On the Formal Relationship between Expected n-call@k and Diversity

Shengbo Guo, Xerox Research Centre Europe
Kar Wai Lim, Australian National University–National ICT Australia
Scott Sanner, National ICT Australia–Australia National University

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 $n \to 1$. Empirical results on three diversity testbeds including the TREC 6-8 Interactive Track, 2009 and 2010 ClueWeb Diversity tasks of the TREC Web Track support our theoretical derivations.

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

General Terms: Algorithms

Additional Key Words and Phrases: diversity, set-based relevance, graphical models, maximal marginal relevance

ACM Reference Format:
Guo, S., Lim, K. W., Sanner, S.. Diverse Retrieval with Expected n-call@k in a Latent Subtopic Relevance Model. ACM Trans. Embedd. Comput. Syst. 9, 4, Article 39 (March 2010), 17 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

One of the basic tenets of set-based information retrieval is to minimize redundancy, hence maximize diversity, in the result set to increase the chance that the results will contain items relevant to the user's query [Goffman 1964]. Hence, *diverse retrieval* can be defined as a *set-level* retrieval objective that takes into account inter-document relevance dependences when producing a result set relevant to a query.

Subtopic retrieval — "the task of finding documents that cover as many *different* subtopics of a general topic as possible" [Zhai et al. 2003] — is a motivating case for diverse retrieval. That is, if a query has multiple facets that should be covered by a result set, or a query has multiple ambiguous interpretations, then a retrieval algorithm should try to "cover" all of these subtopics in its result set.

One of the most popular result set diversification methods is Maximal Marginal Relevance (MMR) [Carbonell and Goldstein 1998]. 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 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 $k-1$

Author's addresses: S. Guo, Xerox Research Centre Europe; K. W., The Australian National University; S. Sanner, National ICT Australia (NICTA) and The Australian National University.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

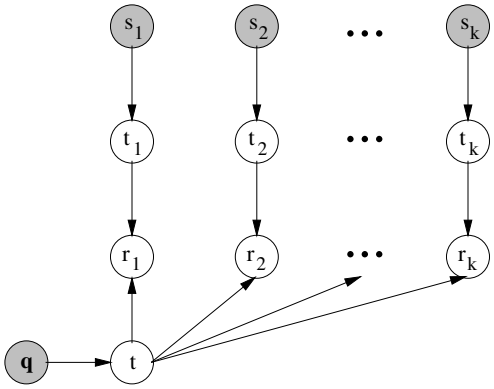

Fig. 1. Latent subtopic binary relevance model.

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:

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

Here, $\lambda \in [0, 1]$, metric Sim_1 measures query-item relevance, and metric Sim_2 measures the similarity between two items. In the case of s^*_1, the $\max$ term is omitted.

In MMR, we note that the λ term explicitly controls the trade-off between relevance and diversity. This λ term has been traditionally set in an ad-hoc manner or in recent work, learned in a query-specific way from data [Santos et al. 2010b].

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 $n \to 1$ [Chen and Karger 2006; Wang and Zhu 2009]. Indeed, Sanner *et al.* [Sanner et al. 2011] 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 [Sanner et al. 2011] 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:

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

Here, $\mathbb{I}[\cdot]$ is 1 when its argument is true and 0 otherwise. 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

$$\text{Exp-}n\text{-Call@}k(S_k, \mathbf{q}) = \mathbb{E}[R_k \geq n | s_1, \ldots, s_k, \mathbf{q}]. \tag{2}$$

Since jointly optimizing Exp-n-Call@$k(S_k, \mathbf{q})$ is NP-hard, we take a greedy approach similar to MMR where we choose the best s_k^* assuming that S_{k-1}^* is given, and present the main derivaiton results in the following section.

3. THEORETICAL RESULTS

3.1. Optimizing Expected 1-call@k

Before we present the result on optimzing Expected n-call@k, we first introdue the our derivation on optimzing Expected 1-call@k. We first formally define the *expected 1-call@k* objective:

$$\text{Exp-1-Call@}k(S_k, \mathbf{q}) = \mathbb{E}\left[\bigvee_{i=1}^{k} r_i = 1 \,\middle|\, s_1, \ldots, s_k, \mathbf{q}\right] \tag{3}$$

Since jointly optimizing Exp-1-Call@$k(S_k, \mathbf{q})$ is NP-hard, we take a greedy approach similar to MMR where we choose the best s_k^* assuming that S_{k-1}^* is given. Then following [Chen and Karger 2006], we can greedily optimize this objective as follows:[1]

