Tutorial - AM4

Statistical Information Retrieval Modelling.

Jun Wang, Kevyn Collins-Thompson

Crowne Plaza Hotel
Glasgow, Scotland
24-28 October 2011
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

- 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

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
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

The history of Probabilistic Retrieval Models

- Probabilistic models
  - Probabilistic indexing
  - Robertson/Spärck Jones Rel Model
  - Two-Poisson model → BM25
  - Bayesian inference networks
  - Statistical language models
- Citation analysis models
  - Hubs & authorities
  - PageRank
  - Clever (1998)
  - Google
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)

Ambiguous queries

- For a given query, we might have different information needs from different users

- By issuing query “Egypt”,
  - A set of ranked Docs
  - 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
  - $k = 1$
  - $k = K_j^v$
**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?**

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

![Diagram showing relevance and assessors](image)

**Ambiguity by Result Type (documents)**

- An overview?
- Tech. Papers?
- Books?
- Software?
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

Risk: statistics of undesirable outcomes

\[ R(x) = \int_{\Theta} L(x, \theta) p(\theta | x) d\theta \]
\[ \sigma^* = \arg \min_{\sigma} R(x | \sigma) \]
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 [This tutorial]
  – Query processing
  – Ranking reliability and diversity

Risk, bias and variance in machine learning

• Across different possible training sets of given size:
  • Bias: how well average prediction of the learning algorithm matches optimal prediction (Bayes rate)
  • Variance: how much the algorithm’s prediction fluctuates
  • Squared error 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 risk 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

The risk of making (multiple) errors in Web search

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

---

**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
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.

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:

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Non-relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term ( t ) Occur</td>
<td>( r_t )</td>
<td>( n_t-r_t )</td>
</tr>
<tr>
<td>Term ( t ) Not Occur</td>
<td>( R-r_t )</td>
<td>( N-R-n_t+r_t )</td>
</tr>
<tr>
<td>Total</td>
<td>( R )</td>
<td>( N-R )</td>
</tr>
</tbody>
</table>

The logit of the probability of relevance:

\[
\text{score}(d) = \log \frac{P(r = 1|q,d)}{P(r = 0|q,d)}
\]

In the end, we get

\[
\text{score}(d) = \sum_{r_t \neq q} \log \left( \frac{(r_t + 0.5)(N - R - n_t + r_t + 0.5)}{(R - r_t + 0.5)(n_t - r_t + 0.5)} \right) \]

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
  \[ R = r_i = 0 \] gives
  \[ score(d) = \sum_{t_i \in q} \log \frac{N - n_t + 0.5}{n_t + 0.5} \]
- As \( n_t \) is much smaller than \( N \), the above is equivalent to IDF

BM25: dealing with Term Frequency

- RSJ model does not consider term frequency
- Saturation Function of Term Frequency
  \[ s(tf) = \frac{s_{MAX} \cdot tf}{tf + K} \]
  \( s_{MAX} \): max score, \( K \) controls the slop
- The BM (Best Math)25 formula
  \[ score(d) = \sum_{t_i \in q} \frac{tf}{tf + K} \log \frac{N - n_t + 0.5}{n_t + 0.5} \]
  \( K = k_1((1 - \lambda) + \lambda L_d) \), \( L_d \) is the normalized doc length
  (i.e. the length of this doc 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

Language Model

Language Model: construct a document model and see how likely a query can be generated from it
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: \(tf_t\)

Terms

Term \(t\) occurs

Language Models (para. esti.)

- Linear Smoothing LM models

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

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

Give the document \(d\):

Observations:
Doc Length: \(L_q\); Num. Occurrences: \(tf_t\)

Terms

Term \(t\) occurs

Give the collection \(c\):

Observations:
Doc Length: \(L_c\); Num. Occurrences: \(tf_{tc}\)
Language Models (para. esti.)

