

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

Shengbo Guo, Xerox Research Centre Europe
 Kar Wai Lim, Australian National University and NICTA
 Scott Sanner, NICTA and Australian National University

In this article, we explore the relationship between result set diversification and the optimization of n -call@ k – a set-based relevance objective that is 1 if at least n documents in a set of k are relevant, otherwise 0. First, we formally quantify the mathematical relationship between diversity and greedy optimization of the *expected* n -call@ k objective in a latent subtopic model of binary relevance. Second, we relate this result to a variety of other diversification approaches proposed in the literature, including deep connections with maximal marginal relevance. The contributions of this work are threefold: (1) in contrast to a variety of diverse retrieval algorithms derived from alternate rank-based relevance criteria such as average precision and reciprocal rank, we provide a complementary theoretical perspective on the emergence of diversity via optimization of n -call@ k in a latent subtopic model of relevance intended to model both ambiguous and faceted subtopic retrieval, (2) we precisely formalize a mathematical relationship between n -call@ k and diversity that confirms empirical observations in the literature, and (3) we provide a theoretical underpinning and comparison to many other (ad-hoc) diversification approaches in the literature.

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. On the Relationship between Expected n -call@ k and Diversity. ACM Trans. Embedd. Comput. Syst. 9, 4, Article 39 (March 2010), 15 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] — has often been noted as 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.

If one wants to optimize a result set to cover all possible query subtopics, the question naturally arises as to what set-level relevance objective should be optimized?

Author’s addresses: S. Guo, Xerox Research Centre Europe; K. W. Lim, Australian National University and NICTA; S. Sanner, NICTA and Australian National University.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2010 ACM 1539-9087/2010/03-ART39 \$15.00

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

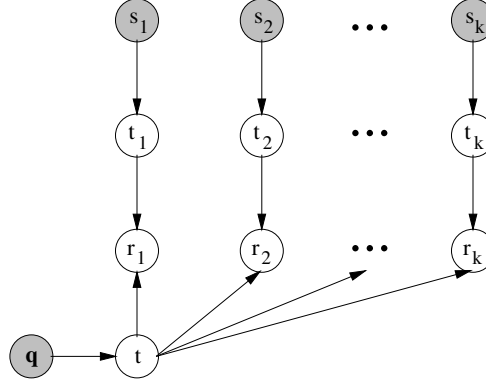


Fig. 1. Latent subtopic binary relevance model.

Wang and Zhu [Wang and Zhu 2010] have shown that natural forms of diversification arise via the optimization of *average precision* [Buckley and Voorhees 2000] and *reciprocal rank* [Voorhees 1999]. While these results directly motivate diverse retrieval via *rank-based* (ordered set) relevance criteria, they do not use the subtopic motivation for diversity. We use this alternate subtopic motivation in this article, where we define binary relevance via a *latent subtopic model* (shown in Figure 1 and formally defined in Section 2). With this definition of relevance, we then optimize the expectation of the *n-call@k set-based* relevance criteria that takes the value 1 if at least n of k documents in a result set are relevant and 0 otherwise [Chen and Karger 2006].

Mathematically, it turns out that the optimization of expected *n-call@k* encourages more diversity as $n \rightarrow 1$, which reflects previous empirical observations in the literature [Wang and Zhu 2009] (Figure 2c). However, it turns out there are also deep connections between this derivation and one of the most popular diversification algorithms in the literature known as maximal marginal relevance (MMR) [Carbonell and Goldstein 1998].

Formally, MMR takes an *item set* D (e.g., a set of documents) where retrieved items are denoted as $s_i \in D$, and aims 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 $k-1$ optimal selections $S_{k-1}^* = \{s_1^*, \dots, s_{k-1}^*\}$ (recursively defining $S_k^* = S_{k-1}^* \cup \{s_k^*\}$ with $S_0^* = \emptyset$) as follows:

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

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]. The derivation presented in this article formally demonstrates that greedy optimization of expected *n-call@k* precisely corresponds to $\lambda = \frac{n}{n+1}$ in MMR. In addition, we also note deep mathematical connections between the optimization of *n-call@k* and a variety of other (often ad-hoc) diversification approaches proposed in the literature.

This article extends our previous works [Guo and Sanner 2010], [Sanner et al. 2011], and [Lim et al. 2012].