$$\begin{aligned}
s_k^* &= \arg\max_{s_k} \ \text{Exp-1-Call@}k(S_{k-1}^* \cup \{s_k\}, \mathbf{q}) \\
&= \arg\max_{s_k} \mathbb{E}\left[\bigvee_{i=1}^{k} r_i = 1 \,\middle|\, S_{k-1}^*, s_k, \mathbf{q}\right] \\
&= \arg\max_{s_k} \mathbb{E}\Bigg[(r_1 = 1) \vee (r_2 = 1 \wedge r_1 = 0) \vee \cdots \vee \\
&\qquad \left(r_k = 1 \wedge \bigwedge_{i=1}^{k-1} r_i = 0\right) \,\Bigg|\, S_{k-1}^*, s_k, \mathbf{q}\Bigg] \\
&= \arg\max_{s_k} \sum_{i=1}^{k} P(r_i = 1, \{r_j = 0\}_{j<i} \,|\, \{s_j^*\}_{j \leq i, j<k}, \{s_k\}_{k=i}, \mathbf{q}) \\
&= \arg\max_{s_k} \sum_{i=1}^{k} P(r_i = 1 \,|\, \{r_j = 0\}_{j<i}, \{s_j^*\}_{j \leq i, j<k}, \{s_k\}_{k=i}, \mathbf{q}) \\
&\qquad P(\{r_j = 0\}_{j<i} \,|\, \{s_j^*\}_{j \leq i, j<k}, \{s_k\}_{k=i}, \mathbf{q}) \\
&= \arg\max_{s_k} P(r_k = 1 \,|\, \{r_j = 0\}_{j<k}, S_{k-1}^*, s_k, \mathbf{q})
\end{aligned} \tag{4}$$

Here, we applied a logical equivalence, exploited additivity of exclusive events, rewrote the expectation of a binary event as its probability, exploited d-separation to remove irrelevant conditions, factorized each joint into a conditional and prior, and removed terms and factors independent of s_k. Thus, we need only maximize s_k's probability of relevance conditioned on the query and previous selections (assumed irrelevant).

[1]The notation $\{\cdot\}_C$ refers to a (possibly empty) set of variables (or variable assignments) $\cdot$ that meet constraints C.

Next we evaluate the final query from (4) w.r.t. our graphical model of subtopic relevance from Figure 1:

$$\begin{aligned}
s_k^* &= \arg\max_{s_k} P(r_k = 1 \mid \{r_j = 0\}_{j<k}, S_{k-1}^*, s_k, \mathbf{q}) \\
&= \arg\max_{s_k} \sum_{t_1,\cdots,t_k,t} P(t|\mathbf{q}) P(t_k|s_k) \mathbb{I}[t_k = t] \prod_{i=1}^{k-1} P(t_i|s_i^*) \mathbb{I}[t_i \neq t] \\
&= \arg\max_{s_k} \sum_t P(t|\mathbf{q}) \sum_{t_k} P(t_k|s_k) \mathbb{I}[t_k = t] \prod_{i=1}^{k-1} \sum_{t_i} P(t_i|s_i^*) \mathbb{I}[t_i \neq t] \\
&= \arg\max_{s_k} \sum_t P(t|\mathbf{q}) P(t_k = t|s_k) \boxed{\prod_{i=1}^{k-1} (1 - P(t_i = t|s_i^*))}
\end{aligned} \tag{5}$$

Defining $\tilde{P}(t|S_{k-1}^*) = 1 - \square = 1 - \prod_{i=1}^{k-1}(1 - P(t_i = t|s_i^*))$, this is the probability that set S_{k-1}^* already *covers* topic t w.r.t. a *noisy-or* interpretation. Substituting $(1 - \tilde{P}(t|S_{k-1}^*))$ for $\square$ since $(1 - \tilde{P}(t|S_{k-1}^*)) = 1 - (1 - \square) = \square$, we obtain

$$\begin{aligned}
s_k^* &= \arg\max_{s_k} \sum_t P(t|\mathbf{q}) P(t_k = t|s_k) \left(1 - \tilde{P}(t|S_{k-1}^*)\right) \\
&= \arg\max_{s_k} \sum_t \underbrace{P(t|\mathbf{q}) P(t_k = t|s_k)}_{\text{query similarity}} \\
&\qquad - \sum_t \underbrace{P(t|\mathbf{q}) P(t_k = t|s_k) \tilde{P}(t|S_{k-1}^*)}_{\text{query-reweighted diversity}}.
\end{aligned} \tag{6}$$

3.2. Optimizing Expected n-call@k

Let us now generalize our results in the Expected 1-call@k to the Expected n-call@k. We cast the optimization of Exp-n-Call@$k(S_k, \mathbf{q})$ in a form similar to MMR in (1) to determine the correspondence between λ and the result of this derivation. Taking MMR's greedy approach, we select s_k assuming that S_{k-1}^* is already chosen:

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

In the second step we have exploited the $\{0, 1\}$ nature of $R_k \geq n$ to rewrite Exp-n-Call@k directly as a probabilistic query. 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:

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

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

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

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^*):

$$=\arg\max_{s_k}\sum_t P(t|\mathbf{q})P(t_k{=}t|s_k)P(R_{k-1}{=}n-1|S_{k-1}^*) \tag{7}$$

By similar reasoning, we can derive that the last probability needed in (16) 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 (16) 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

$$s_k^*=\arg\max_{s_k}\sum_t\Bigg(P(t|\mathbf{q})\,P(t_k=t|s_k)\cdot$$
$$\sum_{j_1,\ldots,j_{n-1}}\prod_{l\in\{j_1,\ldots,j_{n-1}\}}P(t_l{=}t|s_l^*)\prod_{\substack{i=1\\ i\notin\{j_1,\ldots,j_{n-1}\}}}^{k-1}\big(1-P(t_i{=}t|s_i^*)\big)\Bigg) \tag{8}$$

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).