- Linear Smoothing LM models
  \[
  \text{score}_q(d) = \log p(q \mid d) = \log \prod_{t \in q} p(t \mid d) = \sum_{t \in q} \frac{tf_t}{L_d}
  \]

  Give the document
  Term \( t \) occurs
  Terms
  Observations:
  Doc Length: \( L_d \); Num. Occurrences: \( tf_t \)

  Give the collection
  Term \( t \) occurs
  Terms
  Observations:
  Doc Length: \( L_c \); Num. Occurrences: \( tf_{t,c} \)

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
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.*


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: \( P = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Retrieved}|} = \frac{n_j}{n} = \frac{\sum r_j}{n} \)
- Expected Precision@n:
  \[
  E[P] = \frac{\sum [r_j] \cdot \sum p(r_j = 1)}{n} = \frac{\sum [r_j]}{n} \cdot \frac{n}{n} = \frac{p(r_j = 1)}{n}, \quad \text{where } p(r_j = 1) \text{ Prob of rel at rank } j
  \]
  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 \( p(r_j = 1) \)

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:
  \[
  E[L] = \sum_j ((j-1)p(r_j = 1, r_i = 0, ..., r_{j-1} = 0))
  \]
  \[
  = \sum_j ((j-1)p(r_j = 1) \prod_{i=1}^{j-1} p(r_i = 0)) \quad \Leftrightarrow \text{independent assumption}
  \]
  \[
  = 0p(r_1 = 1) + 1p(r_2 = 1)p(r_1 = 0)...
  \]
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:
  \[
  E[L] = \sum_j ((j-1)p(r_j = 1, r_l = 0, ..., r_{j-1} = 0)) \\
  = \sum_j ((j-1)p(r_j = 1) \prod_{i=1}^{j-1} p(r_i = 0)) \quad \leq \text{independent assumption}
  \]

  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:

<table>
<thead>
<tr>
<th>UserClass</th>
<th>D1:Egypt_politics</th>
<th>D2:Egypt_politics</th>
<th>D3:Egypt_travel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egypt_politics</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Egypt_travel</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(p(r))</td>
<td>2/3</td>
<td>2/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>
But why if we are not sure about Information Need (difficulty 1)

<table>
<thead>
<tr>
<th>UserClass</th>
<th>D1:Egypt_politics</th>
<th>D2:Egypt_politics</th>
<th>D3:Egypt_travel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egypt_politics</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Egypt_travel</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>P(r)</td>
<td>2/3</td>
<td>2/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

• PRP gives
  - d1 Egypt_politics
  - d2 Egypt_politics or
d2 Egypt_politics
  - 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
  \[
  E[L] = 0 \cdot p(r(d_i) = 1) + 1 \cdot p(r(d_i) = 1) \cdot p(r(d_i) = 0) \\
  + 2 \cdot p(r(d_i) = 1) \cdot p(r(d_i) = 0) \cdot p(r(d_i) = 0) \\
  = 0(2/3)+1(1/3)(1/3)+2(1/3)(1/3)(1/3)=8/27
  \]

• {d1,d3,d2}
  - d1 Egypt_politics
  - d3 Egypt_travel
  - d2 Egypt_politics
  \[
  E[L] = 0 \cdot p(r(d_i) = 1) + 1 \cdot p(r(d_i) = 1) \cdot p(r(d_i) = 0) \\
  + 2 \cdot p(r(d_i) = 1) \cdot p(r(d_i) = 0) \cdot p(r(d_i) = 0) \\
  = 0(2/3)+1(1/3)(1/3)+2(2/3)(2/3)(1/3)=11/27
  \]

• {d1,d2,d3} is better
Expected Search Length: consider the dependency

- \{d_1, d_2, d_3\}:
  \[ E[L] = 0 \cdot p(r(d_1) = 1) + 1 \cdot p(r(d_2) = 1, r(d_1) = 0) + 2 \cdot p(r(d_3) = 1, r(d_2) = 0, r(d_1) = 0) \]
  \[ = 0 \cdot \frac{2}{3} + 1 \cdot 0 + 2 \cdot \frac{1}{3} = \frac{2}{3} \]

- \{d_1, d_3, d_2\}:
  \[ E[L] = 0 \cdot p(r(d_1) = 1) + 1 \cdot p(r(d_2) = 1, r(d_1) = 0) + 2 \cdot p(r(d_3) = 1, r(d_2) = 0, r(d_1) = 0) \]
  \[ = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} + 2 \cdot 0 = \frac{1}{3} \]

- \{d_1, d_3, d_1\} 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

• 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

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

\[
\lambda Sim_1(d_j,q) - (1 - \lambda) \max_{d_i \in \{1, j-1\}} Sim_2(d_i, d_j),
\]

