Egypt_travel*
Egypt_politics	1	1	0
Egypt_travel	0	0	1
P(r)	2/3	2/3	1/3

- PRP gives

d1 *Egypt_politics* | d2 *Egypt_politics*
d2 *Egypt_politics* or d1 *Egypt_politics*
d3 *Egypt_travel* | d3 *Egypt_travel*

It is NOT optimal as U2 group has to reject two docs before reaching the one it wants

Expected Search Length: assuming independence as in PRP

- {d1,d2,d3}:

d1 *Egypt_politics*
d2 *Egypt_politics*
d3 *Egypt_travel*

$$E[L] = 0p(r(d_1)=1) + 1p(r(d_2)=1)p(r(d_1)=0) + 2p(r(d_3)=1)p(r(d_2)=0)p(r(d_1)=0)$$
$$=0(2/3)+1(2/3)(1/3)+2(1/3)(1/3)(1/3)=8/27$$

- {d1,d3,d2}

d1 *Egypt_politics*
d3 *Egypt_travel*
d2 *Egypt_politics*

$$E[L] = 0p(r(d_1)=1) + 1p(r(d_3)=1)p(r(d_1)=0) + 2p(r(d_2)=1)p(r(d_3)=0)p(r(d_1)=0)$$
$$=0(2/3)+1(1/3)(1/3)+2(2/3)(2/3)(1/3)=11/27$$

- {d1,d2,d3) is better

Expected Search Length: consider the dependency

- {d1,d2,d3}:

d1 *Egypt_politics*

d2 *Egypt_politics*

d3 *Egypt_travel*

$$E[L] = 0p(r(d_1)=1) + 1p(r(d_2)=1, r(d_1)=0) + 2p(r(d_3)=1, r(d_2)=0, r(d_1)=0)$$
$$=0(2/3)+1(0)+2(1/3)=2/3$$

- {d1,d3,d2}

d1 *Egypt_politics*

d3 *Egypt_travel*

d2 *Egypt_politics*

$$E[L] = 0p(r(d_1)=1) + 1p(r(d_2)=1, r(d_1)=0) + 2p(r(d_3)=1, r(d_2)=0, r(d_1)=0)$$
$$=0(2/3)+1(1/3)+2(0)=1/3$$

- {d1,d3,d1) is better!

Individualism (PRP): Summary

- Limitations:
  - Documents are dependent with respect to their relevancy (due to difficulty 1 and/or 2)
- In spite of the limitations, PRP has been influential in retrieval modelling
- Many interesting research questions:
  - How to model uncertainty with probability estimation -> Bayesian approach
  - How to tackle the dependency -> Portfolio theory

Table of Contents

Maximum Marginal Relevance

[Carbonell and Goldstein (1998)]

- When we have many potentially relevant docs, the relevant ones :
  - may be highly redundant with each other
  - might contain partially or fully duplicated information (Instance of IR problem #3)
- Idea: Select documents according to a combined criterion of *query relevance* and *novelty of information*

Maximum Marginal Relevance

- A linear combination of relevancy and novelty:
  - Novelty is measured by dissimilarity between the candidate doc and previously retrieved ones already in the ranked list
  - Relevance is measured by similarity to the query

Find a doc at rank *j* that maximizes

$$\lambda Sim_1(d_j, q) - (1-\lambda)\max_{\forall d_i: i \in \{1, j-1\}} Sim_2(d_i, d_j),$$

where $\lambda \in [0,1]$ is a constant, Sim is similarity measure

- A document has high marginal relevance if it is both relevant to the query and contains minimal similarity to previously selected documents.

Less is more Model

[Chen&Karger 2006]

- A *risk-averse* ranking that ***maximizes the probability that at least one of the documents is relevant.***
- Assumes *previously retrieved documents are non-relevant* when calculating relevance of documents for the current rank position $p(r_j = 1 | r_{j-1} = 0)$, where *j* is the rank
- Metric: *k-call* @ N
  - Binary metric: 1 if top n results has k relevant, 0 otherwise
- Better to satisfy different users with different interpretations, than one user many times over.
- "Equivalent" to maximizing the Reciprocal Rank measure or minimizing the expected Search Length

Less is More

- Suppose we have two documents. The objective to be maximized is:

$$1 - p(r_1 = 0, r_2 = 0)$$
$$= p(r_1 = 1, r_2 = 0) + p(r_1 = 0, r_2 = 1) + p(r_1 = 1, r_2 = 1)$$
$$= p(r_1 = 1) + p(r_1 = 0) p(r_2 = 1 | r_1 = 0))$$

- To maximize it, a greedy approach is to
  - First choose a document that maximizes $p(r_1=1)$;
  - Fix the doc at rank 1, and then select the second doc so as to maximize $p(r_2 = 1 | r_1 = 0)$.

Less is More

- A similar analysis shows that we can select the third document by maximizing
  $p(r_3 = 1 | r_2 = 0, r_1 = 0)$
- In general, we can select the optimal *i*-th document in the greedy approach by choosing the document *d* that maximizes
  $p(r_j = 1 | r_{j-1} = 0, \ldots, r_1 = 0)$
- Intuition: if none of previously retrieved docs is relevant, what else can we get – **keep adding additional insurance!**
- As a result, it diversifies the rank list.
- *Expected Metric Principle* (EMP):
  - maximize $E[\text{metric} | d_1 \ldots d_n]$ for complete result set

Ranking with Quantum `Interference'

- Implicitly captures dependencies between documents through `quantum interference'
- Find a document d that maximizes:

$$S(d) = \left(P(d) - \sum_{d' \in RA} \sqrt{P(d)}\,\sqrt{P(d')}\cos\theta_{d,d'} \right)$$

where *RA* is the set of previous docs in the ranking

- Recent work on connections to portfolio theory [Zuccon, Azzopardi, van Rijsbergen SIGIR 2010]
  - Interference term is like portfolio document correlation term

61

Increasing interest in learning complex structured outputs (including ranking)

- Radlinski *et al.,* ICML '08
  - Minimize abandonment with multi-armed bandits
- Gollapudi *et al.,* WSDM '08
  - Greedy minimization of a submodular formulation based on relevance and utility to user. Assumption that conditional relevance of documents to a query is independent.
- Gollapudi *et al.,* WWW '09
  - 8 desired axioms for diversification (e.g. strength of relevance, strength of similarity), impossibility results for all 8, and investigation of some instantiations

62

Learning to Rank: Drawbacks

Relevance Prediction

Rank Optimization

- Focusing on IR metrics and Ranking
  - bypass the step of estimating the relevance states of individual documents
  - construct a document ranking model from training data by ***directly*** optimizing an IR metric [Volkovs&Zemel 2009]
- However, not all IR metrics necessarily summarize the (training) data well; thus, training data may not be fully explored. [Yilmaz&Robertson2009]

Table of Contents

Bayesian Decision Theory in LM

- Bayesian Decision Theory
  - is a fundamental statistical approach
  - quantifies the tradeoffs between various decisions using probabilities and costs/risk that accompany such decisions
- State of relevance is a random variable
  - r=1 for relevance
  - r=0 for non-relevance
  - P(r=1|d,q) is the probability that the document is relevant to a given query.
  - P(r=0|d,q) = 1 - p(r=1|d,q) is the prob. that the document is not relevant

Bayesian Decision Theory in LM

- We now define a decision *a*
  - *a=1*: retrieve the doc, and *a=0*: not retrieve it
- For a given query, suppose we observe a doc d and take action a=1 (retrieve it)
  - *Note that in this example we do not take other documents into consideration when making a decision*
- If the true state is r, we incur the conditional loss:

$$Loss(a=1|r)=\begin{cases} c1 & r=1 \\ c2 & r=0 \end{cases},\ c2>c1$$

Bayesian Decision Theory in LM

- Then, the expected loss of taking action *a=1*:

$$E[Loss(a=1|q,d)] = \sum_r Loss(a=1|r)p(r|q,d)$$

$$= Loss(a=1|r=1)p(r=1|q,d) + Loss(a=1|r=0)(1-p(r=1|p,q))$$

$$= -(c2-c1)p(r=1|q,d)+c2$$

- Minimizing it would pick up the document which has the highest probability of relevance p(r=1|q,d)
- Thus rank in ascending order of expected loss is equivalent to that in descending order of prob. of relevance

A Generative IR model

[John Lafferty, Chengxiang Zhai SIGIR 2001]

A User first selects a query model

Model selection

Query generation

A query is then generated from that query model

$$\mathcal{U} \xrightarrow[p(\theta_Q|\mathcal{U})]{} \theta_Q \xrightarrow[p(q|\theta_Q)]{} q$$

A Source/ author first selects a doc model

Model selection

Document generation

A doc is then generated from that doc model

$$\mathcal{S} \xrightarrow[p(\theta_D|\mathcal{S})]{} \theta_D \xrightarrow[p(d|\theta_D)]{} d$$

John Lafferty, Chengxiang Zhai Document language models, query models, and risk minimization for information retrieval SIGIR 2001

ChengXiang Zhai, John Lafferty A risk minimization framework for information retrieval, IP&M, 2006

A Generative IR model

The Relevance State depends on the two models

$$
\begin{tikzcd}
\mathcal{U} \arrow[r, "\textit{Model selection}", "p(\theta_Q \mid \mathcal{U})"'] & \theta_Q \arrow[rr, "\textit{Query generation}", "p(\mathbf{q} \mid \theta_Q)"'] \arrow[dr] & & \mathbf{q} \\
& p(R \mid \theta_Q, \theta_D) & R & \\
\mathcal{S} \arrow[r, "\textit{Model selection}", "p(\theta_D \mid \mathcal{S})"'] & \theta_D \arrow[rr, "\textit{Doc generation}", "p(\mathbf{d} \mid \theta_D)"'] \arrow[ur] & & \mathbf{d}
\end{tikzcd}
$$

John Lafferty, Chengxiang Zhai Document language models, query models, and risk minimization for information retrieval SIGIR 2001

ChengXiang Zhai, John Lafferty A risk minimization framework for information retrieval, IP&M, 2006

Understanding Lafferty and Zhai's model

- A general and principled IR model
- A point estimation is used in the formulation.

$$\int_{\theta_d, \theta_d} p(r \mid \theta_q, \theta_d) p(\theta_q \mid q) p(\theta_d \mid d) d\theta_d \, d\theta_a \approx p(r \mid \hat{\theta}_q, \hat{\theta}_d)$$