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 $\max$

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

and by substituting this into (17) 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:

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

Fortuitously, we note that the $\binom{m+1}{n}$ divisor cancelled with the numerators, yielding this elegant and interpretable result. 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 $\lambda = \frac{n}{n+1}$, which achieves our goal of formally expressing the relevance vs. diversity tradeoff as a function of n, k, and m.

As a reality check, we see that this coincides with the published result of $\lambda = 0.5$ in [Sanner et al. 2011] for $n = 1$, $m = 1$. Overall we have achieved our goal and have shown that indeed, diversificiation in expected n-call@k decreases linearly as $n \to 1$.

4. RELATED WORK

As early as 1964, Goffman [Goffman 1964], a mathematical information science pioneer [Harmon 2008], notes that the relevance of documents in a list has to depend on the documents preceding it. More recently, work on MMR [Carbonell and Goldstein 1998] was one of the first to formalize diversification as a mathematical optimization criterion; MMR has proved one of the most popular diversity approaches. Aside from this work, two of the other notable works are [Yue and Joachims 2008], which formalizes a structured SVM loss function based on a set covering objective, and [Wang and Zhu 2009], which borrows concepts from portfolio theory in economics to treat result set diversification as optimization of a risk minimization objective. We note that PLAR

Table I. Dimensions of diversified set-based retrieval systems: *Probabilistic*: uses probabilistic models?; *Latent*: use latent topic models?; *Learning*: uses some form of learning?; *Unsupervised*: does not require labeled topic data or feedback?

Diversity Paper	Probabilistic	Latent	Learning	Unsupervise
Carbonell [Carbonell and Goldstein 1998]				√
Anagnostop [Anagnostopoulos et al. 2005]	√			√
Radlinski [Radlinski and Dumais 2006]				√
Radlinski [Radlinski et al. 2008]	√		√	
Clarke [Clarke et al. 2008]	√	√		√
Yue [Yue and Joachims 2008]		√	√	
Bai [Bai and Nie 2008]			√	√
Sanderson [Sanderson 2008]				√
Yu [Yu et al. 2009]				√
Agrawal [Agrawal et al. 2009]	√	√		√
Gollapudi [Gollapudi and Sharma 2009]				√
Clough [Clough et al. 2009]	√			√
Song [Song et al. 2009]	√		√	
Wang [Wang and Zhu 2009]		√		√
Neal [Lathia et al. 2010]				
Zhao [Zhao et al. 2012]				
Dang [Dang and Croft 2012]	√	√		√
Vargas [Vargas et al. 2012]	√		√	√
Our model (this paper)	√	√	√	√

as derived in the last section formally motivates these last three somewhat ad-hoc diversification approaches and we discuss these connections more deeply in the following sections.

Before we discuss specifics though, we provide a general summary of the result set diversification literature in Table I. Here we breakdown the different proposals according to whether they are probabilistic, use latent models for determining similarity, use adaptive learning techniques, and finally whether they are unsupervised, i.e., they do not require labeled data or feedback. We note that our model is the first proposal (that we are aware of) to combine all four traits in the affirmative.

4.1. MMR

The result in (6) is strikingly similar to MMR — it contains two terms, one for query similarity and the other for result set diversification, where each term represents a similarity kernel — more specifically a *probability product kernel* (PPK) [Jebara et al. 2004] that is an inner product of probability vectors (or more generally, functions). More formally, let $\mathbf{T}'$, $\mathbf{T}_k$, and $\mathbf{T}_{S^*_{k-1}}$ be respective topic probability vectors $P(t' = t|\mathbf{q})$, $P(t_k = t|s_k)$ and $\tilde{P}(t_k = t|S^*_{k-1})$ with vector indices for each topic $t \in T$. Then the similarity and diversity terms from (6) can be respectively written as

$$\sum_{t \in T} P(t' = t|\mathbf{q})P(t_k = t|s_k) \;=\; \langle \mathbf{T}', \mathbf{T}_k \rangle \text{ and} \tag{9}$$

$$\sum_{t \in T} P(t|\mathbf{q})P(t_k = t|s_k)\tilde{P}(t|S^*_{k-1}) \;=\; \langle \mathbf{T}_k, \mathbf{T}_{S^*_{k-1}} \rangle_{\mathbf{T}'}. \tag{10}$$

Here, we let $\langle \cdot, \cdot \rangle$ denote an inner product of two vectors and $\langle \cdot, \cdot \rangle_{\mathbf{v}}$ a $\mathbf{v}$-*reweighted* inner product, defined as in (10).

While having similarity and diversity terms similar to MMR, Exp-1-call@k in (6) clearly differs from MMR:

(1) While MMR's definition allows for any similarity function, not just PPKs, we note that *equating words to subtopics*, popular kernels like TF and TFIDF [Salton and

McGill 1983] can be viewed directly as PPKs if the TF and TFIDF vectors are L_1 normalized to represent probability vectors.

(2) MMR uses a maximization term for diversity, whereas optimization of Exp-1-call@k instead calls for a product (noisy-or) diversity term $\tilde{P}(t|S^*_{k-1})$. We note that a noisy-or reduces to a max when the subtopic probabilities are deterministic (0 or 1).

