On the Foundations of Diverse Information Retrieval

Scott Sanner, Kar Wai Lim, Shengbo Guo,
Thore Graepel, Sarvnaz Karimi, Sadegh Kharazmi

Outline

- Need for diversity
- The answer: MMR
- But what was the question?
  - Expected n-call@k

Search Result Ranking

Full coverage

NAB to customers: you're the voice on security
Sydney Morning Herald - 1 hour ago
National Australia Bank will begin using voice recognition **technology** to identify its phone customers in the latest move towards the use of biometric security among the big banks. The company said that the **technology**, which identifies a person by their speech ...

NAB speaks loud and clear on voice biometrics
Technology Spectator - 2 hours ago
National Australia Bank (NAB) has joined its peer ANZ Banking Group in touting biometrics as a viable replacement to PINs, with the bank's ambitions focused on voice rather than fingerprint recognition. The move comes hot on the heels of ANZ's recent ...

NAB to shift online banking platform
The Australian - 8 hours ago
NATIONAL Australia Bank's popular internet banking platform could have a new home within six months thanks to a significant **technology** upgrade, a senior company executive said. The development comes as the bank announced plans to further cement its ...

Voice recognition **technology** for NAB
Ninemsn - 11 hours ago
Voice recognition **technology** for NAB. 2:07am November 21, 2012. National Australia Bank will become the first major Australian company to roll out voice recognition **technology**, with plans to introduce it next year. Close calls for journalists caught on video ...

Money talks in hi-tech banking
Courier Mail - 7 hours ago
The **technology** is expected to save individual customers three minutes each phone call. NAB executive general manager Adam Bennett said, when fully deployed, Speech Security would save the bank's customers a combined 15 million minutes a year.

NAB deploys customer data aggregator
iT News - 7 hours ago
Chief **technology** officer Denis McGee said the bank had struck "consumption-based" managed services contracts with key suppliers IBM and Telstra. He told iTnews that the vendors typically already had excess capacity – such as bandwidth on existing fibre ...

NAB phone banking will match customers' voices
Banking Day (registration) - 6 hours ago
After first experimenting with the **technology** in 2009, NAB has quietly enrolled 140,000 customers to trial its system. Essentially, the system authenticates the identity of a person calling into NAB's contact centre by matching the person's voice against a voice ...

- We query the daily news for "technology"

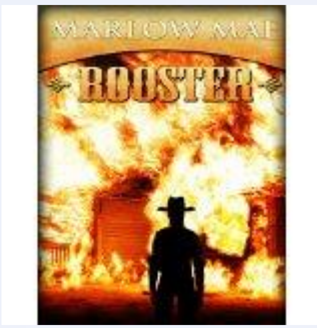
← we get this

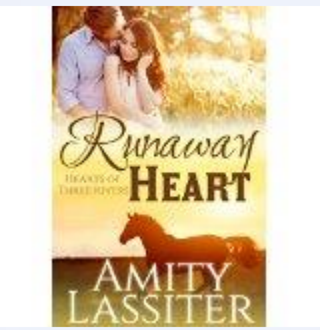
- Is this desirable?

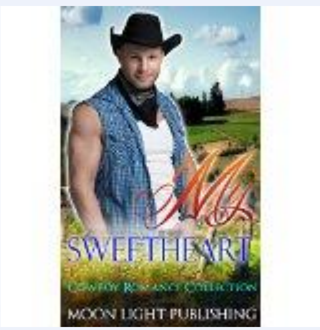
- Note that *de-duplication* does not solve this problem

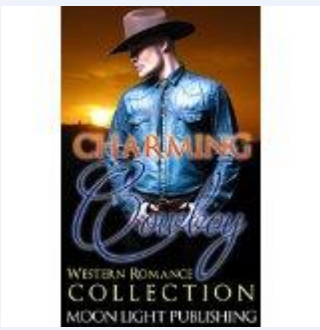
Recommendation

- Book search for “cowboys”*









*These are actual results I got from an e-book search engine.

- Why are they mostly romance books?
  - Will this appeal to all demographics?

Diversity Beyond IR: Machine Learning

- Classifying Computer Science web pages
  - Select top features by some feature scoring metric
    - computer
    - computers
    - computing
    - computation
    - computational
- Certainly all are appropriate
  - But do these cover all relevant web pages well?
  - A better approach? MRMR?

Diversity in IR

- In this talk, focus on diversity from an IR perspective:
  - De-duplication (all search engines handle – locality sensitive hashing)
    - Same page, different URL
    - Different page versions (copied Wiki articles)
  - Source diversity (easy)
    - Web pages vs. news vs. image search vs. Youtube
  - Sense ambiguity (easily addressed through user reformulation)
    - Java, Jaguar, Apple
    - Arguably **not** the main motivation
  - Intrinsic diversity (faceted information needs)
    - Heathrow (checkin, food services, ground transport)
  - Extrinsic diversity (diverse user population)
    - Teens vs. parents, men vs. women, location

How do these relate to previous examples?

Radlinski and Joachims – diverse information needs (SIGIR Forum 2009)

Diversification in IR

- Maximum marginal relevance (MMR)
  - Carbonell & Goldstein, SIGIR 1998
  - *Standard* diversification approach in IR

- MMR Algorithm:
  - S_k is subset of k selected documents from D
  - Greedily build S_k from S_{k-1} where $S_0 = \varnothing$ as follows:

$$s_k^* = \operatorname*{arg\,max}_{s_k \in D \setminus S_{k-1}^*} [\lambda(\mathrm{Sim}_1(\mathbf{q}, s_k)) - (1-\lambda) \max_{s_i \in S_{k-1}^*} \mathrm{Sim}_2(s_i, s_k)]$$

What was the Question?