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

- 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

\[
p(r_j = 1 | r_{j-1} = 0), \text{ where } j \text{ is the rank}
\]

- Metric: \( k\text{-call} @ N \)
  - 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

\[ p(r_i = 1 | r_{i-1} = 0, ..., r_1 = 0) \]

• 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 \( E[\text{metric} | d_1, d_n] \) 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

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
Learning to Rank: Drawbacks

- 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 \textit{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

- 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
- \textbf{Ranking in context and diversity}
  - Loss functions for diverse ranking
  - Less is More, Maximum Marginal Relevance, Diversity Optimization
  - \textbf{Bayesian Decision Theory}
- Portfolio Retrieval
  - Document ranking
  - Risk-reward evaluation methods
  - Query expansion and re-writing
- Future challenges and opportunities
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:
  \[
  Loss(a = 1|r) = \begin{cases} 
  c_1 & r = 1 \\
  c_2 & r = 0 
  \end{cases}, \quad c_2 > c_1
  \]
### Bayesian Decision Theory in LM

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

$$E[\text{Loss}(a=1 \mid q,d)] = \sum_r \text{Loss}(a=1 \mid r)p(r \mid q,d)$$

$$= \text{Loss}(a=1 \mid r = 1)p(r = 1 \mid q,d) + \text{Loss}(a=1 \mid r = 0)(1 - p(r = 1 \mid p,q))$$

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

- Minimizing it would pick up the document which has the highest probability of relevance $p(r=1 \mid q,d)$

- Thus rank in ascending order of expected loss is equivalent to that in descending order of prob. of relevance

---

### A Generative IR model

- A User first selects a query model

- A Source/author first selects a doc model

- A query is then generated from that query model

- A doc is then generated from that doc model

---

**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

\[ U \xrightarrow{\text{Model selection}} p(\theta_Q | U) \xrightarrow{\text{Query generation}} \theta_Q \xrightarrow{p(q | \theta_Q)} q \]

\[ S \xrightarrow{\text{Model selection}} p(\theta_D | S) \xrightarrow{\text{Doc generation}} \theta_D \xrightarrow{p(d | \theta_D)} d \]

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

Understanding Lafferty and Zhai’s model

- A general and principled IR model
- A point estimation is used in the formulation.
  \[ \int_{\hat{\theta}_a, \hat{\theta}_b} p(r | \hat{\theta}_a, \hat{\theta}_b) p(\theta_a | q) p(\theta_b | d) d\theta_a d\theta_b = p(r | \hat{\theta}_a, \hat{\theta}_b) \]
  - 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 \( p(r | \hat{\theta}_a, \hat{\theta}_b) \) (Victor Lavrenko and W. Bruce Croft, Relevance-Based Language Models, SIGIR 2001)
**Another formulation using Bayesian Decision theory**  
[Wang and Zhu SIGIR2010]

- 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

---

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

- Predicting Relevance
- Rank Optimization

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

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

- 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
The statistical document ranking process

The joint probability of relevance given a query:
\[ p(r_1, \ldots, r_N | q) \]

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

\[ \hat{a} = \arg \max_a E(m | q) \]
\[ = \arg \max_{a_1, \ldots, a_N} \left( \sum_{r_1, \ldots, r_N} m(a_1, \ldots, a_N | r_1, \ldots, r_N) p(r_1, \ldots, r_N | q) \right) \]

IR metric:
- Input:
  - A rank order \(a_1, \ldots, a_N\)
  - Relevance of docs. \(r_1, \ldots, r_N\)
- The joint probability of relevance given a query

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

**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

---

**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:

  ![Investment Budget](image1)

  ![Portfolio Effect](image2)

- 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)
Web search (rank positions as investment opportunities)

- Maximize profit
- Maximize users' satisfactions?