(3) While MMR proposes a λ term to explicitly trade off the similarity and diversity terms, the greedy optimization of Exp-1-call@k in (6) yields no such trade-off term (or alternately, an implicit $\lambda = .5$). Although it seems a tunable λ is not needed for maximizing Exp-1-call@k, it may be desirable when maximizing surrogate retrieval objectives (e.g., ranking objectives).

(4) Optimizing Exp-1-call@k introduces query-specific relevance into the diversification term as shown by the query topic ($\mathbf{T}'$) reweighted diversity function in (10).

4.2. Other Diversification Approaches

Recent years have seen numerous proposals for diversification approaches and here we summarize the relationship between optimization of Exp-1-call@k and representatives of these alternative approaches:

Portfolio Theory: [Wang and Zhu 2009] motivates diversification in set-based information retrieval by a risk-minimizing portfolio selection approach. Viewing a result set as an investment portfolio with the objective to maximize return while minimizing risk, the derived result of [Wang and Zhu 2009] mimics both MMR and Exp-1-call@k in that the similarity term may be viewed as *expected portfolio payoff* (relevance) and the diversity term may be viewed as *expected portfolio risk*, which increases as the correlations between documents in the result set increase. One major difference in this framework is that rather than computing the diversity term via a max (MMR) or product (Exp-1-call@k) the portfolio theory derivation uses a summation — we examine the implications of this next.

Set Covering: Yue and Joachims [Yue and Joachims 2008] propose a set covering approach for training SVMs to predict diverse result sets for information retrieval. In their work, they equate subtopics with words and build a loss function for SVM training that penalizes result sets according to the sum of weights of query-relevant words *not* covered by the result set. While their approach provides a "hard" set-covering view of diversity, we note that an expansion of $\tilde{P}(t|S^*_{k-1})$ used in the diversity term of (6) provides a "soft" latent set-covering interpretation; that is, s_k is chosen so as to best cover (in a probabilistic sense) the latent topic space not already covered by $\{s^*_1, \ldots, s^*_{k-1}\}$. Formally, expanding the product in $\tilde{P}(t|S^*_{k-1}) = \prod_{i=1}^{k-1}(1 - P(t_i = t|s^*_i))$, collecting terms and writing it as a series, we arrive at a form that reflects the inclusion-exclusion principle applied to the calculation of probability that topic t is covered by $\{s^*_1, \ldots, s^*_{k-1}\}$:

$$\begin{aligned}
&\prod_{i=1}^{k-1}(1 - P(t_i = t|s^*_i)) \\
&= 1 - \left[\sum_{i=1}^{k-1} P(t_i = t|s^*_i) - \sum_{i=1}^{k-1}\sum_{j=1}^{k-1} P(t_i = t|s^*_i)P(t_j = t|s^*_j)\right. \\
&\qquad \left. + \cdots - (-1)^{k-1}\prod_{i=1}^{k-1} P(t_i = t|s^*_i)\right] \qquad (11)
\end{aligned}$$

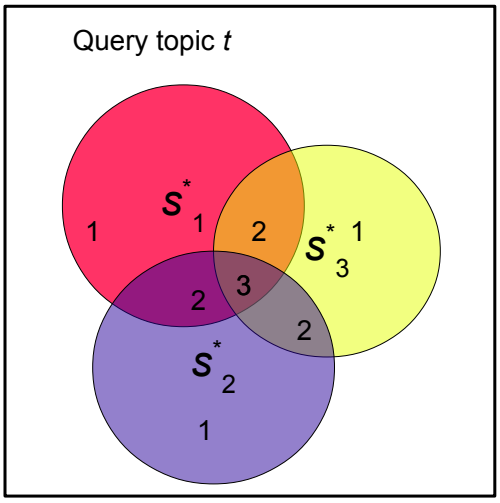

Fig. 2. Inclusion-exclusion principle. The sets represent candidate items s for a query, and the area covered by each set is the "information" covered by that item for query topic t. Numbers on different areas indicates the number of sets that share these areas.

This result has a natural interpretation: the first summation term determines the coverage of topic t by each document s_i ($1 \leq i \leq k-1$) currently in the result set, the second double summation term corrects the first term by removing the joint probability mass from all pairs of documents that was double counted, and so on according to the principle of inclusion-exclusion. (11) not only provides a probabilistic set covering view of Exp-1-call@k, but it also suggests that a portfolio approach to diversity using only the first summation would overcount each document's contribution to the diversity metric according to this set covering perspective.

The inclusion-exclusion principle calculation provided by the second term in Equation 11 is illustrated in Figure 2. In words, this term is calculating the total topic probability coverage of t by all selected items $\{s_1^*, \ldots, s_{k-1}^*\}$ by properly applying the inclusion-exclusion principle to ensure that overlapping probability coverage is not double counted. Then referring back to Equation 5, we note that s_k is chosen by maximizing a weighted sum over topics, where each topic weight is determined by its relevance to the query q, the item s_k, and penalized (i.e., due to the $1-$) by the topic coverage of t by the set $\{s_1^*, \ldots, s_{k-1}^*\}$ to naturally encourage diversity. We note that this is a soft probabilistic version of the "in or out" topic coverage approach of WSL.