- MMR is an **algorithm**, we don't know what underlying objective it is optimizing.
- Previous formalization attempts but *full* question unanswered for 14 years
  - Chen and Karger, SIGIR 2006 came closest
- This talk: a complete derivation of MMR
  - Many assumptions
  - Arguably the assumptions you are making when using MMR!

Where do we start?

Let's try to relate set/ranking objective Precision@k to diversity*

*Note: non-standard IR! IR evaluates these objectives empirically but never derives algorithms to directly optimize them! (Largely because long tail queries & no labels.)

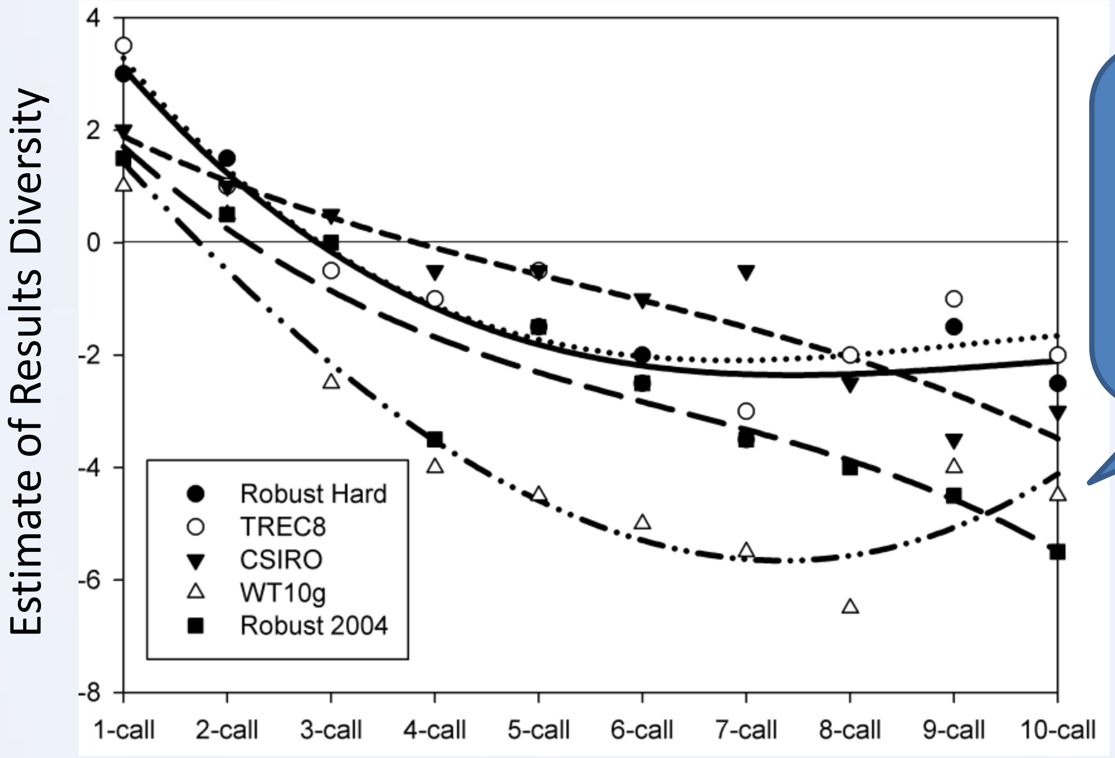
Relating Precision@k Objectives to Diversity

- Chen and Karger, SIGIR 2006: **1-call@k**
  - At least one document in S_k should be relevant (P@k=1)
  - **Very Diverse:** encourages you to "cover your bases" with S_k
    - *Sanner et al*, CIKM 2011: 1-call@k derives MMR with $\lambda = ½$
- van Rijsbergen, 1979: **Probability Ranking Principle (PRP)**
  - Rank items by probability of relevance (e.g., modeled via term freq)
    - PRP relates to k-call@k (P@k=k) which relates to MMR with $\lambda = 1$
  - **Not diverse:** Encourages k^{th} item to be *very similar* to first k-1 items
- So either λ= ½ (1-call@k – very diverse) or λ= 1 (k-call@k – not diverse)?
  - Should really tune λ for MMR based on query ambiguity
    - *Santos, MacDonald, Ounis*, CIKM 2011: Learn best λ given query features
  - So what derives $\lambda \in [½,1]$?
    - Any guesses? ☺

Small fraction of queries have diverse information needs – need good experimental design

Empirical Study of n-call@k

- How does diversity of n-call@k change with n?



Clearly, diversity (pairwise document correlation) decreases with n in n-call@k

J. Wang and J. Zhu. Portfolio theory of information retrieval, SIGIR 2009

Hypothesis

- Let's try optimizing 2-call@k
  - Derivation builds on *Sanner et al*, CIKM 2011
  - Optimizing this leads to MMR with $\lambda = \frac{2}{3}$
- There seems to be a trend relating λ and n:
  - n=1: $\lambda = ½$
  - n=2: $\lambda = \frac{2}{3}$
  - n=k: 1
- Hypothesis
  - Optimizing n-call@k leads to MMR with $\lim_{\{k \to \infty\}} \lambda(k,n) = \frac{n}{n+1}$

Recap

- We wanted to know what objective leads to MMR diversification
- Evidence supports that optimizing n-call@k leads to **diverse** MMR-like behavior where $\lambda = \frac{n}{n+1}$
- Can we derive MMR from n-call@k?