  - **the dependency therefore is modeled by the loss function not relevance probability**
- Various dependent loss functions are defined to incorporate various ranking strategy

  ChengXiang Zhai, John Lafferty A risk minimization framework for information retrieval, IP&M, 2006

- Two challenges are remaining in the model:
  - the risk of understanding user information need is not covered from the point estimation. explore the potential of a full Bayesian treatment
  - explore $p(r \mid \theta_q, \theta_d)$ (Victor Lavrenko and W. Bruce Croft, Relevance-Based Language Models, SIGIR 2001)

Another formulation using Bayesian Decision theory [Wang and Zhu SIGIR2010]

Given an IR Metric

Predicting Relevance

Rank Optimization

- "to estimate the relevance of documents as accurate as possible,
- and to summarize it by the ***joint*** probability of documents' relevance
- dependency between documents is considered

- Rank preference specified, by an IR metric.
- The rank decision making is a ***stochastic*** one due to the ***uncertainty about the relevance***
- As a result, the optimal ranking action is the one that ***maximizes the expected value of the IR metric***

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

Yet Another formulation using Bayesian Decision theory [Wang and Zhu SIGIR2010]

Given an IR Metric

Predicting Relevance

Rank Optimization

- "to estimate the relevance of documents as accurate as possible,
- and to summarize it by the ***joint*** probability of

- Rank preference specified, by an IR metric.
- The rank decision making is a ***stochastic*** one due to the ***uncertainty about the***

- **The cost is best defined by the used IR metric!**

mal one that ***cted value of the IR metric***

The statistical document ranking process

The joint Probability of Relevance given a query: $p(r_1,\ldots,r_N|q)$

$(r_1,\ldots,r_N) \rightarrow$

The effectiveness of a rank action $(a_1,\ldots,a_N)$: $m(a_1,\ldots,a_N|\ r_1,\ldots,r_N\ )$

$$\hat{a} = \arg\max_a E(m\,|\,q)$$

$$= \arg\max_{a_1,\ldots,a_N} \Big(\sum_{r_1,\ldots,r_N} m(a_1,\ldots,a_N\,|\,r_1,\ldots,r_N)\, p(r_1,\ldots,r_N\,|\,q) \Big)$$

IR metric:
Input:
1. **A rank order** $a_1,\ldots,a_N$
2. **Relevance of docs.** $r_1,\ldots,r_N$

The joint probability of relevance given a query

J. Wang and J. Zhu, "On Statistical Analysis and Optimization of Information Retrieval Effectiveness Metrics," in Proc. of SIGIR 2010.

The statistical document ranking process

The joint Probability of Relevance given a query: $p(r_1,\ldots,r_N|q)$

$(r_1,\ldots,r_N) \rightarrow$

The effectiveness of a rank action $(a_1,\ldots,a_N)$: $m(a_1,\ldots,a_N|\ r_1,\ldots,r_N\ )$

$$\hat{a} = \arg\max_a E(m\,|\,q)$$

$$= \arg\max_{a_1,\ldots,a_N} \Big(\sum_{r_1,\ldots,r_N} m(a_1,\ldots,a_N\,|\,r_1,\ldots,r_N)\, p(r_1,\ldots,r_N\,|\,q) \Big)$$

The above equation is computationally expensive! This leads to the *Portfolio theory of IR* using *Mean* and *Variance* to summarize the joint probably of relevance for all the docs in the collection.

y of given a

J. Wang and J. Zhu, "On Statistical Analysis and Optimization of Information Retrieval Effectiveness Metrics," in Proc. of SIGIR 2010.

Table of Contents

Difficulties in IR Modelling: Summary

Difficulty 1 Unclear about the underlying information needs

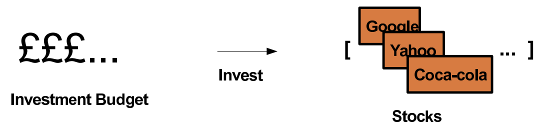
Difficulty 2 The uncertain nature of relevance

Difficulty 3 Documents are correlated

To address them all, *ranking under uncertainty* is not just about picking individual relevant documents

Methodology: Portfolio Retrieval

- A more general methodology: *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

What can we learn from finance?

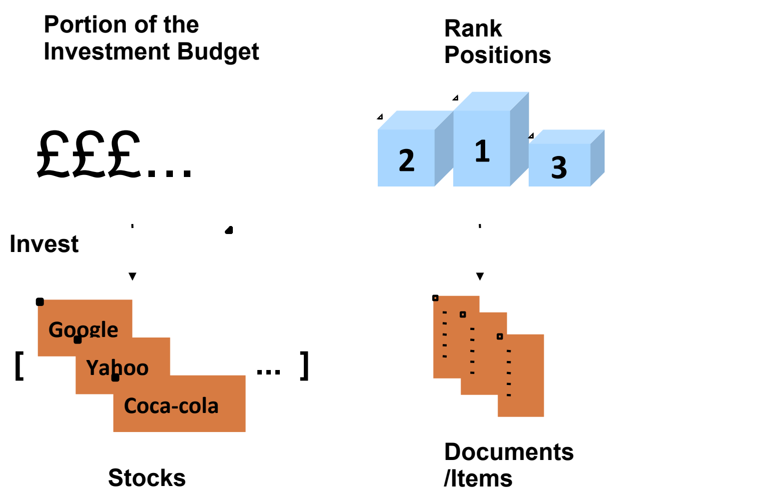
- Financial markets:
  - Place relevant buyers and sellers in one common place
  - make it easy for them to find each other
  - efficient allocation of resources
- The Web essentially does the same thing
  - Information Retrieval: **efficient supply-demand match**
  - expanded accessibility of web resources *by* ***separating the use and ownership (online advertising and search)***

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

Web search (rank positions as investment opportunities)

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

Portfolio Theory of Information Retrieval

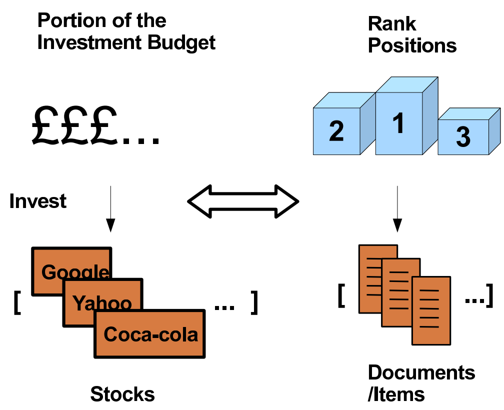


J. Wang and J. Zhu, "Portfolio Theory of Information Retrieval," in SIGIR, 2009.

80

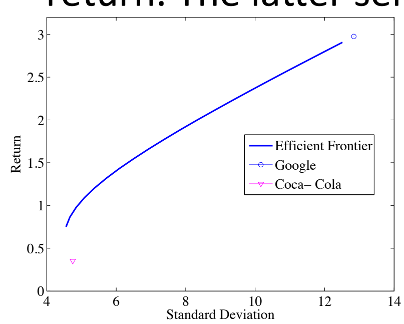
The analogy

- 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)]

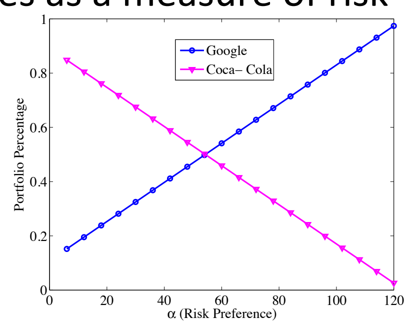


Mean and variance

- 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

Back to the simplified IR problem: using the Portfolio Retrieval formulation

- Suppose an IR system is clever enough to know the relevance state of each doc exactly
- It could then just pick up the relevant docs and show them to the user
- Formulate the process by the portfolio idea

$$\{\hat{w}_j\} = \operatorname{argmax}_{\{w_j\}} o_n = \sum_{j=1}^{n} w_j r_j \ , \ w_i \in \{0,1\}$$

where $r_j \in \{0,1\}$ denotes the relevance state of document j

w_j denotes the decision whether show the document j to the user or not

o_n denotes the number of relevant documents

So the solution: w_j=1 when r_j=1

Portfolio Retrieval formulation (ranked list&graded relevance)

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

$$R_n = \sum_{j=1}^{n} w_j r_j$$

where R_n denotes the overall relevance of a ranked list. Variable w_n differentiates the importance of rank positions. r_j is the rel. score of a doc at j, where j = {1, ..., n}, for each of the rank positions

Portfolio Retrieval formulation (ranked list&graded relevance)

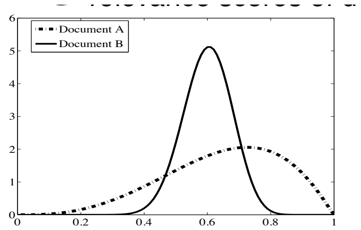
- Weight w_j is similar to the discount factors in IR evaluation in order to penalize late-retrieved relevant documents [Järvelin and Kekäläinen (2002)]

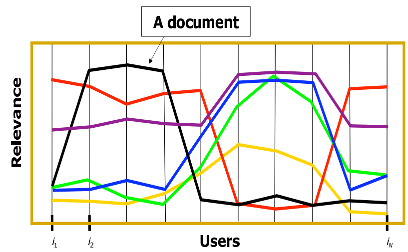
$$w_j = \frac{1}{2^{j-1}} \text{ where } j \in \{1, \ldots, n\}$$

- It can be easily shown that when $w_1 > w_2 \ldots > w_3$, the maximum value of R_n gives the ranking order $r_1 > r_2 \ldots > r_n$
- This follows immediately that maximizing R_n – 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

Difficulties in IR Modelling: Summary

1. Unclear about underlying information needs
2. The uncertain nature of relevance
3. Documents are correlated





Portfolio Retrieval formulation (uncertainty)

- During retrieval, the overall relevance R_n CANOT be calculated with certainty
- Quantify a ranked list based on its expectation (*mean* $E[R_n]$) and its *variance* ($Var(R_n)$):

$$E[R_n] = \sum_{j=1}^{n} E[r_j],\ \mathrm{Var}[R_n] = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j c_{i,j}$$

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

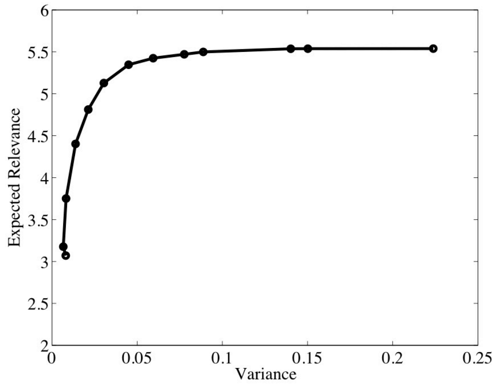
- Now two quantities to summarize a ranked list