Portfolio Theory of Information Retrieval

Portion of the Investment Budget

Invest

Google

Yahoo

Coca-cola

Documents /Items

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.
**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

\[
\{ \hat{w}_j \} = \arg\max_{(w_j)} \quad o_n = \sum_{j=1}^{n} w_j r_j \
\]

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)]
  \[ w_j = \frac{1}{2^{j-1}} \text{ where } j \in \{1,...,n\} \]
- It can be easily shown that when $w_1 > w_2 > ... > w_n$, the maximum value of $R_n$ gives the ranking order $r_1 > r_2 > ... > r_n$
- 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 ($\text{Var}(R_n)$):
  \[
  E[R_n] = \sum_{j=1}^{n} E[r_j], \quad \text{Var}(R_n) = \sum_{j=1}^{n} \sum_{i=1}^{n} w_i w_j c_{i,j}
  \]
  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** $\text{Var}(R_n)$ **regardless of its mean**
3. **Minimize the variance for a specified mean** $t$ (parameter): $\min \text{Var}(R_n),$ subject to $E[R_n] = t$
4. **Maximize the mean for a specified variance** $h$ (parameter): $\max E[R_n],$ subject to $\text{Var}(R_n) = h$
5. **Maximize the mean and minimize the variance by using a specified risk preference parameter** $b$: $\max On = E[R_n] - b \text{Var}(R_n)$
Portfolio Retrieval

• The Efficient Frontier:

- Objective function: \( O_n = E[R_n] - b \text{Var}(R_n) \) 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
A mathematical model of diversification

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

\[
Var(R_p) = \sum_j w_j^2 c_{i,j} + 2 \sum_{j=1} \sum_{i=1} w_jw_i c_{i,j}
\]

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 \( \rightarrow \) negative correlation \( \rightarrow \) reduce the risk

---

The Practical Algorithm

- Unlike in finance, the weight \( w_p \) 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:

\[
E[r_j] - bw_j \sigma_j - 2b \sum_{j=1} w_jw_i \sigma_i \sigma_j \rho_{i,j}, \quad b \text{ is a parameter adjusting the risk}
\]
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

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_
Latent Factor Portfolio

- Its expected value:

\[ \hat{R}_n = \sum_{i=1}^{n} w_i \int_{r} r p(r \mid d_i, q) dr \]

\[ = \sum_{i=1}^{n} w_i \sum_{a=1}^{A} \int_{r} r p(r \mid a, q) dr (p(a \mid d_i)) \]

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

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

- 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
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.

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
**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]

**Risk**
- Downside variance: Net loss of relevant docs due to algorithm failures

**Reward**
- Overall precision gain/loss
Algorithm A dominates algorithm B

Risk-Reward Tradeoff Curves

0 500 1000 1500 2000 2500
0 5 10 15 20 25 30 35 40
Percent MAP Gain
R-Loss (Risk increase)

Curves UP and to the LEFT are better

Why IR Models  Individualism  Rank Context  Portfolio Retrieval

High Relevance  High Relevance
Low Risk       High Risk

Expected Relevance

Low Relevance  Low Relevance
Low Risk       High Risk

Variance (Risk)
The typical risk/reward of query expansion: as interpolation parameter $\alpha$ varies

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

“It’s easier to choose the optimal set of equities to buy if you know your tolerance for risk in the market.

If you want to market your skills you can build your own portfolio of stock photographs by choosing the best ones in your collection.

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

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%

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

![Query expansion: Current state-of-the-art method](image1)

![Robust version](image2)

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’

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>court</td>
<td>0.026</td>
</tr>
<tr>
<td>appeals</td>
<td>0.018</td>
</tr>
<tr>
<td>federal</td>
<td>0.012</td>
</tr>
<tr>
<td>employees</td>
<td>0.010</td>
</tr>
<tr>
<td>case</td>
<td>0.010</td>
</tr>
<tr>
<td>education</td>
<td>0.009</td>
</tr>
<tr>
<td>school</td>
<td>0.008</td>
</tr>
<tr>
<td>union</td>
<td>0.007</td>
</tr>
<tr>
<td>seniority</td>
<td>0.007</td>
</tr>
<tr>
<td>salary</td>
<td>0.006</td>
</tr>
</tbody>
</table>

legal aspect is modeled...

BUT

education & pay aspects thrown away..

