On the Mathematical Relationship between Expected n-call@k and the Relevance vs. Diversity Trade-off

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

Feb 21 2013

Outline

- Need for diversity
- The answer: MMR
- Jeopardy: 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 would not solve this problem

Another example

Query for Apple:

- Is this better?

The Answer: Diversity

- When query is ambiguous, diversity is useful
- How can we achieve this?
  - **Maximum marginal relevance (MMR)**
    - Carbonell & Goldstein, SIGIR 1998
    - S_k is subset of k selected documents from D
    - Greedily build S_k from S_{k-1} where $S_0 = \emptyset$:

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

What was the Question?

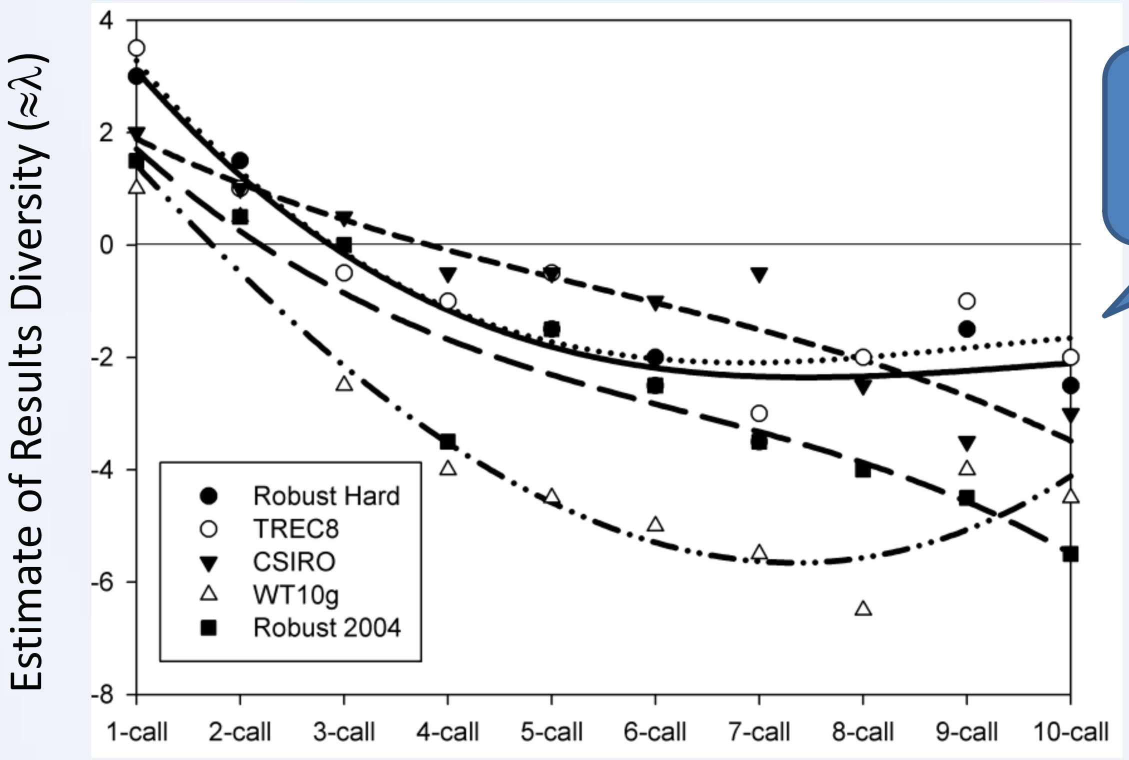
- MMR is an **algorithm**, we don’t know what underlying objective it is optimizing.
- Previous formalization attempts but *full* question unanswered for 14 years
  - Chen and Karger, SIGIR 2006 came closest
- This talk: *one* complete derivation of MMR

What Set-based Objectives Encourage Diversity?