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* $On = E[R_n] - bVar(R_n)$

Portfolio Retrieval

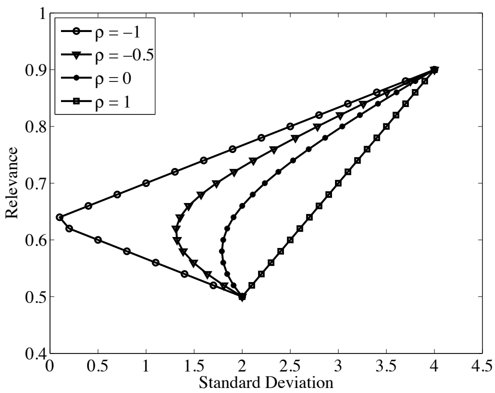
- The Efficient Frontier:



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

A mathematical model of diversification

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



Why IR Models → Individualism → Rank Context → Portfolio Retrieval

A mathematical model of diversification

- **Variance (risk) = uncertainty of individual doc rel predictions + correlations between doc rel predictions**

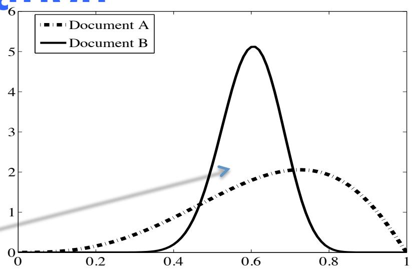
$$Var(R_n) = \sum_j w_j^2 c_{i,i} + 2\sum_i \sum_{j=i+1} w_i w_j c_{i,j}$$

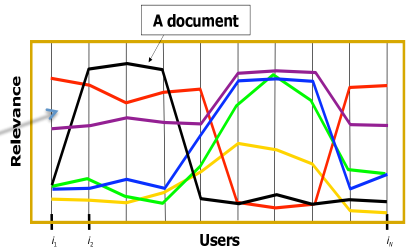
$$= \sum_j w_j^2 \sigma_i^2 + 2\sum_i \sum_{j=i+1} w_i w_j \sigma_i \sigma_j \rho_{i,j}$$

where $\sigma_i = \sqrt{c_{i,i}}$ is the standard

deviation and $\rho_{i,j} = \frac{c_{i,j}}{\sigma_i \sigma_j}$ is

the correlation coefficient





Diversification -> negative correlation -> reduce the risk

91

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

The Practical Algorithm

- Unlike in finance, the weight w_n in IR, representing the discount for each rank position, is a discrete variable
- Therefore, the objective function is *no-smooth*
- A greedy approach: first consider rank 1, and then add docs to the ranked list sequentially until reaching the last rank position *n*
- Select a document at rank *j* that has the maximum value of:

$$E[r_j] - b w_j \sigma_j - 2b \sum_{j=1}^{j-1} w_i w_j \sigma_i \sigma_j \rho_{i,j},\ b \text{ is a parameter adjusting the risk}$$

92

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

The Practical Algorithm

- Select a document at rank *j* that has the maximum value of:

$$E[r_j] - bw_j\sigma_j - 2b\sum_{j=1}^{j-1} w_i w_j \sigma_i \sigma_j \rho_{i,j}$$

b is a parameter adjusting the risk

Relevance Score of a given candidate doc

Uncertainty of your estimation

Correlations between the candidate doc and previously retrieved docs

Could be the probability of relevance or rel scores from any IR models

For the study of variance see Zhu, J. Wang, M. Taylor, and I. Cox, "Risky Business: Modeling and Exploiting Uncertainty in Information Retrieval," in Proc. Of SIGIR, 2009.

Correlation with respect to the topics/terms/ information needs/users

93

Latent Factor Portfolio

- The relevance values of documents are correlated due to the underlying factors, for example
  - if query "earthquake" when Tsunami hits Japan, documents related to that event (topic) are likely to be more relevant than anything else
  - In recommender systems, some people like action movies more than dramas
- It is, thus, interesting to understand how documents are correlated with respect to the underlying topics or factors Portfolio + Latent Topic models (pLSA)
- In addition, the computation of obtaining the covariance matrix can be significantly reduced.

Yue et al. Latent Factor Portfolio for Collaborative Filtering under submission 2011

94

Latent Factor Portfolio

- Its expected value:

$$\hat{R}_n = \sum_{i=1}^{n} w_i \int_r rp(r \mid d_i, q)dr$$

$$= \sum_{i=1}^{n} w_i \sum_{a=1}^{A} \int_r rp(r \mid a, q)dr(p(a \mid d_i))$$

$$= \sum_{a=1}^{A} \hat{r}(a,q)(\sum_{i=1}^{n} w_i p(a \mid d_i))$$

The distribution of relevant topics

The contribution from the docs

where *a* denotes topics – we have *A* number of topics. *q is* the query and d_j denotes the doc at rank *j*

Yue et al. Latent Factor Portfolio for Collaborative Filtering under submission 2011

Table of Contents

Evaluation methods that account for risk and variance

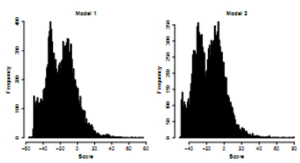
- My new query re-writing algorithm gets:
  - An average 2-point NDCG gain! Ship it! Right?
  - OR: No perceptible NDCG gain over the old one. Scrap it! Right?
- Measuring average gain across queries is not enough when deploying a risky retrieval algorithm.
- Including variance is essential to understand likely effect on users.

97

Reducing variance in Web search ranking

[Ganjisaffar, Caruana, Lopes. SIGIR 2011]

- Core ranking uses boosting: high accuracy, high variance
- Use bagging to reduce variance
  - Train different models on different sampled training sets
  - Then normalize & combine their outputs

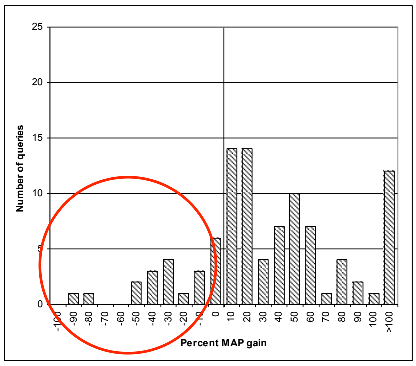


Distribution of ranker scores on validation set
for two different subsamples of the same training set and size

98

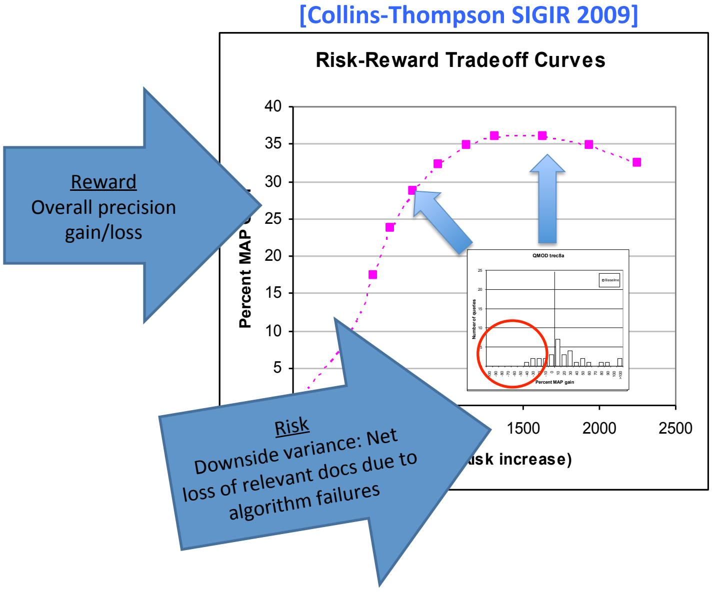
Helped-Hurt Histograms: Distribution of success/failure, with focus on downside variance

- Net loss of relevant docs to algorithm failures
- Many flavors possible:
  - R-Loss @ *k*: Net loss in top *k* documents
  - R-Loss: averaged R-Loss @ *k* (analogous to AP)

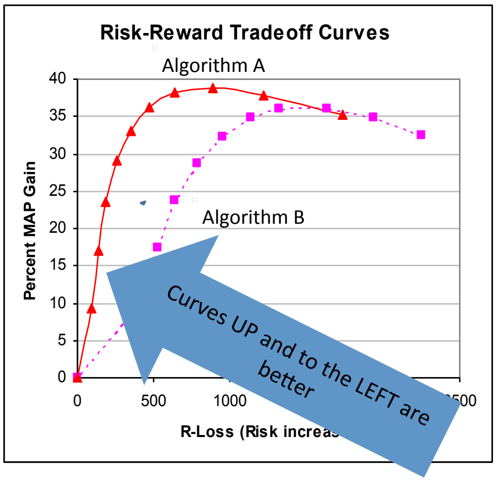


Risk-reward curves capture mean-variance tradeoff achieved by a retrieval algorithm

[Collins-Thompson SIGIR 2009]



Algorithm A dominates algorithm B



High Relevance
Low Risk

High Relevance
High Risk

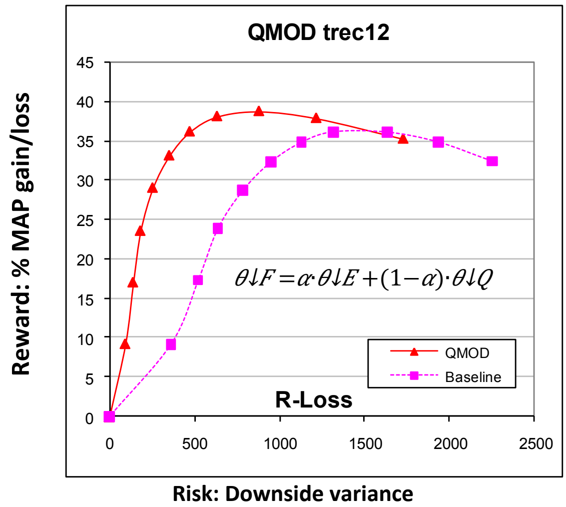
Expected
Relevance

Low Relevance
Low Risk

Low Relevance
High Risk

Variance (Risk)

The typical risk/reward of query expansion: as interpolation parameter α varies



103

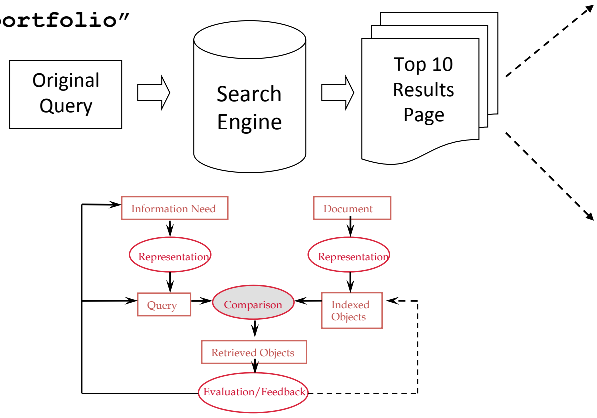
Table of Contents

- Background
  - The need for mathematical IR models
  - Key IR problems and motivation for risk management
- Individualism in IR
  - RSJ model, Language modeling
  - Probability Ranking Principle
- Ranking in context and diversity
  - Loss functions for diverse ranking
  - Less is More, Maximum Marginal Relevance, Diversity Optimization
  - Bayesian Decision Theory
- **Portfolio Retrieval**
  - Document ranking
  - Risk-reward evaluation methods
  - **<u>Query expansion and re-writing</u>**
- Future challenges and opportunities

104

What if the user's query and the author's document terms don't match? Or match incorrectly?