Subtopic Relevance Models: We use a subtopic relevance model that is a simplified version of the model in [Guo and Sanner 2010] with fewer dependence assumptions. In other work, Zhai *et al* [Zhai et al. 2003] present an empirical risk minimization view of dependent document retrieval from a subtopic perspective, where they derive a formalization of the *greedy* selection step that is similar to MMR and to a lesser extent, Exp-1-call@k.

Set-based Relevance Objectives: Chen and Karger [Chen and Karger 2006], whose derivation we extended, directly optimize 1-call@k, but their intention is not to formal-

ize MMR and instead use naïve Bayes to directly evaluate (16). Agrawal et al [Agrawal et al. 2009] and Santos et al (xQuad) [Santos et al. 2010a] both specify set-based diversity metrics *very* similar to Exp-1-call@k but do not provide formal derivations as we have done in this work.

Ranking Based Objectives: Finally, returning to our introductory motivation, Wang and Zhu [Wang and Zhu 2010] have shown that natural forms of result set diversification arise via the optimization of average precision [Buckley and Voorhees 2000] and reciprocal rank [Voorhees 1999]. Both of these methods share the view of directly optimizing a *ranking-based* objective, whereas this paper proposes a novel derivation from the alternate view of optimizing a *set-based* objective w.r.t. a subtopic model of relevance. However, even though Exp-1-call@k is a set-based objective, an indirect consequence of (and motivation for) greedily optimizing it is that documents added earlier yield a greater increase in objective than those added later; this yields a natural rank ordering on the greedy Exp-1-call@k result set.

5. CONCLUSION

In this paper, we presented a new derivation of diverse retrieval by directly optimizing the expected n-call@k set-based retrieval objective w.r.t. a latent subtopic model of binary relevance. This result both motivates and contrasts with various related diversification approaches, providing a new theoretical basis for the investigation of diverse retrieval. Empirical results on three real-world data sets reflected our theoretical results.

Future work includes the study of efficiently optimizing the n-call@k objecive function by possibly exploiting the submodularity [Borodin et al. 2012]. Additional, it is also important to study the temporal diveristy as suggested in [Lathia et al. 2010] and [Zhao et al. 2012].

APPENDIX

A. FULL DERIVATION

A.1. Optimizing objective

We want to choose S_k^* that maximizes the objective:

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

By taking a greedy approach, we select s_k^* given S_{k-1}^*:

$$
\begin{aligned}
s_k^* &= \arg\max_{s_k} \mathbb{E}[R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}] \\
&= \arg\max_{s_k} P(R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}) && (12) \\
&= \arg\max_{s_k} \sum_{T_k} \Big(P(t|\mathbf{q})\, P(t_k|s_k) \Big(\prod_{i=1}^{k-1} P(t_i|s_i^*) \Big) \cdot P(R_k \geq n | T_k, S_{k-1}^*, s_k, \mathbf{q}) \Big) && (13) \\
&= \arg\max_{s_k} \sum_{T_k} P(t|\mathbf{q})\, P(t_k|s_k) \Big(\prod_{i=1}^{k-1} P(t_i|s_i^*) \Big) \cdot \Big(\underbrace{P(r_k \geq 0 | R_{k-1} \geq n, t_k, t)}_{1} P(R_{k-1} \geq n | T_{k\text{-}1}) \\
&\qquad + P(r_k = 1 | R_{k-1} = n-1, t_k, t) P(R_{k-1} = n-1 | T_{k\text{-}1}) \Big) && (14) \\
&= \arg\max_{s_k} \Bigg(\sum_{T_{k\text{-}1}} \underbrace{\Big[\sum_{t_k} P(t_k|s_k) \Big]}_{1} P(t|\mathbf{q}) \Big(\prod_{i=1}^{k-1} P(t_i|s_i^*) \Big) P(R_{k-1} \geq n | T_{k\text{-}1}) + \\
&\qquad \sum_{T_{k\text{-}1}} \Big[\sum_{t_k} P(t_k|s_k) P(r_k = 1 | t_k, t) \Big] P(t|\mathbf{q}) \Big(\prod_{i=1}^{k-1} P(t_i|s_i^*) \Big) P(R_{k-1} = n-1 | T_{k\text{-}1}) \Bigg) \\
&= \arg\max_{s_k} \sum_{t} P(t|\mathbf{q}) P(t_k = t | s_k) \Big[\sum_{t_1, \ldots, t_{k-1}} P(R_{k-1} = n-1 | T_{k\text{-}1}) \prod_{i=1}^{k-1} P(t_i|s_i^*) \Big] && (15) \\
&= \arg\max_{s_k} \sum_{t} P(t|\mathbf{q}) P(t_k = t | s_k) P(R_{k-1} = n-1 | S_{k-1}^*) && (16)
\end{aligned}
$$

Note:
(1) Since $(R_k \geq n)$ can only be zero or one in probability.
(2) Marginalize out T_k.
(3) Split $(R_k \geq n)$ into two disjoint events $(r_k \geq 0, R_{k-1} \geq n)$, $(r_k = 1, R_{k-1} = n-1)$, conditioned on R_{k-1}.
(4) Drop the first line as it does not involve s_k and has no influence in determining s_k^*. Note that $\sum_{t_k} P(t_k|s_k) P(r_k = 1 | t_k, t) = \sum_{t_k} P(t_k|s_k) \mathbb{I}[t_k = t] = P(t_k = t | s_k)$, where t is implicitly conditioned and is not explicitly shown here.
(5) This objective is recursively defined.

