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

Kar Wai Lim
ANU & NICTA
Canberra, Australia
karwai.lim@anu.edu.au

Scott Sanner
NICTA & ANU
Canberra, Australia
ssanner@nicta.com.au

Shengbo Guo
Xerox Research Centre Europe
Grenoble, France
shengbo.guo@xrce.xerox.com

ABSTRACT

It has been previously noted that optimization of the n-call@k relevance objective (i.e., a set-based objective that is 1 if at least n documents in a set of k are relevant, otherwise 0) encourages more result set diversification for smaller n, but this statement has never been formally quantified. In this work, we explicitly derive the mathematical relationship between *expected n-call@k* and the *relevance vs. diversity trade-off* — through fortuitous cancellations in the resulting combinatorial optimization, we show the trade-off is a simple and intuitive function of n (notably independent of the result set size $k \geq n$), where diversification increases as $n \to 1$.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

Keywords

diversity, set-based relevance, maximal marginal relevance

1. RELEVANCE VS. DIVERSITY

Subtopic retrieval — "the task of finding documents that cover as many *different* subtopics of a general topic as possible" [5] — is a motivating case for diverse retrieval. One of the most popular result set diversification methods is Maximal Marginal Relevance (MMR) [1]. Formally, given an *item set* D (e.g., a set of documents) where retrieved items are denoted as $s_i \in D$, we aim to select an optimal subset of items $S_k^* \subset D$ (where $|S_k^*| = k$ and $k < |D|$) *relevant* to a given query $\mathbf{q}$ (e.g., query terms) with some level of *diversity* among the items in S_k^*. MMR builds S_k^* in a greedy manner by choosing the next optimal selection s_k^* given the set of $k-1$ optimal selections $S_{k-1}^* = \{s_1^*, \ldots, s_{k-1}^*\}$ (recursively defining $S_k^* = S_{k-1}^* \cup \{s_k^*\}$ with $S_0^* = \emptyset$) as follows:

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

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.

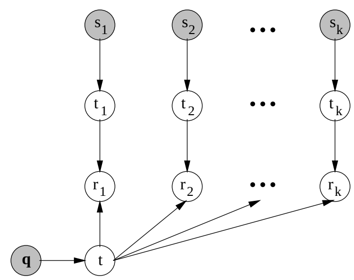

Figure 1: Latent subtopic binary relevance model.

Here, $\lambda \in [0, 1]$, metric Sim_1 measures query-item relevance, and metric Sim_2 measures the similarity between two items.

Presently, little is formally known about how a particular selection of λ relates to the overall *set-based relevance objective* being optimized. However, it has been previously noted that the n-call@k set-based relevance metric (which is 1 if at least n documents in a set of k are relevant, otherwise 0) encourages diversity as $n \to 1$ [2, 4]. Indeed, Sanner *et al.* [3] have shown that optimizing *expected n-call@k* for $n = 1$ corresponds to $\lambda = 0.5$ — we extend this derivation to show that $\lambda = \frac{n}{n+1}$ for arbitrary $n \geq 1$ (independent of result set size $k \geq n$). This result precisely formalizes a relationship between n-call@k and the relevance vs. diversity trade-off.

2. RELEVANCE MODEL AND OBJECTIVE

We review the *probabilistic subtopic model of binary relevance* from [3] shown as a directed graphical model in Figure 1. Shaded nodes represent observed variables, unshaded nodes are latent. Observed variables are the query terms $\mathbf{q}$ and selected items s_i (where for $1 \leq i \leq k$, $s_i \in D$). For the subtopic variables, let T be a discrete subtopic set. Then $t_i \in T$ represent subtopics for respective s_i and $t \in T$ represents a subtopic for query $\mathbf{q}$. The r_i are $\{0, 1\}$ variables that indicate if respective selected items s_i are relevant ($r_i = 1$).

The conditional probability tables (CPTs) are as follows: $P(t_i|s_i)$ and $P(t|\mathbf{q})$ respectively represent the subtopic distribution for item s_i and query $\mathbf{q}$. For the r_i CPTs, using $\mathbb{I}[\cdot]$ as a $\{0, 1\}$ indicator function (1 if $\cdot$ is true), item s_i is deemed *relevant iff its subtopic t_i matches query subtopic t*:

$$P(r_i = 1|t, t_i) \; = \; \mathbb{I}[t_i = t]$$

We next define $R_k = \sum_{i=1}^{k} r_i$, where R_k is the number of relevant items from the first k selections. Reading $R_k \geq n$ as $\mathbb{I}[R_k \geq n]$, we express the *expected n-call@k* objective as

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

3. MAIN DERIVATION AND RESULT

Taking MMR's greedy approach, we select s_k given S^*_{k-1}:[1]

$$s^*_k = \arg\max_{s_k} \mathbb{E}[R_k \geq n | S^*_{k-1}, s_k, \mathbf{q}] = \arg\max_{s_k} P(R_k \geq n | S^*_{k-1}, s_k, \mathbf{q})$$