"picking the best stock market portfolio"



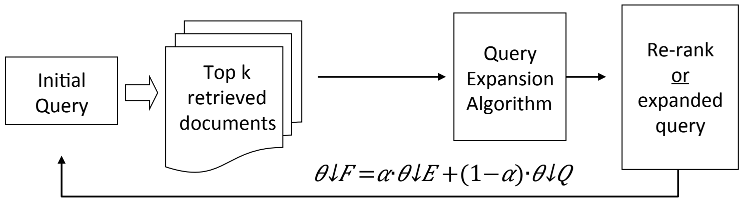
It's easier to <u>**choose**</u> the **optimal set** of <u>**equities**</u> to <u>buy</u> if you know your tolerance for risk in the <u>market</u>

If you want to <u>market</u> your skills you can build your own <u>**portfolio**</u> of <u>**stock**</u> photographs by <u>**choosing**</u> the <u>**best**</u> ones in your collection…

Goal: Improve retrieval quality by estimating a more complete representation of the user's information need

"mercedes car repair"

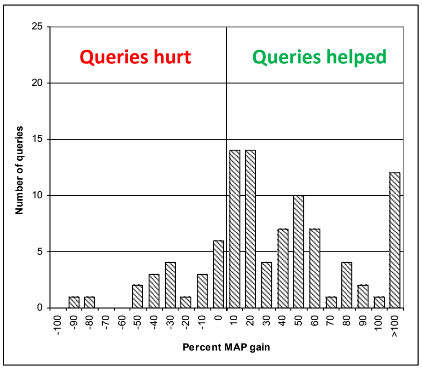
\+ auto
automobile
shop
benz
service



$$\theta_F = \alpha \cdot \theta_E + (1-\alpha) \cdot \theta_Q$$

Current methods perform a type of simple risk-reward tradeoff by interpolating the expansion model with the initial query model

Current query expansion methods work well on average…

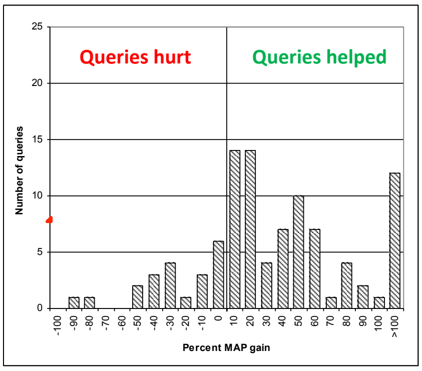


Query expansion:
Current state-of-the-art method

Mean Average Precision gain: +30%

107

…but exhibit high variance across individual queries

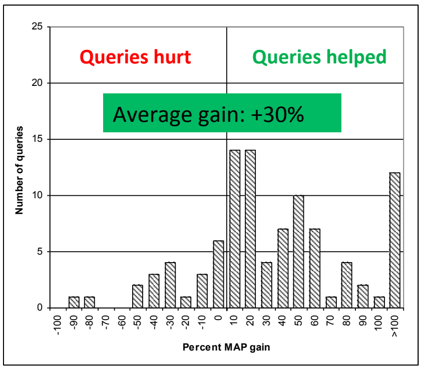


Query expansion:
Current state-of-the-art method

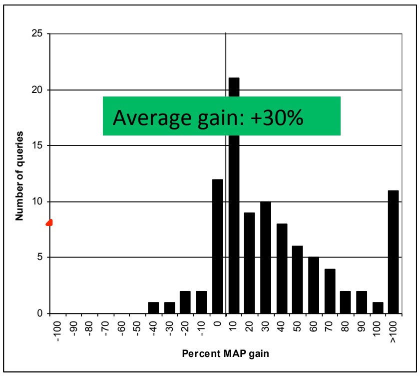
This is one of the reasons that even state-of-the-art algorithms are impractical for many real-world scenarios.

108

We want a robust query algorithm that almost never hurts, while preserving large average gains



Query expansion:
Current state-of-the-art method



Robust version

Current query expansion algorithms still have basic problems

- They ignore evidence of risky scenarios & data uncertainty
  - e.g. query aspects not balanced in expansion model
  - Result: unstable algorithms with high downside risk
- Existing methods cannot handle increasingly complex estimation problems with multiple task constraints
  - Personalization, computation costs, implicit/explicit feedback…
- We need a better algorithmic framework that..
  - Optimizes for both relevance and variance
  - Solves for the optimal set of terms, not just individual selection
  - Makes it easy to account for multiple sources of domain knowledge
  - Restricts or avoids expansion in risky situations
- Is there a generic method that can be applied to improve the output from existing algorithms?

Example: Ignoring aspect balance increases algorithm risk

Hypothetical query: 'merit pay law for teachers'

court	0.026
appeals	0.018
federal	0.012
employees	0.010
case	0.010
education	0.009
school	0.008
union	0.007
seniority	0.007
salary	0.006

legal aspect is modeled…

BUT

education & pay aspects thrown away..

111

A better approach is to optimize selection of terms as a set

Hypothetical query: 'merit pay law for teachers'

court	0.026
appeals	0.018
federal	0.012
employees	0.010
case	0.010
education	0.009
school	0.008
union	0.007
seniority	0.007
salary	0.006

More balanced query model

Empirical evidence: Udupa, Bhole and Bhattacharya. ICTIR 2009

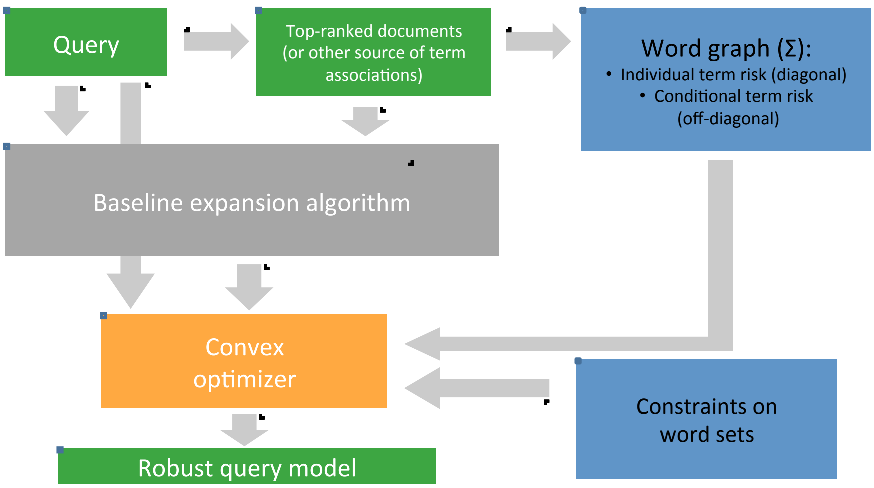
112

A portfolio theory approach to query expansion

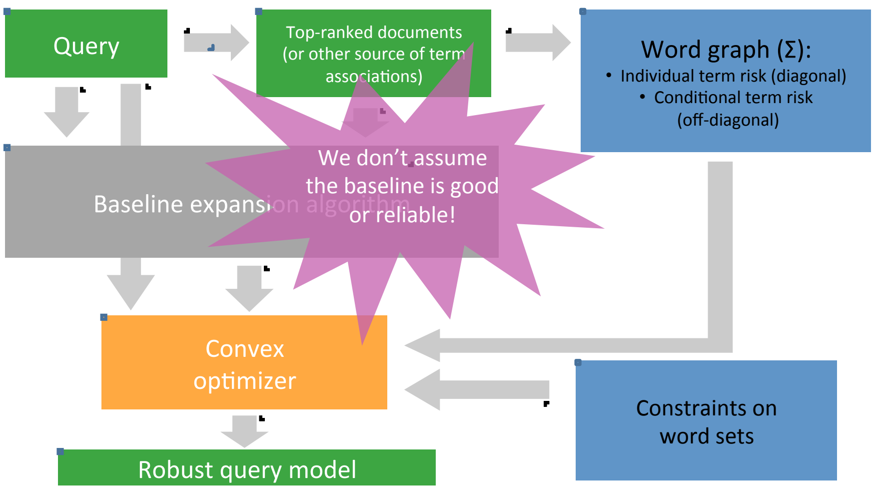
[Collins-Thompson NIPS 2008, CIKM 2009]

1. Cast query expansion as a constrained convex optimization problem:
   - Risk and reward captured in objective function
   - Allows rich constraint set to capture domain knowledge
2. Robust optimization gives more conservative solutions by accounting for uncertainty:
   - Define uncertainty set U around data (term weights)
   - Then minimize worst-case loss over U
   - Simple QP regularization form

Our approach refines initial estimates from a baseline expansion algorithm



Our approach refines initial estimates from a baseline expansion algorithm



115

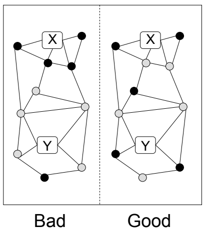
Portfolio theory suggests a good objective function for query expansion

- Reward:
  - Baseline provides initial weight vector *c*
  - Prefer words with higher c_i values: $R(x) = c^T x$
- Risk:
  - Model uncertainty in *c* using a covariance matrix Σ
  - Model uncertainty in Σ using regularized $\Sigma_\gamma = \Sigma + \gamma D$
  - Diagonal: captures individual term variance (centrality)
  - Off-diagonal: term covariance (co-occurrence)
- Combined objective:

$$U(x) = -R(x) + \kappa V(x) = -c^T x + \frac{\kappa}{2} x^T (\Sigma + \gamma D) x$$

116

What are good constraints for query expansion? Visualization on a word graph:

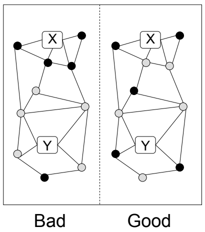


Aspect balance

- Vertices: Words
  Query terms X and Y
- Edges: word similarity, e.g. term association or co-occurrence measure

Query term support: the expanded query should not be too ‘far’ from the initial query. The initial query terms should have high weight in the expanded query.