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

Hypothetical query: ‘merit pay law for teachers’

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>court</td>
<td>0.026</td>
</tr>
<tr>
<td>appeals</td>
<td>0.018</td>
</tr>
<tr>
<td>federal</td>
<td>0.012</td>
</tr>
<tr>
<td>employees</td>
<td>0.010</td>
</tr>
<tr>
<td>case</td>
<td>0.010</td>
</tr>
<tr>
<td>education</td>
<td>0.009</td>
</tr>
<tr>
<td>school</td>
<td>0.008</td>
</tr>
<tr>
<td>union</td>
<td>0.007</td>
</tr>
<tr>
<td>seniority</td>
<td>0.007</td>
</tr>
<tr>
<td>salary</td>
<td>0.006</td>
</tr>
</tbody>
</table>

More balanced query model

Empirical evidence: Udupa, Bhole and Bhattacharya. ICTIR 2009
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

1. Query
2. Top-ranked documents (or other source of term associations)
3. Baseline expansion algorithm
4. Convex optimizer
5. Robust query model
6. Word graph ($\mathcal{S}$):
   - Individual term risk (diagonal)
   - Conditional term risk (off-diagonal)
7. Constraints on word sets
Our approach refines initial estimates from a baseline expansion algorithm

- **Query**
- **Top-ranked documents (or other source of term associations)**
- **Baseline expansion algorithm**
- **Word graph ($\Sigma$):**
  - Individual term risk (diagonal)
  - Conditional term risk (off-diagonal)
- **Convex optimizer**
- **Constraints on word sets**

We don’t assume the baseline is good or reliable!

---

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 $\Sigma$
  - Model uncertainty in $\Sigma$ using regularized $\Sigma_y = \Sigma + \gamma D$
  - **Diagonal:** captures individual term variance (centrality)
  - **Off-diagonal:** term covariance (co-occurrence)

- **Combined objective:**

$$U(x) = -R(x) + \kappa V(x) = -c^T x + \frac{\kappa}{2} x^T (\Sigma + \gamma D) x$$
What are good constraints for query expansion?
Visualization on a word graph:

- 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

- 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

\[
\text{minimize } -c^T x + \frac{1}{2} x^T \Sigma x + \lambda y \\
\text{subject to } \begin{align*}
A x &\leq \mu + \xi, & \text{Risk & reward: Budget} \\
g_i^T x &\geq z_i, & \text{Aspect balance} \\
1 &\leq x &\leq n, & \text{ Aspect coverage} \\
w^T x &\leq y, & \text{Query term support} \\
0 &\leq x &\leq 1 & \text{Budget / sparsity}
\end{align*}
\]

Example solution output

Query: parkinson’s disease

<table>
<thead>
<tr>
<th>Baseline expansion</th>
<th>Convex REXP expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td>parkinson 0.996</td>
<td>parkinson 0.9900</td>
</tr>
<tr>
<td>disease 0.848</td>
<td>disease 0.9900</td>
</tr>
<tr>
<td>syndrome 0.495</td>
<td>syndrome 0.2077</td>
</tr>
<tr>
<td>disorders 0.492</td>
<td>parkinsons 0.1350</td>
</tr>
<tr>
<td>parkinsons 0.491</td>
<td>patients 0.0918</td>
</tr>
<tr>
<td>patient 0.483</td>
<td>brain 0.0256</td>
</tr>
<tr>
<td>brain 0.360</td>
<td>(All other terms zero)</td>
</tr>
<tr>
<td>patients 0.313</td>
<td></td>
</tr>
<tr>
<td>treatment 0.289</td>
<td></td>
</tr>
<tr>
<td>diseases 0.153</td>
<td></td>
</tr>
<tr>
<td>alzheimers 0.114</td>
<td></td>
</tr>
<tr>
<td>... and 90 more...</td>
<td></td>
</tr>
</tbody>
</table>
Convex REXP version dominates the strong baseline version (MAP)
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

Confidence-weighted classifiers treat decision boundary as a random variable

1. Estimate feature weight variance matrix $\Sigma$

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

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

$$\Pr[y_i(w \cdot x_i) \geq 0] \geq \eta$$
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 $t=1:m$

1. Receive example $x_t$ and label $y_t$.
2. If $y_t \mu^T x_t < 1$ then make the following updates:
   \[
   \mu_{t+1} = \mu_t + \alpha_t \Sigma_{t-1} y_t x_t
   \]
   \[
   \Sigma_{t+1} = \Sigma_t - \beta_t \Sigma_{t-1} y_t y_t^T \Sigma_{t-1}
   \]
   where
   \[
   \alpha_t = \frac{\ell_t(\mu^T x_t)}{y_t^2 - 1 - \beta_t}
   \]
   \[
   \beta_t = \frac{1}{x^T x + r}
   \]