By similar reasoning, the probability needed in (16) is recursively defined as

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

For expected n-call@k where $n \leq k/2$, by unrolling its recursive definition in (16), the explicit objective is

$$s_k^* = \arg\max_{s_k} \sum_t \left(P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j_1,\ldots,j_{n-1}} \prod_{l \in \{j_1,\ldots,j_{n-1}\}} P(t_l = t|s_l^*) \prod_{\substack{i=1 \\ i \notin \{j_1,\ldots,j_{n-1}\}}}^{k-1} \left(1 - P(t_i = t|s_i^*)\right) \right) \quad (17)$$

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).

Similarly, for expected n-call@k where $n > k/2$, the explicit objective is

$$s_k^* = \arg\max_{s_k} \sum_t \left(P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j_n,\ldots,j_{k-1}} \prod_{l \in \{j_n,\ldots,j_{k-1}\}} \left(1 - P(t_l = t|s_l^*)\right) \prod_{\substack{i=1 \\ i \notin \{j_n,\ldots,j_{k-1}\}}}^{k-1} P(t_i = t|s_i^*) \right) \quad (18)$$

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

A.2. Relation to MMR: expected n-call@k when $n > k/2$

Assuming that $\forall i\ P(t_i|s_i) \in \{0, 1\}$ and $P(t|\mathbf{q}) \in \{0, 1\}$. It is possible to write

$$\prod_{l \in \{j_n,\ldots,j_{k-1}\}} \left(1 - P(t_l = t|s_l^*)\right) = 1 - \left(1 - \prod_{l \in \{j_n,\ldots,j_{k-1}\}} \left(1 - P(t_l = t|s_l^*)\right)\right) = 1 - \left(\max_{l \in \{j_n,\ldots,j_{k-1}\}} P(t_l = t|s_l^*)\right)$$

This allows us to rewrite (18)

$$\begin{aligned} s_k^* = \arg\max_{s_k} \sum_t \Bigg(& P(t|\mathbf{q}) P(t_k = t|s_k) \sum_{j_n,\ldots,j_{k-1}} \prod_{\substack{i=1 \\ i \notin \{j_n,\ldots,j_{k-1}\}}}^{k-1} P(t_i = t|s_i^*) \\ & - P(t|\mathbf{q}) P(t_k = t|s_k) \sum_{j_n,\ldots,j_{k-1}} \prod_{\substack{i=1 \\ i \notin \{j_n,\ldots,j_{k-1}\}}}^{k-1} P(t_i = t|s_i^*) \max_{l \in \{j_n,\ldots,j_{k-1}\}} P(t_l = t|s_l^*) \Bigg) \end{aligned} \quad (19)$$

Assuming m relevant documents are already selected in the $k-1$ collection, then the top term (specifically $\prod_i$) is non-zero $\binom{m}{n-1}$ times. For the bottom term, it takes $n-1$ relevant documents to satisfy its $\prod_i$, and one additional relevant document to satisfy the $\max_l$ 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), (19) becomes

$$s_k^* = \arg\max_{s_k} \left[\sum_t P(t|\mathbf{q})P(t_k=t|s_k)\binom{m}{n-1}\right] - \left[\sum_t P(t|\mathbf{q})P(t_k=t|s_k)\binom{m}{n}\underbrace{\max_{s_i\in S_{k-1}^*} P(t_i=t|s_i)}_{1}\right]$$

$$= \arg\max_{s_k} \binom{m}{n-1}\underbrace{\sum_t P(t|\mathbf{q})P(t_k=t|s_k)}_{\text{relevance: } \mathrm{Sim}_1(s_k,\mathbf{q})} - \binom{m}{n}\max_{s_i\in S_{k-1}^*}\underbrace{\sum_t P(t_i=t|s_i)P(t|\mathbf{q})P(t_k=t|s_k)}_{\text{diversity: } \mathrm{Sim}_2(s_k,s_i,\mathbf{q})} \quad (20)$$

$$= \arg\max_{s_k} \frac{n}{m+1}\mathrm{Sim}_1(s_k,\mathbf{q}) - \frac{m-n+1}{m+1}\max_{s_i\in S_{k-1}^*}\mathrm{Sim}_2(s_k,s_i,\mathbf{q}) \quad (21)$$

Note:
(9) We can rearrange "$\sum_t P(t|\mathbf{q})\max_{s_i}\cdots$" as "$\max_{s_i}\sum_t P(t|\mathbf{q})\cdots$" since the $\sum_t P(t|\mathbf{q})$ 'selects' the only t for which $P(t|\mathbf{q})=1$.
(10) Normalize by dividing the equation by $\binom{m}{n-1}+\binom{m}{n}=\binom{m+1}{n}$ (Pascal's rule).
The result is the same as the case where $n \leq k/2$.

The reason that we do not remove the max term in (20) is that this allows us to compare the objective with MMR directly. Also, leaving the max term suggests an approximate form for the case where the subtopic probabilities are non-deterministic (not strictly 0 or 1), and approaches (20) as the probabilities become more deterministic.