- Chen and Karger, SIGIR 2006: **1-call@k**
  - At least one document in S_k should be relevant
  - **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)
  - **Not diverse:** Encourages k^{th} item to be *very similar* to first k-1 items
  - **k-call@k relates to MMR with $\lambda = 1$, which is PRP**
- So either λ= ½ (1-call@k) or λ= 1 (k-call@k)?
  - 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? ☺

Empirical Study of n-call@k

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



Clearly, λ decreases with n in n-call

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

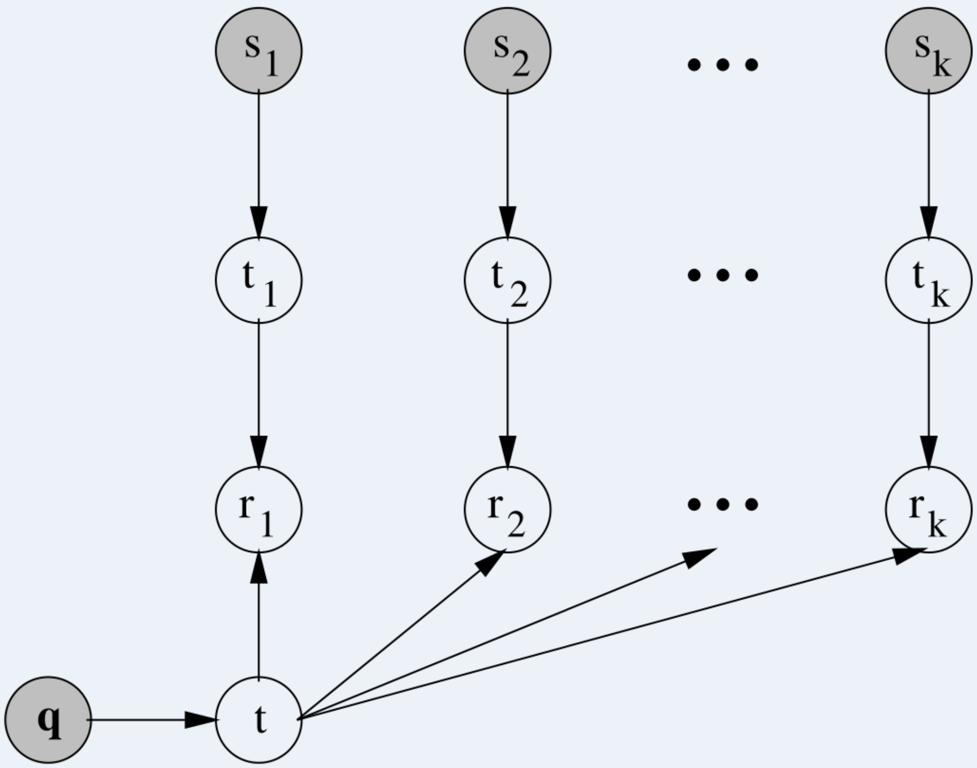
Hypothesis

- 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 $\lim_{\{k\to\infty\}} \lambda(k,n) = \frac{n}{n+1}$

One Detail is Missing…

- We want to optimize n-call@k
  - i.e., at least n of k documents should be relevant
- But what is "relevance"?
  - Need a model for this
  - In particular, one that models query and document ambiguity (via latent topics)
    - Since we hypothesize that topic ambiguity underlies the need for diversity