One Detail is Missing…

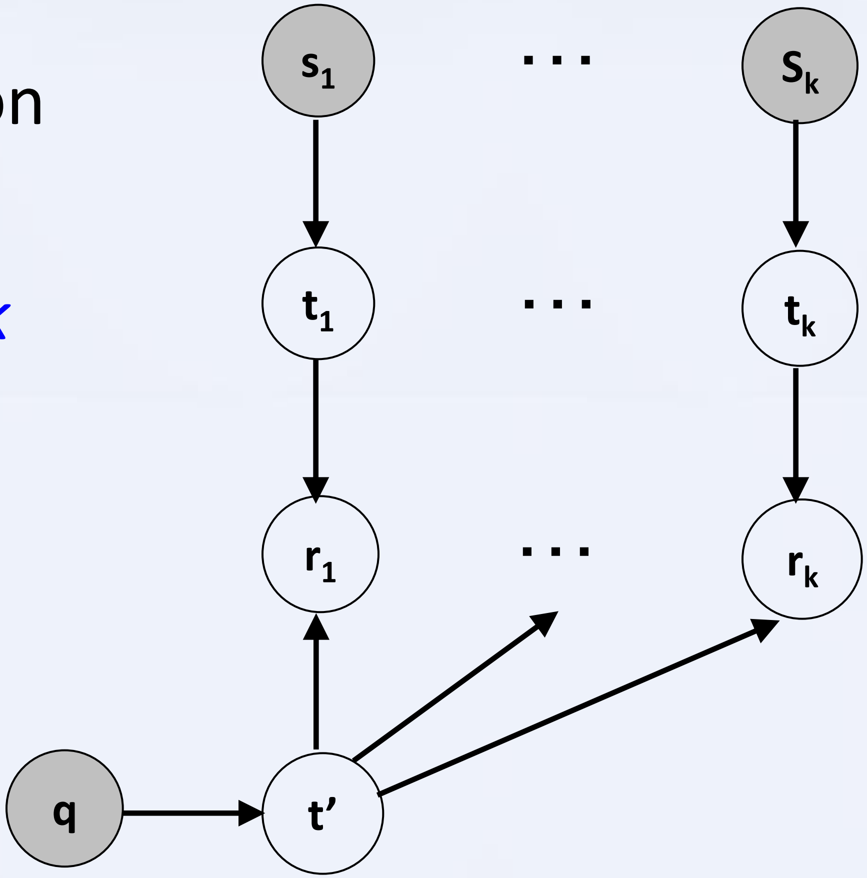
- We want to optimize n-call@k
  - i.e., at least n of k documents should be relevant
  - Great, but given a query and corpus, how do we do this?

- Key question: how to define “relevance”?
  - Need a model for this – probabilistic given PRP connections
  - If diversity needed to cover latent information needs

  → relevance model must include latent query/doc “topics”

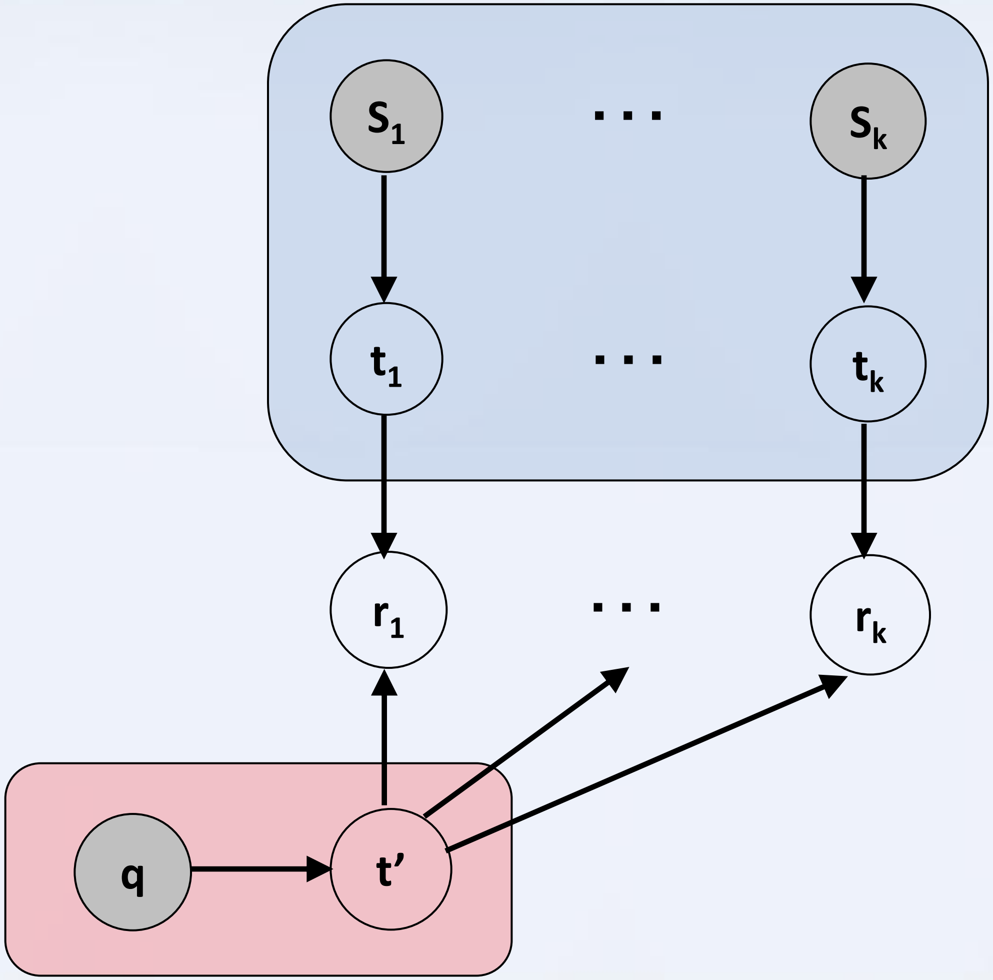
Latent Binary Relevance Retrieval Model

- Formalize as optimization in the graphical model:
  - s_i: doc selection $i=1..k$
  - t_i: topic for i
  - r_i: i relevant?
  - $\boldsymbol{q}$: query
  - t': topic for $\boldsymbol{q}$



How to determine latent topics?