2. RELEVANCE MODEL AND OBJECTIVE

In this article, we motivate diversity through the task of subtopic retrieval [Zhai et al. 2003]. Subtopics may differ according to a variety of *information needs* [Radlinski et al. 2009] such as query facets (e.g., different political party views in a set of blogs) or different ambiguous query interpretations (e.g., *Java* the programming language vs. *Java* the island).

We propose a directed graphical model (i.e., a Bayesian network) in Figure 1 to formalize the independence assumptions in a probabilistic subtopic model of binary relevance. In Figure 1, shaded nodes represent observed variables, whereas unshaded nodes are latent. The observed variables are the query terms 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 finite and discrete subtopic set. Then $t_i \in T$ represent subtopics for respective s_i and $t \in T$ represents a subtopic for query q . The r_i are binary variables that indicate if respective selected items s_i are relevant ($r_i = 1$).

The conditional probability tables (CPTs) associated with each node in Figure 1 are defined as follows: $P(t_i|s_i)$ and $P(t|q)$ respectively represent the subtopic distribution for item s_i and query q . The remaining CPTs are for the relevance variables r_i ; 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.

It is a quite natural at this point to ask (a) how the subtopic set T can be chosen automatically, and (b) how the subtopic CPTs $P(t|q)$ and $P(t_i|s_i)$ are specified or learned for the subtopics $t \in T$. However, we defer this question until Section 3.3 where it will become apparent via the mathematical connections to MMR and the “kernel trick” how the necessary quantities required for optimization can be specified via similarity functions commonly used with MMR. We note this is simply one automated option though — a user is otherwise free to define subtopics and specify or learn their probabilities in any way deemed appropriate — the only required property is that T is finite and all subtopic CPTs are proper conditional probabilities that sum to 1 over $t \in T$.

Given the specification of our latent subtopic binary relevance model, we now formally define the expectation of n -call@ k w.r.t. this model. To facilitate this, we first define $R_k = \sum_{i=1}^k r_i$, where R_k is the number of relevant items from the first k selections. Interpreting $R_k \geq n$ as an indicator random variable $\mathbb{I}[R_k \geq n]$, we can then express the *expected n -call@ k* objective simply as

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

With our relevance model and objective now formally defined, we present our main theoretical results in the following section.

3. THEORETICAL RESULTS AND COMPARISON TO MMR

3.1. Optimizing Expected 1-call@ k

Before we present the result on optimizing expected n -call@ k for general n , we start off from the simplest case, expected 1-call@ k (where $n = 1$):

$$\text{Exp-1-Call@}k(S_k, q) = \mathbb{E}[R_k \geq 1 | s_1, \dots, s_k, q] \quad (3)$$

Since $R_k \geq 1$ is satisfied as long as any one of the r_i is 1, the objective can be rewritten in terms of logical notation:

$$\text{Exp-1-Call@}k(S_k, \mathbf{q}) = \mathbb{E} \left[\bigvee_{i=1}^k r_i = 1 \mid s_1, \dots, s_k, \mathbf{q} \right] \quad (4)$$

Since jointly optimizing $\text{Exp-1-Call@}k(S_k, \mathbf{q})$ 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:

$$\begin{aligned} s_k^* &= \arg \max_{s_k} \text{Exp-1-Call@}k(S_{k-1}^* \cup \{s_k\}, \mathbf{q}) \\ &= \arg \max_{s_k} \mathbb{E} \left[\bigvee_{i=1}^k r_i = 1 \mid S_{k-1}^*, s_k, \mathbf{q} \right] \end{aligned}$$

Next we can apply a logical equivalence and exploit the additivity of mutually exclusive events to split $\bigvee_{i=1}^k r_i = 1$ into mutually exclusive disjoint subsets:

$$\begin{aligned} s_k^* &= \arg \max_{s_k} \mathbb{E} \left[(r_1 = 1) \vee (r_2 = 1 \wedge r_1 = 0) \vee (r_3 = 1 \wedge r_2 = 0 \wedge r_1 = 0) \vee \dots \vee \right. \\ &\quad \left. \left(r_k = 1 \wedge \bigwedge_{i=1}^{k-1} r_i = 0 \right) \mid S_{k-1}^*, s_k, \mathbf{q} \right] \quad (5) \end{aligned}$$