Output: $\mu_m$, $\Sigma_m$

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

<table>
<thead>
<tr>
<th>Alteration rule</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;weeble&quot; S-lyrics</td>
<td>0.737</td>
<td>0.153</td>
</tr>
<tr>
<td>&quot;weeble&quot; S-lyrics</td>
<td>0.672</td>
<td>0.139</td>
</tr>
<tr>
<td>&quot;(0)* S-lyrics</td>
<td>0.659</td>
<td>0.134</td>
</tr>
<tr>
<td>&quot;(0)* S-lyrics</td>
<td>0.631</td>
<td>0.119</td>
</tr>
<tr>
<td>&quot;(westing) S-lyrics</td>
<td>0.635</td>
<td>0.180</td>
</tr>
<tr>
<td>&quot;(westing) S-lyrics</td>
<td>0.559</td>
<td>0.135</td>
</tr>
<tr>
<td>&quot;(40) S-lyrics</td>
<td>0.651</td>
<td>0.207</td>
</tr>
<tr>
<td>&quot;(broad) S-lyrics</td>
<td>0.449</td>
<td>0.222</td>
</tr>
<tr>
<td>&quot;(dance) S-lyrics</td>
<td>-0.503</td>
<td>0.188</td>
</tr>
<tr>
<td>&quot;(dance) S-lyrics</td>
<td>-0.521</td>
<td>0.197</td>
</tr>
<tr>
<td>&quot;(dinner) S-lyrics</td>
<td>-0.570</td>
<td>0.181</td>
</tr>
<tr>
<td>&quot;(dance) S-lyrics</td>
<td>-0.520</td>
<td>0.134</td>
</tr>
<tr>
<td>&quot;(dinner) S-lyrics</td>
<td>-0.631</td>
<td>0.190</td>
</tr>
</tbody>
</table>

Adding "lyrics" to a query

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]
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.

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

Understanding IR metrics under uncertainty

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


No free lunch

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

Risking brand: Exploration vs. exploitation

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

Paris Population and Demographics (Paris, TX)
Paris complete population and statistics ... find local info, yellow pages, white pages, demographics and more using AreaConnect Paris
paris.areaconnect.com/statistics.htm [last accessed 2023]

• 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.

[Image of a digital camera search]

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

State-of-the-art personalization is still risky

[Bar chart showing 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]
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.

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, ...)
On-going research directions

- Multimedia retrieval

- Content analysis and fusion

- Advertising

- Collaborative Filtering

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
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

Acknowledgments

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

- History of relevance and IR models

- Classical probabilistic IR model and extensions

Bibliography (cont’d)

- Portfolio Theory and IR applications
Bibliography (cont’d)

- Language modeling for IR

- Federated search / distributed IR / meta-search

- Recommender systems

- Learning-to-rank and rank diversity
Appendix A

The derivation of RSJ Model

RSJ Model (joint prob.)

Given term \( t \), we could have a contingency table to summarize our observation:

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Non-relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term ( t ) Occur</td>
<td>( r_t )</td>
<td>( n_t-r_t )</td>
</tr>
<tr>
<td>Term ( t ) Not Occur</td>
<td>( R-r_t )</td>
<td>( N-R-n_t+r_t )</td>
</tr>
<tr>
<td>( R )</td>
<td>( N-R )</td>
<td>( N )</td>
</tr>
</tbody>
</table>

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

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:

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Non-relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term $t$ Occur</td>
<td>$r_t$</td>
<td>$n_t-r_t$</td>
</tr>
<tr>
<td>Term $t$ Not Occur</td>
<td>$R-r_t$</td>
<td>$N-R-n_t+r_t$</td>
</tr>
<tr>
<td>$R$</td>
<td>$N-R$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

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

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

For simplicity, we define two parameters for $t$

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

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

We thus have

$$P(t \mid q, r = 1) = a'(1 - a)^{1-t}, P(t \mid q, r = 0) = b'(1 - b)^{1-t}$$
RSJ Model (scoring)