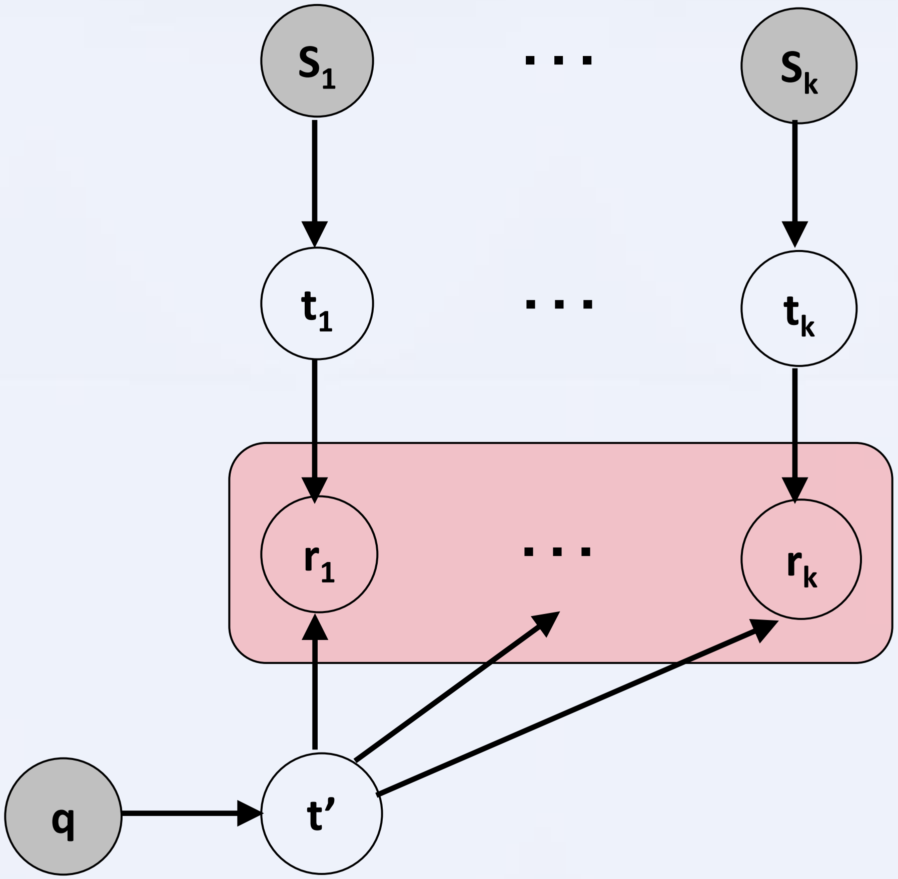
- observed
  latent
- Need CPTs for
  - $P(t_i \mid s_i)$
  - $P(t' \mid \boldsymbol{q})$
- Can…
  - Set arbitrarily
    - Topics are words
    - L1-norm TF
      or TF-IDF!
  - Topic modeling
    (not quite LDA)



Defining Relevance

- Adapt 0-1 loss model of PRP:

$$P(r_i|t', t_i) = \begin{cases} 0 & \text{if } t_i \neq t' \\ 1 & \text{if } t_i = t' \end{cases}$$



Optimizing Expected 1-call@k

$$S^* = \operatorname*{argmax}_{S=\{s_1,\ldots,s_k\}} \text{Exp-1-call@k}(S, \vec{q})$$

All disjuncts mutually exclusive

$$\begin{aligned}
&\text{Exp-1-call@k}(S, \vec{q}) \\
&= \mathbb{E}\left[\bigvee_{i=1}^{k} r_i = 1 \,\middle|\, s_1, \ldots, s_k, \vec{q} \right] = P\left(\bigvee_{i=1}^{k} r_i = 1 | s_1, \ldots, s_k, \vec{q}\right) \\
&= P(\boxed{r_1 = 1} \vee \boxed{[r_1 = 0 \wedge r_2 = 1]} \vee \boxed{[r_1 = 0 \wedge r_2 = 0 \wedge r_3 = 1]} \vee \ldots | s_1, \ldots, s_k, \vec{q}) \\
&= \sum_{i=1}^{k} P(r_i = 1, r_1 = 0, \ldots, r_{i-1} = 0 | s_1, \ldots, s_k, \vec{q}) \\
&= \sum_{i=1}^{k} P(r_i = 1 | r_1 = 0, \ldots, r_{i-1} = 0, s_1, \ldots, s_k, \vec{q}) \boxed{\xcancel{P(r_1 = 0, \ldots, r_{i-1} = 0 | S, \vec{q})}}
\end{aligned}$$

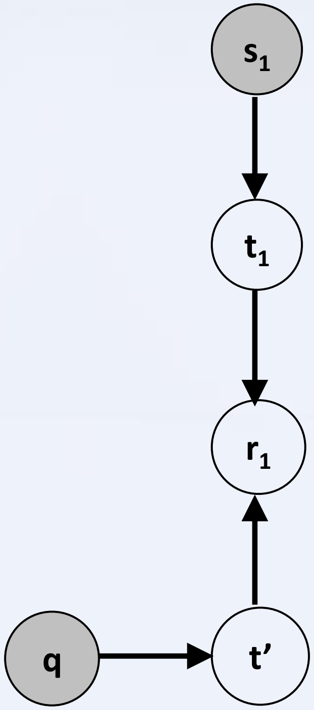
s_k D-separated from $r_1 \ldots r_{k-1}$; so can ignore when greedy!