Since these events are binary and disjoint, we can rewrite the expectation as probability. This gives us the sum of the probabilities of each individual event. We further simplify it by grouping all $r_j = 0$.¹ We factorize each joint probability into a conditional and prior, and remove terms and factors that are do not contain s_k , noting that these terms only act as constants when optimizing for s_k :

$$\begin{aligned} s_k^* &= \arg \max_{s_k} \sum_{i=1}^k P(r_i = 1, \{r_j = 0\}_{j < i} \mid \{s_j^*\}_{j \leq i, j < k}, \{s_k\}_{k=i}, \mathbf{q}) \\ &= \arg \max_{s_k} \sum_{i=1}^k P(r_i = 1 \mid \{r_j = 0\}_{j < i}, \{s_j^*\}_{j \leq i, j < k}, \{s_k\}_{k=i}, \mathbf{q}) \\ &\quad P(\{r_j = 0\}_{j < i} \mid \{s_j^*\}_{j \leq i, j < k}, \mathbf{q}) \\ &= \arg \max_{s_k} P(r_k = 1 \mid \{r_j = 0\}_{j < k}, S_{k-1}^*, s_k, \mathbf{q}) \quad (6) \end{aligned}$$

From (6), to optimize for s_k , we need only to maximize s_k 's probability of relevance conditioned on the previous selections (which are assumed irrelevant, $r_j = 0$) and the query.

¹The 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 (6) w.r.t. our actual graphical model of subtopic relevance from Figure 1:

$$\begin{aligned}
s_k^* &= \arg \max_{s_k} P(r_k = 1 \mid \{r_j = 0\}_{j < k}, S_{k-1}^*, s_k, \mathbf{q}) \\
&= \arg \max_{s_k} \sum_{t, t_1, \dots, t_k} P(t \mid \mathbf{q}) P(t_k \mid s_k) \mathbb{I}[t_k = t] \prod_{i=1}^{k-1} P(t_i \mid s_i^*) \mathbb{I}[t_i \neq t] \\
&= \arg \max_{s_k} \sum_t P(t \mid \mathbf{q}) \sum_{t_k} P(t_k \mid s_k) \mathbb{I}[t_k = t] \prod_{i=1}^{k-1} \sum_{t_i} P(t_i \mid s_i^*) \mathbb{I}[t_i \neq t] \\
&= \arg \max_{s_k} \sum_t P(t \mid \mathbf{q}) P(t_k = t \mid s_k) \left[\prod_{i=1}^{k-1} (1 - P(t_i = t \mid s_i^*)) \right] \tag{7}
\end{aligned}$$

Here we have used the following equality:

$$\begin{aligned}
\sum_{t_i} P(t_i \mid s_i) \mathbb{I}[t_i = t] &= P(t_i = t \mid s_i) \\
\sum_{t_i} P(t_i \mid s_i) \mathbb{I}[t_i \neq t] &= 1 - P(t_i = t \mid s_i)
\end{aligned}$$

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

$$\begin{aligned}
s_k^* &= \arg \max_{s_k} \sum_t P(t \mid \mathbf{q}) P(t_k = t \mid s_k) \left(1 - \tilde{P}(t \mid S_{k-1}^*) \right) \\
&= \arg \max_{s_k} \underbrace{\sum_t P(t \mid \mathbf{q}) P(t_k = t \mid s_k)}_{\text{query similarity}} - \underbrace{\sum_t P(t \mid \mathbf{q}) P(t_k = t \mid s_k) \tilde{P}(t \mid S_{k-1}^*)}_{\text{query-reweighted diversity}}. \tag{8}
\end{aligned}$$

From (8), we can see that optimizing expected 1-call@k give us a greedy algorithm that somewhat resembles MMR in (1) with an implicit equal balancing of the similarity and diversity objectives (i.e., $\lambda = \frac{1}{2}$). After presenting our general derivation for expected n -call@k in the next subsection for $n \geq 1$, we return to a general comparison between expected n -call@k and MMR.

3.2. Optimizing Expected n-call@k

Having analyzed the special case of n -call@k for $n = 1$, we now attempt to generalize our approach to general $n \geq 1$. In contrast to $n = 1$, however, we note that it is not straightforward to use the logical notation of the expected n -call@k objective for $n = 1$ given in (4) when $n > 1$: even for $n = 2$, the number of disjoint events that satisfy $R_k \geq 2$ grows by a factor of $(k-1)/2$.²

Next we provide our main result on optimizing expected n -call@k for general $n \geq 1$. For the interested reader, we note that a simpler (logic-based) derivation for the special case of $n = 2$ is provided in Appendix A, which provides some guiding intuitions for the general derivation developed in this section.