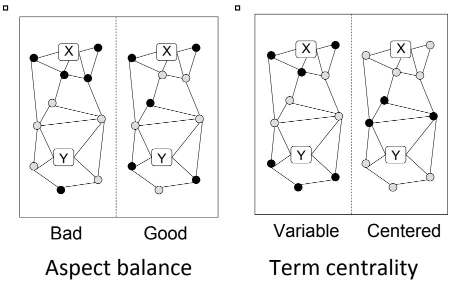
Aspect balance means that both concepts X and Y are well-represented



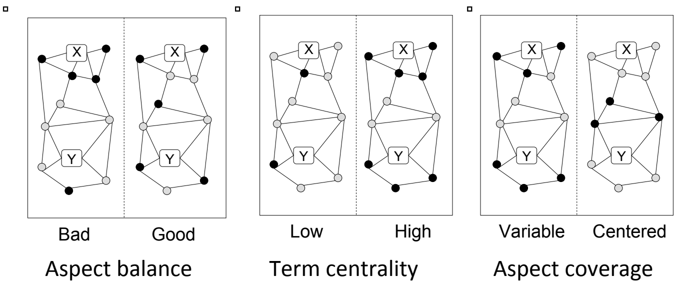
Aspect balance

- Vertices: Words
  Query terms X and Y
- Edges: word similarity, e.g. term association or co-occurrence measure

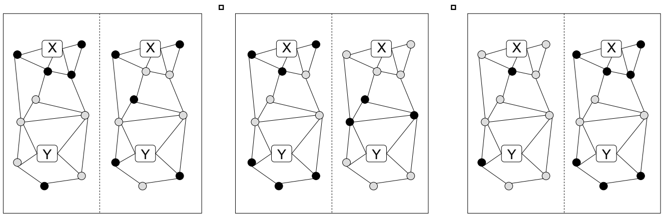
Term centrality prefers words related to multiple query terms



Aspect coverage controls the level of support for each query concept



These conditions are complementary and can be combined with the objective into quadratic program



Bad	Good	Variable	Centered	Low	High
Aspect balance		Term centrality		Aspect coverage	

REXP algorithm

$$
\begin{array}{lll}
\text{minimize} & -c^T x + \frac{\kappa}{2} x^T \Sigma_\gamma x + \lambda y & \textit{Risk \& reward; Budget} \\
\text{subject to} & Ax \le \mu + \varsigma_\mu & \textit{Aspect balance} \\
 & g_i^T x \ge \varsigma_i, \quad w_i \in Q & \textit{Aspect coverage} \\
 & l_i \le x_i \le u_i, \quad w_i \in Q & \textit{Query term support} \\
 & w^T x \le y & \textit{Budget / sparsity} \\
 & 0 \le x \le 1 &
\end{array}
$$

121

Example solution output

Query: `parkinson's disease`

<u>Baseline expansion</u>

```
parkinson   0.996
disease  0.848
syndrome 0.495
disorders   0.492
parkinsons  0.491
patient  0.483
brain    0.360
patients 0.313
treatment   0.289
diseases 0.153
alzheimers  0.114
...and 90 more...
```

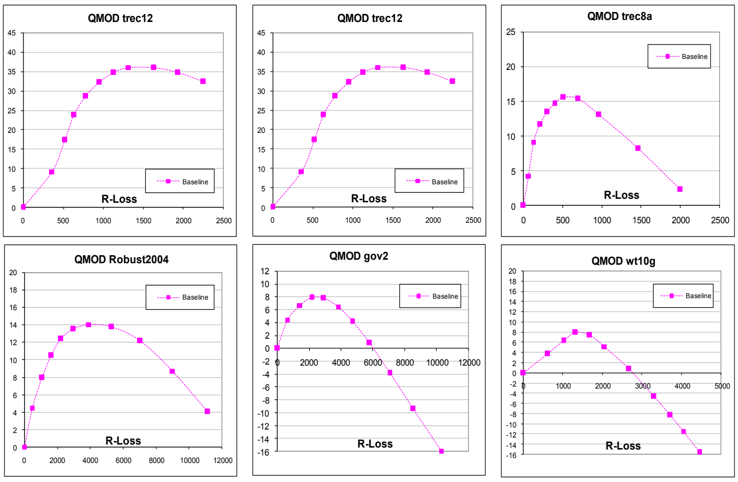
<u>Convex REXP expansion</u>

```
parkinson   0.9900
disease  0.9900
syndrome 0.2077
parkinsons  0.1350
patients 0.0918
brain    0.0256
```

(All other terms zero)

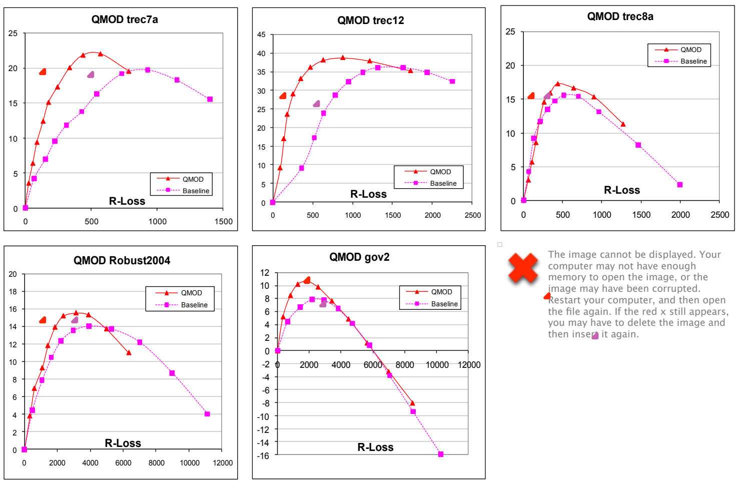
122

Convex REXP version dominates the strong baseline version (MAP)



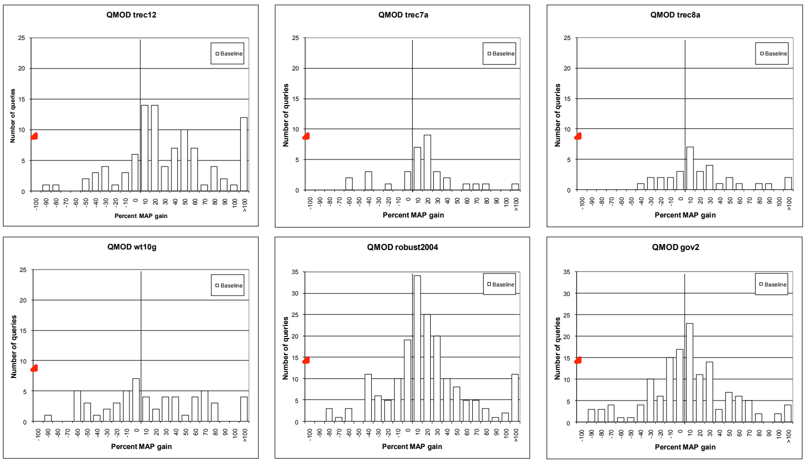
123

Convex REXP version dominates the strong baseline version (MAP)



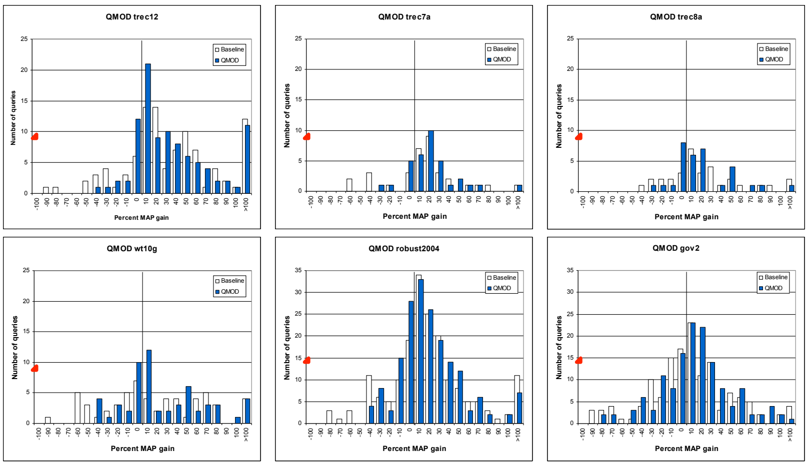
124

REXP significantly reduces the worst expansion failures



125

REXP significantly reduces the worst expansion failures



126

Summary: Avoiding risk in query expansion

- Formulate as an optimization problem that selects the best set of terms, with some constraints.
  - *Portfolio theory* provides effective framework

- Both the objective and constraints play a critical role in achieving more reliable overall performance:
  - Objective:
    - Select the best overall set
    - Penalize solutions in directions of high uncertainty
  - Constraints: Prune likely bad solutions completely

From query expansion to automatic query rewriting ('alteration')

- It can be hard for a user to formulate a good query
  - Misspellings: fored → ford
  - Synonyms: pictures → photos
  - Over-specifying: directions for irs tax form 1040ez → 1040ez directions
  - Under-specifying: sis file →mobile sis file
  - etc.