Greedy: $s_i^* = \operatorname*{argmax}_{s_i} P(r_i = 1 | r_1 = 0, \ldots, r_{i-1} = 0, s_1^*, \ldots, s_{i-1}^*, s_i, \vec{q})$

Objective to Optimize: $s_1{}^*$

- Take a greedy approach (like MMR)
- Choose s_1 via AccRel first

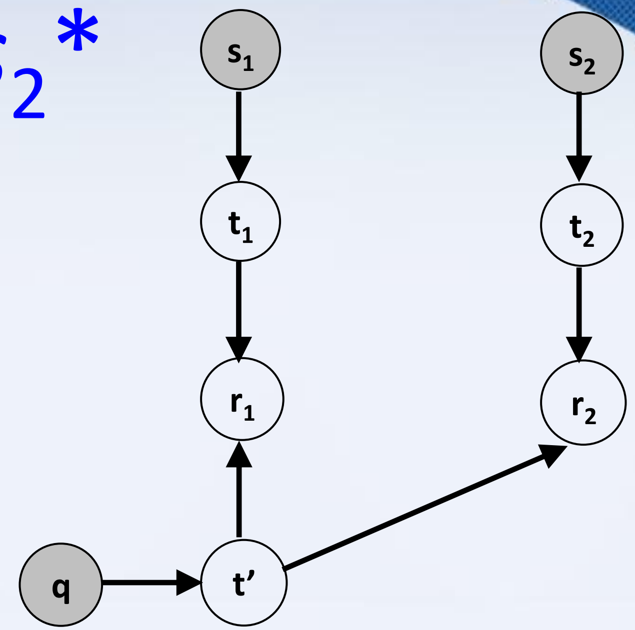
$$
\begin{aligned}
s_1^* &= \arg\max_{s_1} P(r_1 | s_1, \vec{q}) \\
&= \arg\max_{s_1} \sum_{t_1, t'} \mathbb{I}[t' = t_1] P(t' | \vec{q}) P(t_1 | s_1) \\
&= \arg\max_{s_1} \underbrace{\sum_{t'} P(t' | \vec{q}) P(t_1 = t' | s_1)}_{\begin{bmatrix} \vdots \end{bmatrix} \cdot \begin{bmatrix} \vdots \end{bmatrix}}
\end{aligned}
$$



Can derive numerous kernels including TF, TD-IDF, LSI

Objective to Optimize: s_2^*

- Choose s_2 via AccRel next
  - Condition on chosen s_1^* and r_1=0



$$
\begin{aligned}
s_2^* &= \arg\max_{s_2} P(r_2 = 1 | r_1 = 0, s_1^*, s_2, \vec{q}) \\
&= \arg\max_{s_2} \sum_{t_1, t_2, t'} \mathbb{I}[t_2 = t'] P(t_1 | s_1^*) \mathbb{I}[t_1 \neq t'] P(t_2 | s_2) P(t' | \vec{q}) \\
&= \arg\max_{s_2} \sum_{t'} p(t' | \vec{q}) P(t_2 = t' | s_2)(1 - P(t_1 = t' | s_1^*)) \\
&= \arg\max_{s_2} \underbrace{\left[\sum_{t'} p(t' | \vec{q}) P(t_2 = t' | s_2) \right]}_{\text{relevance}} - \underbrace{\left[\sum_{t'} P(t' | \vec{q}) p(t_1 = t' | s_1^*) P(t_2 = t' | s_2) \right]}_{\text{non-diversity penalty}}
\end{aligned}
$$

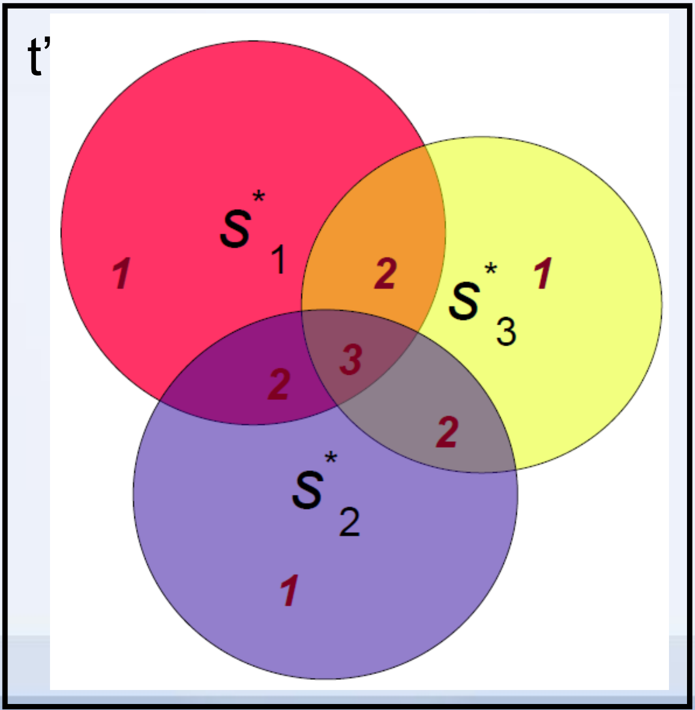
Query-topic weighted diversity!

What is λ? ½.

Objective to Optimize: *s_k*, k>2*

$$s_k^* = \underset{s_k \in D \backslash S_{k-1}}{\arg\max} \sum_{t'} P(t_k = t'|s_k) P(t'|\vec{q}) \boxed{\prod_{i=1}^{k-1} [1 - P(t_i = t'|s_i^*)]}$$

$$\boxed{\prod_{i=1}^{k-1} [1 - P(t_i = t'|s_i^*)]} = 1 - \left[\sum_{i=1}^{k-1} P(t_i = t'|s_i^*) - \sum_{i=1}^{k-1} \sum_{j=1, j \neq i}^{k-1} P(t_i = t'|s_i^*) P(t_j = t'|s_j^*) + \ldots \right]$$



Derives topic t' coverage by **Principle of Inclusion, Exclusion**!