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

\[
score(d) = \log \frac{P(r = 1 \mid d, q)}{P(r = 0 \mid d, q)}
\]
\[
= \log \frac{P(d \mid r = 1, q)P(r = 1 \mid q)}{P(d \mid r = 0, q)P(r = 0 \mid q)}
\]
\[
\propto \log \frac{P(d \mid r = 1, q)}{P(d \mid r = 0, q)}
\]

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 (\( t \not\in d \))
  - Independent: terms are conditionally independent with each other given relevance/non-relevance

\[
P(d \mid r = 1, q) = \prod_t P(t \mid q, r = 1) \quad P(d \mid r = 0, q) = \prod_t P(t \mid q, r = 0)
\]
**RSJ Model (scoring)**

- We thus have  \( \text{score}(d) = \log \frac{P(d \mid r = 1, q)}{P(d \mid r = 0, q)} = \sum \log \frac{P(t \mid q, r = 1)}{P(t \mid q, r = 0)} \)

- Replacing the probabilities with the defined parameters gives

\[
\text{score}(d) = \sum \log \frac{a^t (1-a)^{1-r}}{b^t (1-b)^{1-r}} = \sum \log \frac{a^t (1-b)^{1-r}}{b^t (1-a)^{1-r}}
\]

\[
= \sum t \log \frac{a^{1-b}}{b^{1-a}} + \sum \log \frac{1-a}{1-b} \cdot \sum t \log \frac{a^{1-b}}{b^{1-a}} = \sum t \log \frac{a^{1-b}}{b^{1-a}}
\]

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

\[
\text{score}(d) = \sum_{t \in d \cap q} \log \frac{a^{1-b}}{b^{1-a}}
\]

---

**RSJ Model (Bayes’ Rule)**

<table>
<thead>
<tr>
<th>Term t Occur</th>
<th>Relevant</th>
<th>Non-relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term t Not Occur</td>
<td>R-t _</td>
<td>N-R-nt _</td>
</tr>
</tbody>
</table>

Finally, we get

\[
\text{score}(d) = \sum_{t \in d \cap q} \log \frac{a^{1-b}}{b^{1-a}}
\]

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

The derivation of Lafferty and Zhai’s Bayesian Decision Theory of IR
**A Generative IR model**

\[ U \xrightarrow{\text{Model selection}} \theta_Q \xrightarrow{\text{Query generation}} q \]

\[ S \xrightarrow{\text{Model selection}} \theta_D \xrightarrow{\text{Doc generation}} d \]

\[ p(\theta_Q | U) \quad p(q | \theta_Q) \quad p(R | \theta_Q, \theta_D) \quad p(d | \theta_D) \]

John Lafferty, Chengxian 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 \( a=1 \):

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

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

\[
\approx \sum_r \text{Loss}(a = 1 | \hat{\theta}_q, \hat{\theta}_d) p(r | \hat{\theta}_q, \hat{\theta}_d) \quad \text{point estimation } p(\hat{\theta}_q | q) \text{ and } p(\hat{\theta}_d | d) \approx 1
\]

- If a distance-based loss function is used

\[
\text{Loss}(a = 1 | r, \hat{\theta}_q, \hat{\theta}_d) = KL(\hat{\theta}_q, \hat{\theta}_d) = \sum_r p(r | \hat{\theta}_q) \log \frac{p(r | \hat{\theta}_q)}{p(r | \hat{\theta}_d)}
\]

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

- This results in:

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

---
Bayesian Decision Theory in LM

- A further development can show that
  \[ E[\text{Loss}(a = 1|q,d)] \approx KL(\hat{\theta}_q, \hat{\theta}_d) \]
  \[
  = \sum_t p(t|\hat{\theta}_q) \log \frac{p(t|\hat{\theta}_q)}{p(t|\hat{\theta}_d)} \\
  \propto -\sum_t p(t|\hat{\theta}_q) \log p(t|\hat{\theta}_d) + \sum_t p(t|\hat{\theta}_q) \log p(t|\hat{\theta}_q) \\
  \propto -\frac{1}{l_q} \sum_{req} \log p(t|\hat{\theta}_d), \text{ where } l_q \text{ is query length}
  \]
  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