Graphical Model of Relevance



Latent subtopic binary relevance model

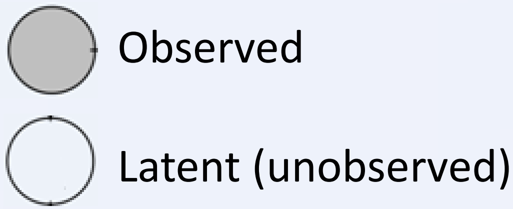
s = selected docs

t = subtopics ∈ T

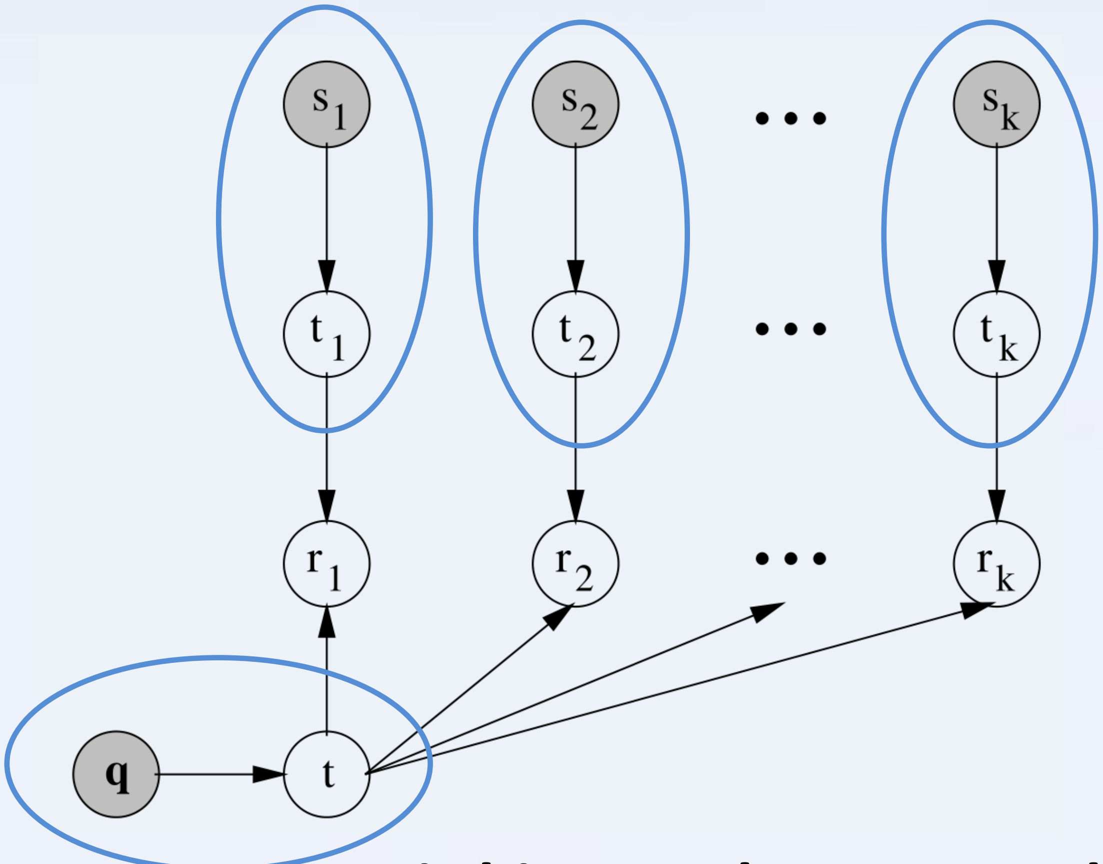
r = relevance ∈ {0, 1}

q = observed query

T = discrete subtopic set
{apple-fruit, apple-inc}



Graphical model of Relevance



Latent subtopic binary relevance model

$P(t_i = C | s_i)$

= prob. of document s belongs to subtopic C

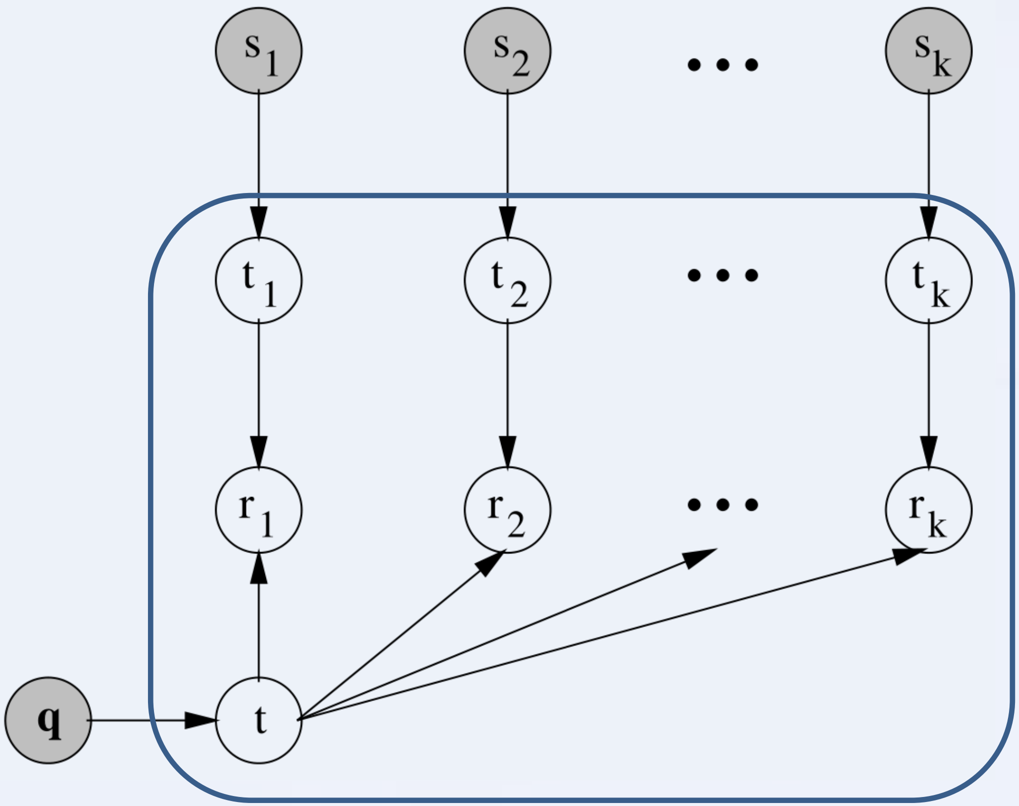
$P(t = C | \mathbf{q})$

= prob. query **q** refers to subtopic C

Observed

Latent (unobserved)

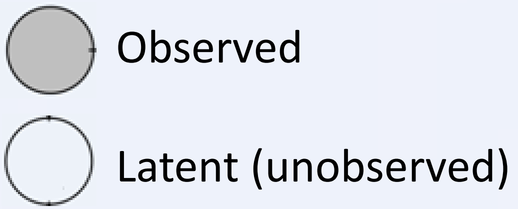
Graphical model of Relevance



Latent subtopic binary relevance model

$P(r_i=1|t_i=t) = 1$

$P(r_i=1|t_i\neq t) = 0$



Optimising Objective

- Now we can compute expected relevance
  - So need to use Expected n-call@k objective:

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

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

- For given query **q**, we want the maximizing S_k
  - Intractable to jointly optimize

Greedy approach

- Like MMR, we'll take a greedy approach
  - 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

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

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

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

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

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

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.

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

$$n \leq k/2$$

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

$$n > k/2$$

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

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

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

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

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

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

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

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

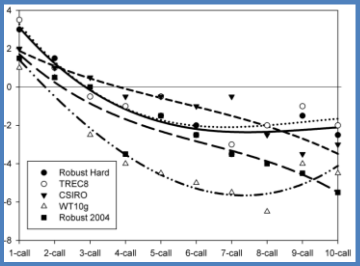
Expectation of m

- Under expected n-call@k's greedy algorithm, after choosing k-1 documents (note that $k \geq n$ and $m \geq n$), we would expect $m \approx n$.

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

m is corpus dependent, but can leave in if wanted; since $m \geq n$ it follows that $\lambda = \frac{n}{n+1}$ is an upper bound on $\lambda = \frac{n}{m+1}$

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



Summary of Contributions

- We showed *the first* derivation of MMR from first principles:
  - MMR optimizes expected n-call@k under the given graphical model of relevance and assumptions
  - After 14 years, gives insight as to what MMR is optimizing!
- This framework can be used to derive *new diversification (or retrieval)* algorithms by changing
  - the graphical model of relevance
  - the set- or rank-based objective criterion
  - the assumptions