Provides set-covering view of diversity.

So far…

- We've seen hints of MMR from E[1-call@k]
  - Need a few more assumptions to get to MMR

- Let's also generalize to E[n-call@k] for general λ:

$$\text{Exp-}n\text{-Call@}k(S_k, \mathbf{q}) = \mathbb{E}[R_k \geq n | s_1, \ldots, s_k, \mathbf{q}]$$

where $R_k = \sum_{i=1}^{k} r_i$

Optimization Objective

- Continue with greedy approach for E[n-call@k]
  - Select the next document s_k* given all previously chosen documents S_{k-1}:

$$s_k^* = \arg\max_{s_k} \mathbb{E}[R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}]$$

Derivation

- Nontrivial
  - Only an overview of “key tricks” here

- For full details, see
  - Sanner et al, CIKM 2011: 1-call@k (gentler introduction)
    - http://users.cecs.anu.edu.au/~ssanner/Papers/cikm11.pdf
  - Lim et al, SIGIR 2012: n-call@k
    - http://users.cecs.anu.edu.au/~ssanner/Papers/sigir12.pdf

    and online SIGIR 2012 appendix
    - http://users.cecs.anu.edu.au/~ssanner/Papers/sigir12_app.pdf

Derivation

$$
\begin{aligned}
s_k^* &= \arg\max_{s_k} \mathbb{E}[R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}] \\
&= \arg\max_{s_k} P(R_k \geq n | S_{k-1}^*, s_k, \mathbf{q})
\end{aligned}
$$

Derivation

$$
\begin{aligned}
s_k^* &= \arg\max_{s_k} \mathbb{E}[R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}] \\
&= \arg\max_{s_k} P(R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}) \\
&= \arg\max_{s_k} \sum_{T_k} \Big(P(t|\mathbf{q})\, P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s_i^*) \\
&\qquad\qquad \cdot P(R_k \geq n | T_k, S_{k-1}^*, s_k, \mathbf{q}) \Big)
\end{aligned}
$$

Marginalise out all subtopics (using conditional probability)

$T_k = \{t, t_1, \ldots, t_k\}$ and $\sum_{T_k} \circ = \sum_t \sum_{t_1} \cdots \sum_{t_k} \circ$

Derivation

$$
\begin{aligned}
s_k^* &= \arg\max_{s_k} \mathbb{E}[R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}] \\
&= \arg\max_{s_k} P(R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}) \\
&= \arg\max_{s_k} \sum_{T_k} \Big(P(t|\mathbf{q})\, P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s_i^*) \\
&\qquad\qquad \cdot P(R_k \geq n | T_k, S_{k-1}^*, s_k, \mathbf{q}) \Big) \\
&= \arg\max_{s_k} \sum_{T_k} P(t|\mathbf{q})\, P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s_i^*) \\
&\qquad \cdot \Big(\underbrace{P(r_k \geq 0 | R_{k-1} \geq n, t_k, t)}_{1} P(R_{k-1} \geq n | T_{k\text{-}1}) \\
&\qquad\quad + P(r_k = 1 | R_{k-1} = n-1, t_k, t) P(R_{k-1} = n-1 | T_{k\text{-}1}) \Big)
\end{aligned}
$$

We write r_k as conditioned on R_{k-1}, where it decomposes into two independent events, hence the +

Derivation

$$
\begin{aligned}
s_k^* &= \arg\max_{s_k} \mathbb{E}[R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}] \\
&= \arg\max_{s_k} P(R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}) \\
&= \arg\max_{s_k} \sum_{T_k} \Big(P(t|\mathbf{q})\, P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s_i^*) \\
&\qquad \cdot P(R_k \geq n | T_k, S_{k-1}^*, s_k, \mathbf{q}) \Big) \\
&= \arg\max_{s_k} \sum_{T_k} P(t|\mathbf{q})\, P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s_i^*) \\
&\qquad \cdot \Big(\underbrace{P(r_k \geq 0 | R_{k-1} \geq n, t_k, t)}_{1} P(R_{k-1} \geq n | T_{k\text{-}1}) \\
&\qquad + P(r_k = 1 | R_{k-1} = n-1, t_k, t) P(R_{k-1} = n-1 | T_{k\text{-}1}) \Big) \\
&= \arg\max_{s_k} \Bigg(\sum_{T_{k\text{-}1}} \underbrace{\Big[\sum_{t_k} P(t_k|s_k) \Big]}_{1} P(R_{k-1} \geq n | T_{k\text{-}1}) P(t|\mathbf{q}) \prod_{i=1}^{k-1} P(t_i|s_i^*) + \\
&\qquad \sum_t P(t|\mathbf{q}) P(t_k = t | s_k) \sum_{t_1,\ldots,t_{k-1}} P(R_{k-1} = n-1 | T_{k\text{-}1}) \prod_{i=1}^{k-1} P(t_i|s_i^*) \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&\textstyle\sum_{t_k} P(t_k|s_k) P(r_k{=}1 | t_k, t) \\
&= \textstyle\sum_{t_k} P(t_k|s_k) \mathbb{I}[t_k{=}t] = P(t_k{=}t | s_k)
\end{aligned}
$$

Start to push latent topic marginalizations as far in as possible.