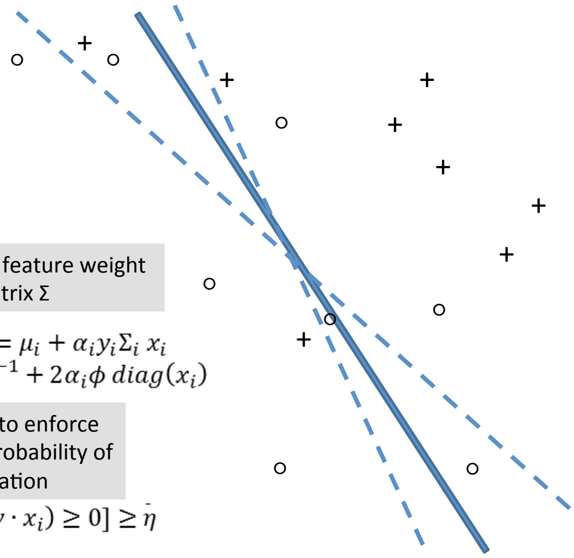
- We want to modify user's query automatically to improve their search results
  - Oracle: best single-word alteration gives gains
  - Synonyms affect 70 percent of user searches across 100 languages [Google study]

Robust classification: Query re-writing as binary classification under uncertainty

129

Confidence-weighted classifiers treat decision boundary as a random variable

[Dredze, Crammer, Pereira ICML 2008]



1. Estimate feature weight variance matrix Σ

$$\mu_{i+1} = \mu_i + \alpha_i y_i \Sigma_i \, x_i$$
$$\Sigma_{i+1}^{-1} = \Sigma_i^{-1} + 2\alpha_i \phi \, diag(x_i)$$

2. Attempt to enforce bound on probability of mis-classification

$$\Pr[y_i(w \cdot x_i) \geq 0] \geq \bar{\eta}$$

130

CW classifiers: AROW algorithm

Adaptive Regularization Of Weights [Crammer, Kulesza, Dredze NIPS 2009]

- On-line algorithm
- Large margin training
- Confidence weighting
- Handles non-separable data

Input: r

For i=1:m

1. Receive example $\boldsymbol{x}_i$ and label y_i.
2. If $y_i, \boldsymbol{\mu}^T \boldsymbol{x}_i < 1$ then make the following updates:

$$\boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_i + \alpha_i \boldsymbol{\Sigma}_{i-1} y_i \boldsymbol{x}_i$$

$$\boldsymbol{\Sigma}_{i+1} = \boldsymbol{\Sigma}_i - \beta_i \boldsymbol{\Sigma}_{i-1} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\Sigma}_{i-1}$$

where

$$\alpha_i = \ell_h(y_i, \boldsymbol{\mu}_{i-1}^T \boldsymbol{x}_i)\beta_i$$

$$\beta_i = \frac{1}{\boldsymbol{x}_i^T \boldsymbol{\Sigma}_{i-1} \boldsymbol{x}_i + r}$$

Output: $\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_\mathrm{m}$

Examples of high- and low-risk query rewriting rules learned with AROW

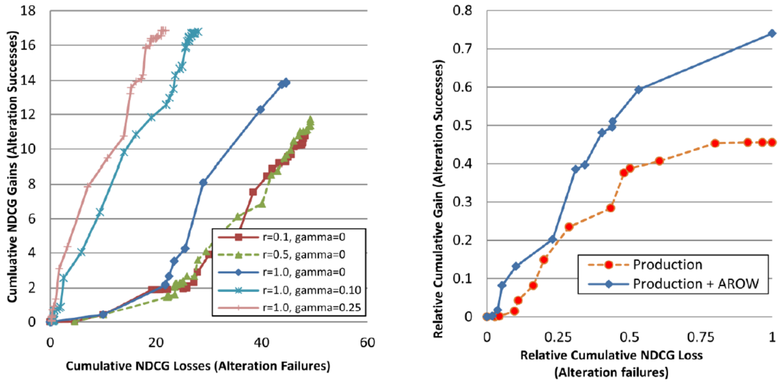
Alteration rule	*Mean*	*Variance*
["welcome] §→lyrics	0.737	0.153
["my] §→lyrics	0.672	0.139
["wish] §→lyrics	0.669	0.184
[18] §→lyrics	0.631	0.119
["wasting] §→lyrics	0.635	0.190
[cent] §→lyrics	0.569	0.135
[move] §→lyrics	0.551	0.207
[broadway] §→lyrics	-0.486	0.222
[dance] §→lyrics	-0.503	0.186
[matter"] §→lyrics	-0.521	0.197
[morning_] §→lyrics	-0.570	0.181
[1_] §→lyrics	-0.626	0.134
[killer] §→lyrics	-0.631	0.190

Adding “lyrics” to a query

Alteration rule	*Mean*	*Variance*
!970 → 1970	0.31	0.03
<1>,usa[ace → myspace	0.52	0.02
[arctic_]act → cat	0.24	0.05
[airbrush]pain → paint	0.13	0.03
[adobe]win → windows	0.14	0.05
[405]win → winchester	0.23	0.01
[7]win → +win	-0.22	0.02
[andersen]win → windows	0.10	0.03
[calculator]fiance → finance	0.09	0.05
[alicia]fiance → boyfriend	0.16	0.04
[adriana]fiance → finance	-0.11	0.05

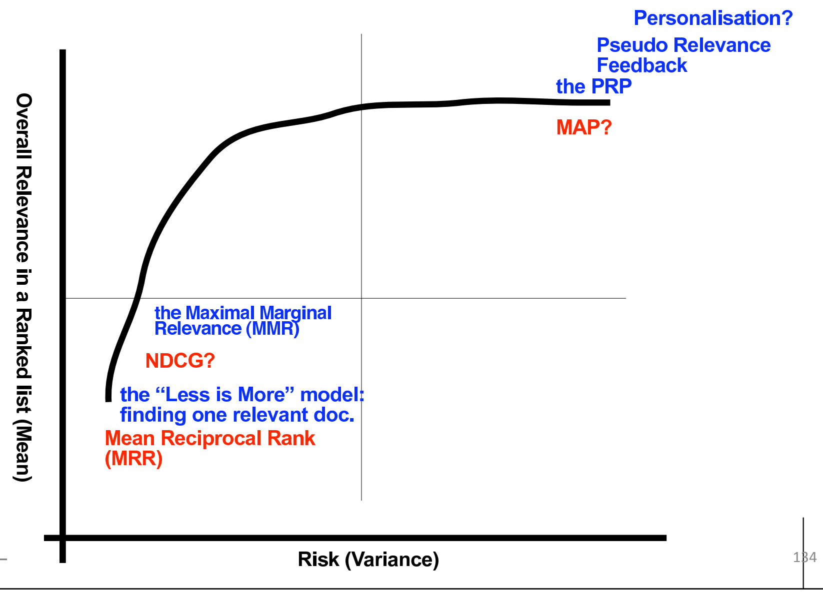
Detecting mis-typed words

Reducing the risk of automatic query reformulation with AROW alter/no-alter features



[Collins-Thompson, Lao, Ali, Teevan. Learning Low-Risk Rules for Altering Queries. In Submission]

Risk vs. reward



Why IR Models → Individualism → Rank Context → Portfolio Retrieval

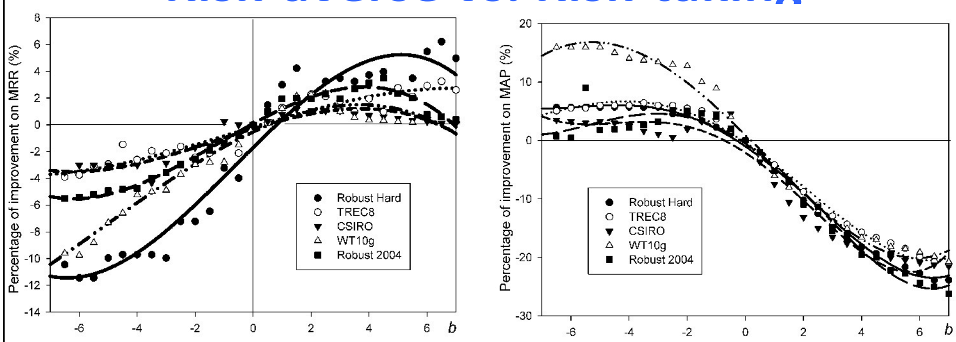
The advantages

- Theoretically explained the need for diversification -> reduce the risk/uncertainty
- Explanations of some empirical retrieval results
  - the trade-off between MAP and MRR, and
  - the justification for pseudo-relevance feedback
  - but also help us develop useful retrieval techniques such as risk-aware query expansion and optimal document ranking.

135

Why IR Models → Individualism → Rank Context → Portfolio Retrieval

Risk-averse vs. Risk-taking



(a) Mean Reciprocal Rank (MRR) (b) Mean Average Precision (MAP)

(a) positive b (minus variance): "invest" into different docs. increases the chance of early returning the first rel. docs -> **Risk-averse**

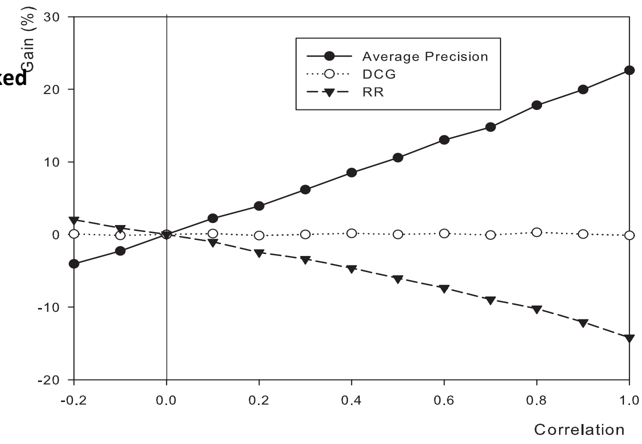
(b) negative b (add variance): "invest" in "similar" docs (big variance) might hurt the MRR but on average increases the performance of the entire ranked list -> **Risk-taking**

J. Wang and J. Zhu, "Portfolio Theory of Information Retrieval," in SIGIR, 2009.

136

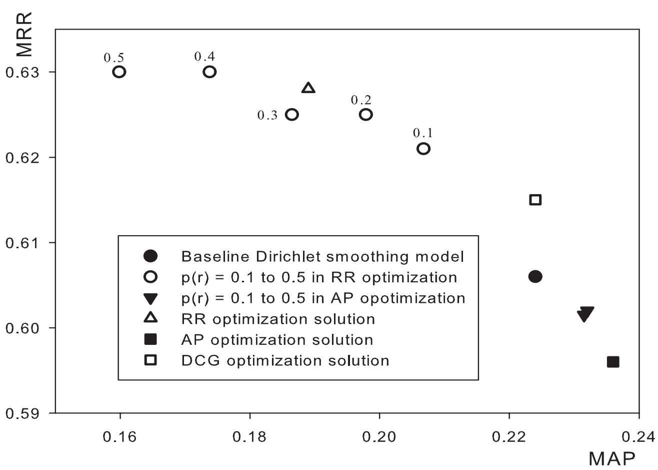
Understanding IR metrics under uncertainty

Simulation: change correlations in the ranked list.
Neg correlation -> increase RR
Positive Correlation -> increase Average Precision



[Wang and Zhu 2010] J. Wang and J. Zhu, "On Statistical Analysis and Optimization of Information Retrieval Effectiveness Metrics," in Proc. of SIGIR 2010.