In practice, under the greedy approach of the expected n-call@k in selecting S_k^*, we expect that there are already n relevant documents chosen in the set $S_{k-1}^* = \{s_1^*,\ldots,s_{k-1}^*\}$ (where $n << k$). In expectation, $m = n$ and hence the optimizing objective can be thought to be

$$s_k^* = \arg\max_{s_k} \frac{n}{n+1}\mathrm{Sim}_1(s_k,\mathbf{q}) - \frac{1}{n+1}\max_{s_i\in S_{k-1}^*}\mathrm{Sim}_2(s_k,s_i,\mathbf{q}) \quad (22)$$

From (22), it is simple to see that the diversification level decreases with n.

B. ADDITIONAL DERIVATION

B.1. Alternative derivation for expected 2-call@k

$$
\begin{aligned}
s_k^* &= \arg\max_{s_k} \mathbb{E}\big[R_k \geq 2 \,\big|\, S_{k-1}^*, s_k, \mathbf{q}\big] \\
&= \arg\max_{s_k} \mathbb{E}\Big[(r_1 = 1 \wedge r_2 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 1) \vee (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 1) \vee \\
&\qquad (r_1 = 0 \wedge r_2 = 0 \wedge r_3 = 1 \wedge r_4 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 0 \wedge r_4 = 1) \vee \\
&\qquad (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 0 \wedge r_4 = 1) \vee \cdots \vee \\
&\qquad (r_1 = 0 \wedge \cdots \wedge r_{k-2} = 0 \wedge r_{k-1} = 1 \wedge r_k = 1) \vee \\
&\qquad (r_1 = 0 \wedge \cdots \wedge r_{k-3} = 0 \wedge r_{k-2} = 1 \wedge r_{k-1} = 0 \wedge r_k = 1) \vee \cdots \vee \\
&\qquad (r_1 = 1 \wedge r_2 = 0 \wedge \cdots \wedge r_{k-1} = 0 \wedge r_k = 1) \Big| S_{k-1}^*, s_k, \mathbf{q} \Big] \\
&= \arg\max_{s_k} \mathbb{E}\Big[(r_1 = 1 \wedge r_2 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 1) \vee (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 1) \vee \\
&\qquad (r_1 = 0 \wedge r_2 = 0 \wedge r_3 = 1 \wedge r_4 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 0 \wedge r_4 = 1) \vee \\
&\qquad (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 0 \wedge r_4 = 1) \vee \cdots \vee \\
&\qquad \bigvee_{j=1}^{k-1} \left(r_k = 1 \wedge \bigwedge_{\substack{i=1 \\ i \neq j}}^{k-1} r_i = 0 \wedge r_j = 1 \right) \Big| S_{k-1}^*, s_k, \mathbf{q} \Big] \\
&= \arg\max_{s_k} \sum_{j=1}^{k-1} P\left(r_k = 1 \wedge \bigwedge_{\substack{i=1 \\ i \neq j}}^{k-1} r_i = 0 \wedge r_j = 1 \,\middle|\, S_{k-1}^*, s_k, \mathbf{q} \right) \\
&= \arg\max_{s_k} \sum_{j=1}^{k-1} \left(\sum_{t_1, \ldots, t_k, t} P(t|\mathbf{q})\, P(t_k|s_k) \mathbb{I}[t_k = t]\, P(t_j|s_j^*) \mathbb{I}[t_j = t] \prod_{\substack{i=1 \\ i \neq j}}^{k-1} P(t_i|s_i^*) \mathbb{I}[t_i \neq t] \right) \\
&= \arg\max_{s_k} \sum_{t} P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} \left(P(t_j = t|s_j^*) \prod_{\substack{i=1 \\ i \neq j}}^{k-1} (1 - P(t_i = t|s_i^*)) \right)
\end{aligned}
$$

Assuming that $\forall i\ P(t_i|s_i) \in \{0,1\}$ and $P(t|\mathbf{q}) \in \{0,1\}$, the objective becomes:

$$
\begin{aligned}
s_k^* &= \arg\max_{s_k} \sum_t P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} \left(P(t_j = t|s_j^*) \prod_{\substack{i=1 \\ i\neq j}}^{k-1} (1 - P(t_i = t|s_i^*)) \right) \\
&= \arg\max_{s_k} \sum_t P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} \left(P(t_j = t|s_j^*) \left[1 - \left(1 - \prod_{\substack{i=1 \\ i\neq j}}^{k-1} (1 - P(t_i = t|s_i^*)) \right) \right] \right) \\
&= \arg\max_{s_k} \sum_t P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} \left[P\big(t_j = t|s_j^*\big) - P\big(t_j = t|s_j^*\big) \left(1 - \prod_{\substack{i=1 \\ i\neq j}}^{k-1} (1 - P(t_i = t|s_i^*)) \right) \right] \\
&= \arg\max_{s_k} \sum_t P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} P(t_j = t|s_j^*) \\
&\qquad - \sum_t P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} P(t_j = t|s_j^*) \max_{\substack{i\in[1,k-1] \\ i\neq j}} P(t_i = t|s_i^*)
\end{aligned}
$$

Noting that this is of the same form as (19), albeit much simpler.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

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

REFERENCES

AGRAWAL, R., GOLLAPUDI, S., HALVERSON, A., AND IEONG, S. 2009. Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, USA, 5–14.