This query can be evaluated w.r.t. our latent subtopic binary relevance model in Figure 1 as follows, where we marginalize out all non-query, non-evidence variables T_k and define $T_k = \{t, t_1, \ldots, t_k\}$ and $\sum_{T_k} \circ = \sum_t \sum_{t_1} \cdots \sum_{t_k} \circ$:

$$= \arg\max_{s_k} \sum_{T_k} \Big(P(t|\mathbf{q})\, P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s^*_i) \cdot P(R_k \geq n | T_k, S^*_{k-1}, s_k, \mathbf{q}) \Big)$$

We split $R_k \geq n$ into two disjoint (additive) events ($r_k \geq 0, R_{k-1} \geq n$), ($r_k = 1, R_{k-1} = n-1$) where all r_i are D-separated:

$$= \arg\max_{s_k} \sum_{T_k} P(t|\mathbf{q})\, P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s^*_i) \cdot \Big(\underbrace{P(r_k \geq 0 | R_{k-1} \geq n, t_k, t)}_{1} P(R_{k-1} \geq n | T_{k-1}) + P(r_k = 1 | R_{k-1} = n-1, t_k, t) P(R_{k-1} = n-1 | T_{k-1}) \Big)$$

We distribute initial terms over the summands noting that $\sum_{t_k} P(t_k|s_k) P(r_k = 1 | t_k, t) = \sum_{t_k} P(t_k|s_k) \mathbb{I}[t_k = t] = P(t_k = t | s_k)$:

$$= \arg\max_{s_k} \Bigg(\sum_{T_{k-1}} \Big[\underbrace{\sum_{t_k} P(t_k|s_k)}_{1} \Big] P(R_{k-1} \geq n | T_{k-1}) P(t|\mathbf{q}) \prod_{i=1}^{k-1} P(t_i|s^*_i) + \sum_t P(t|\mathbf{q}) P(t_k = t|s_k) \sum_{t_1, \ldots, t_{k-1}} P(R_{k-1} = n-1 | T_{k-1}) \prod_{i=1}^{k-1} P(t_i|s^*_i) \Bigg)$$

Next we proceed to drop the first summand since it is not a function of s_k (i.e., it has no influence in determining s^*_k):

$$= \arg\max_{s_k} \sum_t P(t|\mathbf{q}) P(t_k = t|s_k) P(R_{k-1} = n-1 | S^*_{k-1}) \qquad (2)$$

By similar reasoning, we can derive that the last probability needed in (2) is recursively defined as $P(R_k = n | S_k, t) =$

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

We can now rewrite (2) by unrolling its recursive definition. For expected n-call@k where $n \leq k/2$ (a symmetrical result holds for $k/2 < n \leq k$), the explicit unrolled objective is

$$s^*_k = \arg\max_{s_k} \sum_t \Bigg(P(t|\mathbf{q})\, P(t_k = t|s_k) \cdot \sum_{j_1, \ldots, j_{n-1}} \prod_{l \in \{j_1, \ldots, j_{n-1}\}} P(t_l = t|s^*_l) \prod_{\substack{i=1 \\ i \notin \{j_1, \ldots, j_{n-1}\}}}^{k-1} \big(1 - P(t_i = t|s^*_i)\big) \Bigg) \qquad (3)$$

where $j_1, \ldots, j_{n-1} \in \{1, \ldots, k-1\}$ satisfy that $j_i < j_{i+1}$ (i.e., an ordered permutation of $n-1$ result set indices).

[1]We present a derivation summary; A full derivation may be found in an online appendix at the authors' web pages.

If we assume each document covers a single subtopic of the query (e.g., a subtopic represents an intent of an ambiguous query) then we can assume that $\forall i\ P(t_i|s_i) \in \{0,1\}$ and $P(t|\mathbf{q}) \in \{0,1\}$. This allows us to convert a $\prod$ to a max

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

and by substituting this into (3) and distributing, we get

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

Assuming m selected documents S^*_{k-1} are relevant then the top term (specifically $\prod_l$) is non-zero $\binom{m}{n-1}$ times. For the bottom term, it takes $n-1$ relevant S^*_{k-1} to satisfy its $\prod_l$, and one additional relevant document to satisfy the $\max_i$ making it non-zero $\binom{m}{n}$ times. Factoring out the max element from the bottom and pushing the $\sum_t$ inwards (all legal due to the $\{0,1\}$ subtopic probability assumption) we get

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

From here we can normalize by $\binom{m}{n-1} + \binom{m}{n} = \binom{m+1}{n}$ (Pascal's rule), leading to fortuitous cancellations and the result:

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

Comparing to MMR in (1), we can clearly see that $\lambda = \frac{n}{m+1}$. Assuming $m \approx n$ since Exp-n-Call@k optimizes for the case where n relevant documents are selected, then $\lambda = \frac{n}{n+1}$.

Acknowledgements

NICTA is funded by the Australian Government via the Dept. of Broadband, Comm. and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

4. REFERENCES

[1] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reording documents and producing summaries. In *SIGIR-98*. ACM, 1998.

[2] H. Chen and D. R. Karger. Less is more: Probabilistic models for retrieving fewer relevant documents. In *SIGIR-06*. ACM, 2006.

[3] S. Sanner, S. Guo, T. Graepel, S. Kharazmi, and S. Karimi. Diverse retrieval via greedy optimization of expected 1-call@k in a latent subtopic relevance model. In *CIKM-11*. ACM, 2011.

[4] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR-09*. ACM, 2009.

[5] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *SIGIR-03*. ACM, 2003.