No free lunch

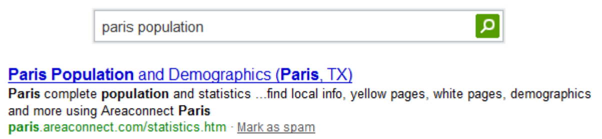


[Wang and Zhu 2010] J. Wang and J. Zhu, "On Statistical Analysis and Optimization of Information Retrieval Effectiveness Metrics," in Proc. of SIGIR 2010.

Table of Contents

Risking brand: Exploration vs. exploitation

- Should you display potentially irrelevant items to determine if they are relevant?



- Showing irrelevant items risks lowering user perception of search engine's quality.
- Potentially more susceptible to spamming
- Open Area:
  - Models that learn risk and reward and integrate that into a risk/reward tradeoff framework.
  - Identifying low risk-scenarios for exploring relevance.
  - Predicting query difficulty

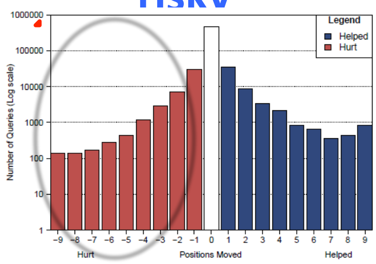
Choosing when and how to personalize search results

- The same query means different things to different people.
- The same results therefore have different relevance value to two issuers of the same query.

- Hypothesis: many forms of ambiguity would disappear if we could condition on the user.

141

State-of-the-art personalization is still risky



Reading level personalization**:** re-ranking gains and losses
(**Note log scale.**)
[Collins-Thompson et al. CIKM 2011]

- Similar distributions for personalization by:
  - Location [Bennett et al., SIGIR 2011]
  - Long-term user profiles [In submission]

142

The risk of personalization

- Personalization can help significantly, but when and how to apply?
  - All the time?
    - Data sparsity challenge: building a profile to cover all queries.
    - Often people search "outside" of their profiles.
  - When the query matches the user's profile?
    - How should the profile be built? Topically? Demographic? Locale?
- Predicting when to personalize is likely to have a high payoff if done with a high accuracy.
  - Early results indicate reasonable accuracy can be attained via machine learning [Teevan *et al.*, SIGIR 2008].
- Open area for machine learning researchers to contribute more methods and approaches.

143

Federated search optimization

- Search results can come from different resources, which are then combined
  - Reward: Relevance estimates for individual resources
  - Risk: estimates of resource overlap
  - Budget and other constraints
- Web search is becoming federated search
  - Instant answer
  - Vertical search (topic experts: sports, health, …)
  - Fact databases (people, places, …)

144

On-going research directions

- Multimedia retrieval
  - Aly, R.B.N., Aiden, D., Hiemstra, D., Smeaton, A. (2010) Beyond Shot Retrieval: Searching for Broadcast News Items Using Language Models of Concepts. In ECIR 2010.
- Content analysis and fusion
  - Xiangyu Wang, Mohan Kankanhalli: Portfolio theory of multimedia fusion. ACM Multimedia 2010: 723-726.
- Advertising
  - D. Zhang, J. Lu. Batch-mode Computational Advertising based on Modern Portfolio Theory. ICTIR 2009.
- Collaborative Filtering
  - J. Wang, "Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval," in Proc. of ECIR, 2009.

145

Future directions

- Broad applicability for robust risk frameworks to improve reliability and precision in IR
  - More stable, reliable solutions based on accounting for variance and uncertainty
  - Query reformulation, when to personalize, federated resources, document ranking…
- Learn effective parameters for objectives, feasible sets for selective operation
- New objectives, constraints, approximations, computational tradeoffs for scalability
- Structured prediction problems in high dimensions with large number of constraints

146

Thank you!

Jun Wang

jun.wang@cs.ucl.ac.uk
http://www.cs.ucl.ac.uk/staff/jun.wang/blog/

Kevyn Collins-Thompson

kevynct@microsoft.com
http://research.microsoft.com/~kevynct

147

Acknowledgments

- Thanks to Jianhan Zhu, Paul Bennett, Misha Bilenko and Filip Radlinski for contributions to this presentation.

148

Bibliography

- History of relevance and IR models
  - M.E. Maron and J. L. Kuhns. (1960) On relevance, probabilistic indexing, and information retrieval. Journal of the ACM 7:216-244.
  - Mizarro, S. Relevance: The whole history. *Journal of the American Society for Information Science 48*, 9 (1997), 810.832.
  - William S. Cooper. The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. *University of California, Berkeley*, 1971.
- Classical probabilistic IR model and extensions
  - S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, pages 294-304, 1977.
  - S.E. Robertson and K. Spärck Jones, Relevance weighting of search terms. *Journal of the American Society for Information Science* **27**, 129-46 (1976). http://www.soi.city.ac.uk/~ser/papers/RSJ76.pdf
  - S.E. Robertson. (1990) On term selection for query expansion. Journal of Documentation. 46, 359-364. http://www.soi.city.ac.uk/~ser/papers/on_term_selection.pdf
  - Robertson, S. E. and Walker, S. (1999). Okapi/keenbow at TREC-8. In Voorhees, E. M. and Harman, D. K., editors, *The Eighth Text REtrieval Conference (TREC 8)*. NIST Special Publication 500-246.
  - Michael D. Gordon and Peter Lenk. A utility theoretic examination of the probability ranking principle in information retrieval. *JASIS*, 42(10):703-714, 1991.
  - 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.
  - [Järvelin and Kekäläinen(2002)] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 2002.

Bibliography (cont'd)

- Portfolio Theory and IR applications
  - H Markowitz. Portfolio selection. *Journal of Finance*, 1952
  - [Collins-Thompson 2008] K. Collins-Thompson. "Robust model estimation methods for information retrieval". Ph.D. thesis (LTI Technical Report CMU-LTI-08-010) Carnegie Mellon University, 2008.
  - [Collins-Thompson 2008] K. Collins-Thompson. "Estimating robust query models with convex optimization". *Advances in Neural Information Processing Systems 21* (NIPS), 2008. pg. 329-336
  - [Collins-Thompson 2009] K. Collins-Thompson. "Reducing the risk of query expansion via robust constrained optimization". *Proceedings of the Eighteenth International Conference on Information and Knowledge Management (CIKM 2009)*. ACM. Hong Kong. pg. 837-846.
  - [Collins-Thompson 2009] K. Collins-Thompson. "Accounting for stability of retrieval algorithms using risk-reward curves". *Proceedings of SIGIR 2009 Workshop on the Future of Evaluation in Information Retrieval*, Boston. pg. 27-28.
  - [Collins-Thompson and Dillon 2010] K. Collins-Thompson and J. Dillon. "Controlling the search for expanded query representations by constrained optimization in latent variable space." *SIGIR 2010 Workshop on Query Representation and Understanding.*
  - [Dillon and Collins-Thompson 2010] J. Dillon and K. Collins-Thompson. "A unified optimization framework for robust pseudo-relevance feedback algorithms." *Proceedings of the Nineteenth International Conference on Information and Knowledge Management (CIKM 2010)*, Toronto, Canada.
  - [Wang 2009] J. Wang, ""Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval," in Proc. of ECIR, 2009.
  - [Zhu et al 2009] Zhu, J. Wang, M. Taylor, and I. Cox, "Risky Business: Modeling and Exploiting Uncertainty in Information Retrieval," in Proc. Of SIGIR, 2009.
  - [Wang and Zhu 2009] J. Wang and J. Zhu, "Portfolio Theory of Information Retrieval," in SIGIR, 2009.
  - [Wang and Zhu 2010] J. Wang and J. Zhu, "On Statistical Analysis and Optimization of Information Retrieval Effectiveness Metrics," in Proc. of SIGIR 2010.
  - [Zuccon, Azzopardi, van Rijsbergen SIGIR 2010]

Bibliography (cont'd)

- Language modeling for IR
  - J.M. Ponte and W.B. Croft. 1998. A language modeling approach to information retrieval. In *SIGIR 21.* pp. 275-281.
  - A. Berger and J. Lafferty. 1999. Information retrieval as statistical translation. *SIGIR 22*, pp. 222–229.
  - Workshop on Language Modeling and Information Retrieval, CMU 2001. http://sigir.org/forum/S2001/LM.pdf
  - The Lemur Toolkit for Language Modeling and Information Retrieval. Open-source toolkit from CMU/Umass. LM and IR system in C(++) http://www.lemurproject.org/~lemur/
  - C. Zhai. 2008. Statistical language models for information retrieval: a critical review. Foundations and Trends in Information Retrieval Vol. 2, No. 3.
  - V. Lavrenko. A Generative Theory of Relevance. Doctoral dissertation. Univ. of Massachusetts Amherst, 2004.
  - Metzler, D. and Croft, W.B., "A Markov Random Field Model for Term Dependencies," *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR 2005), 472-479, 2005

Bibliography (cont'd)

- Federated search / distributed IR / meta-search
  - Callan, J. (2000). Distributed information retrieval. In W.B. Croft, editor, *Advances in Information Retrieval*. (pp. 127-150). Kluwer Academic Publishers.
  - L. Si. (2006). Federated Search of Text Search Engines in Uncooperative Environments. Doctoral dissertation. Carnegie Mellon University.
  - F. A. D. Neves, E. A. Fox, and X. Yu. Connecting topics in document collections with stepping stones and pathways. In CIKM, pages 91–98, 2005.
  - Aslam, J. A. and Montague, M. 2001. Models for metasearch. In *Proceedings of SIGIR 2001*(New Orleans, Louisiana, United States). SIGIR '01. ACM, New York, NY, 276-284.
- Recommender systems
  - Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the User Experience of Recommender Systems. *Paper in second round of reviews for a special issue of User Modeling and User-Adapted Interaction (UMUAI) on User Interfaces of Recommender Systems.* http://www.usabart.nl/portfolio/KnijnenburgWillemsen-UMUAI2011_UIRecSy.pdf
- Learning-to-rank and rank diversity
  - Yasser Ganjisaffar, Rich Caruana, Cristina Videira Lopes: Bagging gradient-boosted trees for high precision, low variance ranking models. SIGIR 2011: 85-94

Appendix A

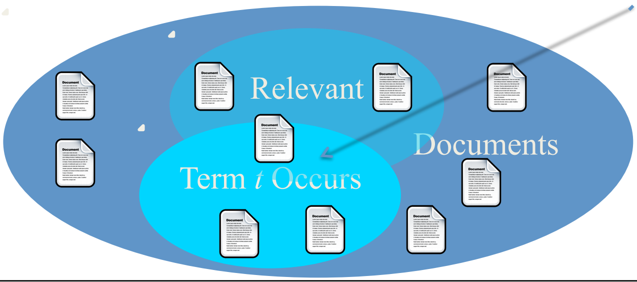
The derivation of RSJ Model

RSJ Model (joint prob.)