Derivation

$$
\begin{aligned}
s_k^* &= \arg\max_{s_k} \mathbb{E}[R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}] \\
&= \arg\max_{s_k} P(R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}) \\
&= \arg\max_{s_k} \sum_{T_k} \Big(P(t|\mathbf{q})\, P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s_i^*) \\
&\qquad\qquad \cdot P(R_k \geq n | T_k, S_{k-1}^*, s_k, \mathbf{q}) \Big) \\
&= \arg\max_{s_k} \sum_{T_k} P(t|\mathbf{q})\, P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s_i^*) \\
&\qquad \cdot \Big(\underbrace{P(r_k \geq 0 | R_{k-1} \geq n, t_k, t)}_{1} P(R_{k-1} \geq n | T_{k\text{-}1}) \\
&\qquad\quad + P(r_k = 1 | R_{k-1} = n-1, t_k, t) P(R_{k-1} = n-1 | T_{k\text{-}1}) \Big) \\
&= \arg\max_{s_k} \Big(\sum_{T_{k\text{-}1}} \underbrace{\Big[\sum_{t_k} P(t_k|s_k) \Big]}_{1} P(R_{k-1} \geq n | T_{k\text{-}1}) P(t|\mathbf{q}) \prod_{i=1}^{k-1} P(t_i|s_i^*) + \\
&\qquad \sum_t P(t|\mathbf{q}) P(t_k = t|s_k) \sum_{t_1,\ldots,t_{k-1}} P(R_{k-1} = n-1 | T_{k\text{-}1}) \prod_{i=1}^{k-1} P(t_i|s_i^*) \Big) \\
&= \arg\max_{s_k} \sum_t P(t|\mathbf{q}) P(t_k = t|s_k) P(R_{k-1} = n-1 | S_{k-1}^*)
\end{aligned}
$$

First term in + is independent of s_k so can remove from max!

Derivation

- We arrive at the simplified

$$s_k^* = \arg\max_{s_k} \mathbb{E}[R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}]$$

$$= \arg\max_{s_k} \sum_t P(t|\mathbf{q}) P(t_k = t | s_k) P(R_{k-1} = n-1 | S_{k-1}^*)$$

- This is still a complicated expression, but it can be expressed recursively…

Recursion

$$P(R_k = n|S_k, t) =$$

$$\begin{cases} n \geq 1, k > 1: & \big(1 - P(t_k = t|s_k)\big) P(R_{k-1} = n|S_{k-1}, t) \\ & \quad + P(t_k = t|s_k) P(R_{k-1} = n-1|S_{k-1}, t) \\ n = 0, k > 1: & \big(1 - P(t_k = t|s_k)\big) P(R_{k-1} = 0|S_{k-1}, t) \\ n = 1, k = 1: & P(t_1 = t|s_1) \\ n = 0, k = 1: & 1 - P(t_1 = t|s_1) \\ n > k: & 0 \end{cases}$$

Very similar conditional decomposition as done in first part of derivation.

Unrolling the Recursion

- We can unroll the previous recursion, express it in closed-form, and substitute:

Where's the max? MMR has a max.

$$s_k^* = \arg\max_{s_k} \sum_t \left(P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j_1,\ldots,j_{n-1}} \prod_{l \in \{j_1,\ldots,j_{n-1}\}} P(t_l = t|s_l^*) \prod_{\substack{i=1 \\ i \notin \{j_1,\ldots,j_{n-1}\}}}^{k-1} (1 - P(t_i = t|s_i^*)) \right)$$

$$n \le k/2$$

$$s_k^* = \arg\max_{s_k} \sum_t \left(P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j_n,\ldots,j_{k-1}} \prod_{l \in \{j_n,\ldots,j_{k-1}\}} (1 - P(t_l = t|s_l^*)) \prod_{\substack{i=1 \\ i \notin \{j_n,\ldots,j_{k-1}\}}}^{k-1} P(t_i = t|s_i^*) \right)$$

$$n > k/2$$

where $j_1, \ldots, j_{n-1} \in \{1, \ldots, k-1\}$ satisfy that $j_i < j_{i+1}$

Deterministic Topic Probabilities

- We assume that the topics of each document are known (deterministic), hence:

$$P(t_i|s_i) \in \{0, 1\}$$

  – Likewise for $\mathrm{P}(\mathrm{t}|\mathbf{q})$
  – This means that a document refers to exactly one topic and likewise for queries, e.g.,
    - If you search for "Apple" you meant *the fruit* OR *the company*, but not both
    - If a document refers to "Apple" *the fruit*, it does not discuss *the company* Apple Computer

Deterministic Topic Probabilities

- Generally:

$$\begin{bmatrix} P(t_i = C_1|s_i) \\ P(t_i = C_2|s_i) \\ \vdots \\ P(t_i = C_{|T|}|s_i) \end{bmatrix} = \begin{bmatrix} 0.24 \\ 0.62 \\ \vdots \\ 0.01 \end{bmatrix}$$

- Deterministic:

$$\begin{bmatrix} P(t_i = C_1|s_i) \\ P(t_i = C_2|s_i) \\ \vdots \\ P(t_i = C_{|T|}|s_i) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Convert a $\prod$ to a max

- Assuming deterministic topic probabilities, we can convert a $\prod$ to a max and vice versa

- For $x_i \in \{0 \text{ (false)}, 1 \text{ (true)}\}$

$$\begin{aligned} \max_i &= \vee_i\, x_i \\ &= \neg \wedge_i (\neg x_i) \\ &= 1 - \wedge_i (1 - x_i) \\ &= 1 - \textstyle\prod_i (1 - x_i) \end{aligned}$$