ANAGNOSTOPOULOS, A., BRODER, A. Z., AND CARMEL, D. 2005. Sampling search-engine results. In *Proceedings of the 14th International Conference on World Wide Web*. ACM, Chiba, Japan, 245–256.

BAI, J. AND NIE, J.-Y. 2008. Adapting information retrieval to query contexts. *Information Processing and Management 44,* 6, 1901–1922.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

BORODIN, A., LEE, H. C., AND YE, Y. 2012. Max-sum diversification, monotone submodular functions and dynamic updates. In *Proceedings of the 31st symposium on Principles of Database Systems*. PODS '12. ACM, New York, NY, USA, 155–166.

BUCKLEY, C. AND VOORHEES, E. M. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '00. ACM, New York, NY, USA, 33–40.

CARBONELL, J. AND GOLDSTEIN, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 335–336.

CHAPELLE, O., METLZER, D., ZHANG, Y., AND GRINSPAN, P. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. CIKM '09. ACM, New York, NY, USA, 621–630.

CHEN, H. AND KARGER, D. R. 2006. Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 429–436.

CLARKE, C. L. A., KOLLA, M., CORMACK, G. V., AND VECHTOMOVA, O. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 659–666.

CLOUGH, P., SANDERSON, M., ABOUAMMOH, M., NAVARRO, S., AND PARAMITA, M. 2009. Multiple approaches to analysing query diversity. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 734–735.

DANG, V. AND CROFT, W. B. 2012. Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '12. ACM, New York, NY, USA, 65–74.

DEERWESTER, S., DUMAISAND, S. T., FURNAS, G. W., LANDAUER, T. K., AND HARSHMAN, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*, 391–407.

GOFFMAN, W. 1964. On relevance as a measure. *Information Storage and Retrieval 2,* 3, 201–203.

GOLLAPUDI, S. AND SHARMA, A. 2009. An axiomatic approach for result diversification. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 381–390.

GUO, S. AND SANNER, S. 2010. Probabilistic latent maximal marginal relevance. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Geneva, Switzerland.

HARMON, G. 2008. Remembering William Goffman: Mathematical information science pioneer. *Information Processing & Management 44,* 4, 1634–1647.

JEBARA, T., KONDOR, R., AND HOWARD, A. 2004. Probability product kernels. *Journal of Machine Learning Research 5*, 819–844.

LATHIA, N., HAILES, S., CAPRA, L., AND AMATRIAIN, X. 2010. Temporal diversity in recommender systems. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 210–217.

RADLINSKI, F. AND DUMAIS, S. 2006. Improving personalized web search using result diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 691–692.

RADLINSKI, F., KLEINBERG, R., AND JOACHIMS, T. 2008. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, 784–791.

ROBERTSON, S. E. AND WALKER, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '94. Springer-Verlag New York, Inc., New York, NY, USA, 232–241.

SALTON, G. AND MCGILL, M. 1983. *Introduction to modern information retrieval*. McGraw-Hill.

SANDERSON, M. 2008. Ambiguous queries: Test collections need more sense. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 499–506.

SANNER, S., GUO, S., GRAEPEL, T., KHARAZMI, S., AND KARIMI, S. 2011. Diverse retrieval via greedy optimization of expected 1-call@k in a latent subtopic relevance model. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, New York, NY, USA, 1977–1980.

SANTOS, R. L., MACDONALD, C., AND OUNIS, I. 2010a. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*. WWW '10. ACM, New York, NY, USA, 881–890.

SANTOS, R. L., MACDONALD, C., AND OUNIS, I. 2010b. Selectively diversifying web search results. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, New York, NY, USA, 1179–1188.

SONG, R., LUO, Z., NIE, J.-Y., YU, Y., AND HON, H.-W. 2009. Identification of ambiguous queries in web search. *Information Processing and Management 45,* 2, 216–229.

VARGAS, S., CASTELLS, P., AND VALLET, D. 2012. Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '12. ACM, New York, NY, USA, 75–84.

VOORHEES, E. M. 1999. TREC-8 question answering track report. In *Proceedings of the 8th Text Retrieval Conference*. 77–82.

WANG, J. AND ZHU, J. 2009. Portfolio theory of information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Boston, Massachusetts, USA, 115–122.

WANG, J. AND ZHU, J. 2010. On statistical analysis and optimization of information retrieval effectiveness metrics. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '10. ACM, New York, NY, USA, 226–233.

YU, C., LAKSHMANAN, L., AND AMER-YAHIA, S. 2009. It takes variety to make a world: diversification in recommender systems. In *Proceedings of the 12th International Conference on Extending Database Technology*. Vol. 360. ACM, 368–378.

YUE, Y. AND JOACHIMS, T. 2008. Predicting diverse subsets using structural SVMs. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, New York, NY, USA, 1224–1231.

ZHAI, C., COHEN, W. W., AND LAFFERTY, J. 2003. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development In Informaion Retrieval*. ACM, 10–17.

ZHAO, G., LEE, M. L., HSU, W., AND CHEN, W. 2012. Increasing temporal diversity with purchase intervals. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '12. ACM, New York, NY, USA, 165–174.