Given term *t*, we could have a contingency table to summarize our observation:

	Relevant	Non-relevant	
Term t Occur	r_t	n_t-r_t	n_t
Term t Not Occur	R-r_t	N-R-n_t+r_t	N-n_t
	R	N-R	N

$$P(t=1, r=1 \mid q) = \frac{r_t}{N}$$



The joint probability presents the chance that a document will fall into that region.

RSJ Model (conditional prob.)

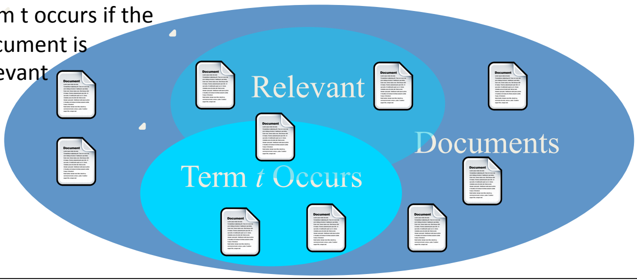
Given term *t*, we could have a contingency table to summarize our observation:

	Relevant	Non-relevant	
Term t Occur	r_t	n_t-r_t	n_t
Term t Not Occur	R-r_t	N-R-n_t+r_t	N-n_t
	R	N-R	N

The cond probability presents the chance that term t occurs if the document is relevant

$$P(t=1|r=1,q)=\frac{P(t=1,r=1|q)}{\sum_t P(t=1,r=1|q)}=\frac{r_t}{R} \quad P(t=1|r=0,q)=\frac{n_r-r_t}{N-R}$$

The cond probability presents the chance that term t occurs if the document is not relevant



RSJ Model (conditional prob.)

Given term *t*, we could have a contingency table to summarize our observation:

	Relevant	Non-relevant	
Term t Occur	r_t	n_t-r_t	n_t
Term t Not Occur	R-r_t	N-R-n_t+r_t	N-n_t
	R	N-R	N

- For simplicity, we define two parameters for *t*

$$a(t)\equiv P(t=1|q,r=1)=\frac{r_t}{R}$$

$$b(t)\equiv P(t=1|q,r=0)=\frac{n_r-r_t}{N-R}$$

- We thus have

$$P(t|q,r=1)=a^t(1-a)^{1-t}, P(t|q,r=0)=b^t(1-b)^{1-t}$$

RSJ Model (scoring)

- Now we could score a document based on its terms (whether occur or not in the document)

$$score(d) = \log \frac{P(r=1|d,q)}{P(r=0|d,q)}$$
$$= \log \frac{P(d|r=1,q)P(r=1|q)}{P(d|r=0,q)P(r=0|q)}$$
$$\propto \log \frac{P(d|r=1,q)}{P(d|r=0,q)}$$

RSJ Model (scoring)

- Binary independent assumption
  - *Binary*: either a term occurs or not in a document (term frequency is not considered)

$d = [t_1, t_2, \ldots,]$,

$t = 1$ means that term t occurs in the document $(t \in d)$

$t = 0$ otherwise $(t \notin d)$

  - *Independent*: terms are *conditionally independent* with each other given relevance/non-relevance

$$P(d|r=1,q) = \prod_t P(t|q,r=1) \quad P(d|r=0,q) = \prod_t P(t|q,r=0)$$

RSJ Model (scoring)

- We thus have $score(d) = \log \frac{P(d \mid r=1, q)}{P(d \mid r=0, q)} = \sum_t \log \frac{P(t \mid q, r=1)}{P(t \mid q, r=0)}$
- Replacing the probabilities with the defined parameters gives

$$score(d) = \sum_t \log \frac{a^t (1-a)^{1-t}}{b^t (1-b)^{1-t}} = \sum_t \log \frac{a^t (1-b)^t (1-a)}{b^t (1-a)^t (1-b)}$$

$$= \sum_t t \log \frac{a(1-b)}{b(1-a)} + \sum_t \log \frac{(1-a)}{(1-b)} \propto \sum_t t \log \frac{a(1-b)}{b(1-a)} = \sum_{t \in d} \log \frac{a(1-b)}{b(1-a)}$$

- Consider only the terms occurring in both doc and query, we get

$$score(d) = \sum_{t \in d \cap q} \log \frac{a(1-b)}{b(1-a)}$$

RSJ Model (Bayes' Rule)

	Relevant	Non-relevant	
Term t Occur	r_t	n_t-r_t	n_t
Term t Not Occur	R-r_t	N-R-n_t+r_t	N-n_t
	R	N-R	N

Finally, we get

$$score(d) = \sum_{t \in d \cap q} \log \frac{a(1-b)}{b(1-a)}$$

$$= \sum_{t \in d \cap q} \log \frac{(r_t + 0.5)(N - R - n_t + r_t + 0.5)}{(R - r_t + 0.5)(n_t - r_t + 0.5)}$$

Appendix B

The derivation of Lafferty and Zhai's Bayesian Decision Theory of IR

A Generative IR model

[John Lafferty, Chengxiang Zhai SIGIR 2001]

A User first selects a query model

Model selection *Query generation*

A query is then generated from that query model

$$\mathcal{U} \longrightarrow \theta_Q \longrightarrow q$$

$$p(\theta_Q \,|\, \mathcal{U}) \qquad p(q \,|\, \theta_Q)$$

A Source/ author first selects a doc model

Model selection *Document generation*

A doc is then generated from that doc model

$$\mathcal{S} \longrightarrow \theta_D \longrightarrow d$$

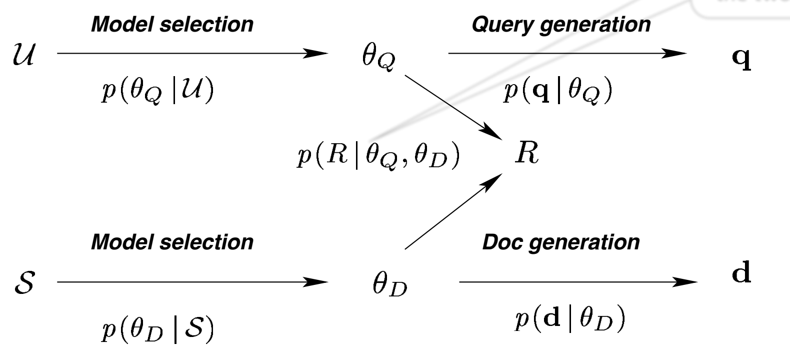
$$p(\theta_D \,|\, \mathcal{S}) \qquad p(d \,|\, \theta_D)$$

John Lafferty, Chengxiang Zhai Document language models, query models, and risk minimization for information retrieval SIGIR 2001

ChengXiang Zhai, John Lafferty A risk minimization framework for information retrieval, IP&M, 2006

A Generative IR model

The Relevance State depends on the two models



John Lafferty, Chengxiang Zhai Document language models, query models, and risk minimization for information retrieval SIGIR 2001

ChengXiang Zhai, John Lafferty A risk minimization framework for information retrieval, IP&M, 2006

Bayesian Decision Theory in LM

- Thus, the expected loss of taking action *a=1*:

$$E[Loss(a=1|q,d)] = \sum_r \int_{\theta_q,\theta_d} d\theta_d \, d\theta_a Loss(a=1|r,\theta_q,\theta_d) p(r,\theta_q,\theta_d|q,d)$$

$$= \sum_r Loss(a=1|r,\theta_q,\theta_d) \int_{\theta_d,\theta_d} p(r|\theta_q,\theta_d) p(\theta_q|q) p(\theta_d|d) d\theta_d \, d\theta_a$$

$$\approx \sum_r Loss(a=1|r,\hat\theta_q,\hat\theta_d) p(r|\hat\theta_q,\hat\theta_d) \quad \text{<- point estimation } p(\hat\theta_q|q) \text{ and } p(\hat\theta_d|d) \approx 1$$

- If a distance-based loss function is used

$$Loss(a=1|r,\hat\theta_q,\hat\theta_d) \equiv KL(\hat\theta_q,\hat\theta_d) \equiv \sum_t p(t|\hat\theta_q) \log \frac{p(t|\hat\theta_q)}{p(t|\hat\theta_d)}$$

the Kullback–Leibler divergence is a non-symmetric measure of the difference between two probability distributions

- This results in:

$$E[Loss(a=1|q,d)] \approx KL(\hat\theta_q,\hat\theta_d) \sum_r p(r|\hat\theta_q,\hat\theta_d) = KL(\hat\theta_q,\hat\theta_d)$$

Bayesian Decision Theory in LM

- A further development can show that

$$E[Loss(a=1\,|\,q,d)] \approx KL(\hat{\theta}_q,\hat{\theta}_d)$$

$$=\sum_t p(t\,|\,\hat{\theta}_q)\log\frac{p(t\,|\,\hat{\theta}_q)}{p(t\,|\,\hat{\theta}_d)}$$

$$\propto -\sum_t p(t\,|\,\hat{\theta}_q)\log p(t\,|\,\hat{\theta}_d)+\sum_t p(t\,|\,\hat{\theta}_q)\log p(t\,|\,\hat{\theta}_q)$$

$$\propto -\frac{1}{l_q}\sum_{t\in q}\log p(t\,|\,\hat{\theta}_d),\ \text{where } \mathrm{l}_q \text{ is query length}$$

and the empirical distribution is used for $\hat{\theta}_q$

- It is indeed the language model of IR

John Lafferty, Chengxiang Zhai Document language models, query models, and risk minimization for information retrieval SIGIR 2001