Convert a $\prod$ to a max

- From the optimizing objective when $n \leq k/2$, we can write

$$\prod_{\substack{i=1 \\ i \notin \{j_1, \ldots, j_{n-1}\}}}^{k-1} \big(1 - P(t_i = t | s_i^*)\big) = 1 - \left(1 - \prod_{\substack{i=1 \\ i \notin \{j_1, \ldots, j_{n-1}\}}}^{k-1} \big(1 - P(t_i = t | s_i^*)\big)\right)$$

$$= 1 - \Big(\max_{\substack{i \in [1, k-1] \\ i \notin \{j_1, \ldots, j_{n-1}\}}} P(t_i = t | s_i^*)\Big)$$

Objective After $\prod$ → max

$$s_k^* = \arg\max_{s_k} \sum_t \left(P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j_1,\ldots,j_{n-1}} \prod_{l \in \{j_1,\ldots,j_{n-1}\}} P(t_l = t|s_l^*) \prod_{\substack{i=1 \\ i \notin \{j_1,\ldots,j_{n-1}\}}}^{k-1} \big(1 - P(t_i = t|s_i^*)\big) \right)$$

$$= \arg\max_{s_k} \sum_t \left(P(t|\mathbf{q}) P(t_k = t|s_k) \sum_{j_1,\ldots,j_{n-1}} \prod_{l \in \{j_1,\ldots,j_{n-1}\}} P(t_l = t|s_l^*) \right.$$
$$\left. - P(t|\mathbf{q}) P(t_k = t|s_k) \sum_{j_1,\ldots,j_{n-1}} \prod_{l \in \{j_1,\ldots,j_{n-1}\}} P(t_l = t|s_l^*) \max_{\substack{i \in [1,k-1] \\ i \notin \{j_1,\ldots,j_{n-1}\}}} P(t_i = t|s_i^*) \right)$$

Combinatorial Simplification

- Deterministic topics also permit combinatorial simplification of some of the $\prod$
- Assuming that **m** documents out of the chosen (k-1) are relevant, then

  $\sum_{j_1,\ldots,j_{n-1}} \prod_{l\in\{j_1,\ldots,j_{n-1}\}} P(t_l = t|s_l^*)$ (the top term) are non-zero $\binom{m}{n-1}$ times.
- $\sum_{j_1,\ldots,j_{n-1}} \prod_{l\in\{j_1,\ldots,j_{n-1}\}} P(t_l = t|s_l^*) \max_{\substack{i\in[1,k-1] \\ i\notin\{j_1,\ldots,j_{n-1}\}}} P(t_i = t|s_i^*)$ (bottom term) are non-zero $\binom{m}{n}$ times.

Final form

- After…
  - assuming a deterministic topic distribution,
  - converting $\prod$ to a max, and
  - combinatorial simplification

$$= \arg\max_{s_k} \binom{m}{n-1} \underbrace{\sum_t P(t|\mathbf{q}) P(t_k = t|s_k)}_{\text{relevance: } \mathrm{Sim}_1(s_k, \mathbf{q})} - \binom{m}{n} \max_{s_i \in S^*_{k-1}} \underbrace{\sum_t P(t_i = t|s_i) P(t|\mathbf{q}) P(t_k = t|s_k)}_{\text{diversity: } \mathrm{Sim}_2(s_k, s_i, \mathbf{q})}$$

$$= \arg\max_{s_k} \frac{n}{m+1} \mathrm{Sim}_1(s_k, \mathbf{q}) - \frac{m-n+1}{m+1} \max_{s_i \in S^*_{k-1}} \mathrm{Sim}_2(s_k, s_i, \mathbf{q})$$

Topic marginalization leads to probability product kernel $\mathrm{Sim}_1(\cdot, \cdot)$: this is any kernel that L_1 normalizes inputs, so can use with TF, TF-IDF! MMR drops **q** dependence in $\mathrm{Sim}_2(\cdot, \cdot)$.

argmax invariant to constant multiplier, use Pascal's rule to normalize coefficients to [0,1]:

$$\binom{m}{n-1} + \binom{m}{n} = \binom{m+1}{n}$$

Comparison to MMR

- The optimising objective used in MMR is

$$s_k^* = \operatorname*{arg\,max}_{s_k \in D \setminus S_{k-1}^*} [\lambda(\mathrm{Sim}_1(\mathbf{q}, s_k)) - (1-\lambda) \max_{s_i \in S_{k-1}^*} \mathrm{Sim}_2(s_i, s_k)]$$

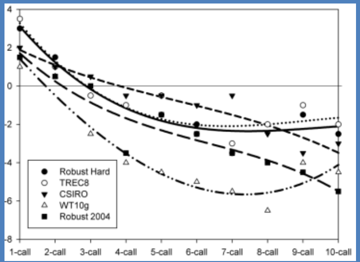
- We note that the optimizing objective for expected n-call@k has the same form as MMR, with $\lambda = \frac{n}{m+1}$.
  – but m is unknown

Expectation of m

- m is expected number of relevant documents ($m \geq n$), we can lower bound m as $m \approx n$.

- With the assumption m=n, we obtain $\lambda = \frac{n}{n+1}$
  - Our hypothesis!

$\lambda = \frac{n}{n+1}$ also roughly follows empirical behavior observed earlier, variation is likely due to m for each corpus



  - If instead m constant, still yields MMR-like algorithm

Summary of Contributions

- We derived MMR from n-call@k!
  - After 14 years, we have insight as to what MMR is optimizing!
  - Don’t like the assumptions?
    - Write down the objective you want
    - Derive the solution!

Bigger Picture: Prob ML for IR

- Search engines are complex beasts
  - Manually optimized
    (which has grown out of empirical IR philosophy)
- But there are probabilistic derivations for popular algorithms in IR
  - TF-IDF, BM25, Language Model
- Opportunity for more modeling, learning, optimization
  - Probabilistic models of (latent) information needs
  - And solutions which autonomously learn and optimize these needs!