In the following derivation, we adopt a more abstract approach (working with R_k directly) while utilizing the same principle as in expected 1-call@k. As done previously,

² $\binom{k}{2} / \binom{k}{1} = (k-1)/2$

we take a greedy approach that selects s_k assuming that S_{k-1}^* is already chosen:

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

Here, we have exploited the binary (0, 1) nature of $R_k \geq n$ to rewrite the objective 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 (define $T_k = \{t, t_1, \dots, t_k\}$ and $\sum_{T_k} \circ = \sum_t \sum_{t_1} \dots \sum_{t_k} \circ$):

$$s_k^* = \arg \max_{s_k} \sum_{T_k} \left(P(t|\mathbf{q}) P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s_i^*) \cdot P(R_k \geq n | T_k, S_{k-1}^*, s_k, \mathbf{q}) \right)$$

We split $R_k \geq n$ into two disjoint (additive) events ($r_k \geq 0, R_{k-1} \geq n$), ($r_k=1, R_{k-1}=n-1$) based on R_{k-1} . (If R_{k-1} is equal to $n-1$, r_k must be 1; if R_{k-1} is greater or equal to n , then r_k can be either 0 or 1).

$$\begin{aligned} s_k^* &= \arg \max_{s_k} \sum_{T_k} P(t|\mathbf{q}) P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s_i^*) \cdot \left(P(r_k \geq 0, R_{k-1} \geq n | T_k, S_{k-1}^*, s_k, \mathbf{q}) \right. \\ &\quad \left. + P(r_k = 1, R_{k-1} = n-1 | T_k, S_{k-1}^*, s_k, \mathbf{q}) \right) \end{aligned}$$

We then write the joint probability into a conditioned and prior, conditioned on R_k :

$$\begin{aligned} s_k^* &= \arg \max_{s_k} \sum_{T_k} P(t|\mathbf{q}) P(t_k|s_k) \prod_{i=1}^{k-1} P(t_i|s_i^*) \cdot \left(\underbrace{P(r_k \geq 0 | R_{k-1} \geq n, t_k, t)}_1 P(R_{k-1} \geq n | T_{k-1}) \right. \\ &\quad \left. + P(r_k = 1 | R_{k-1} = n-1, t_k, t) P(R_{k-1} = n-1 | T_{k-1}) \right) \end{aligned}$$

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

$$\begin{aligned} s_k^* &= \arg \max_{s_k} \left(\sum_{T_{k-1}} \underbrace{\left[\sum_{t_k} P(t_k|s_k) \right]}_1 P(R_{k-1} \geq n | T_{k-1}) P(t|\mathbf{q}) \prod_{i=1}^{k-1} P(t_i|s_i^*) \right. \\ &\quad \left. + \sum_t P(t|\mathbf{q}) P(t_k=t|s_k) \sum_{t_1, \dots, t_{k-1}} P(R_{k-1} = n-1 | T_{k-1}) \prod_{i=1}^{k-1} P(t_i|s_i^*) \right) \end{aligned}$$

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^*):

$$s_k^* = \arg \max_{s_k} \sum_t P(t|\mathbf{q}) P(t_k=t|s_k) P(R_{k-1} = n-1 | S_{k-1}^*, t) \quad (9)$$

By similar reasoning, we can derive that the last probability needed in (9) is recursively defined as

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

We can now rewrite (9) 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

$$s_k^* = \arg \max_{s_k} \sum_t \left(P(t | \mathbf{q}) P(t_k = t | s_k) \cdot \sum_{j_1, \dots, j_{n-1}} \prod_{l \in \{j_1, \dots, j_{n-1}\}} P(t_l = t | s_l^*) \prod_{\substack{i=1 \\ i \notin \{j_1, \dots, j_{n-1}\}}}^{k-1} (1 - P(t_i = t | s_i^*)) \right) \quad (10)$$

where $j_1, \dots, j_{n-1} \in \{1, \dots, k-1\}$ satisfy that $j_i < j_{i+1}$ (i.e., an ordered permutation of $n-1$ result set indices). Note that (10) reduces to (7) when $n = 1$.

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

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

and by substituting this into (10) and distributing, we get

$$s_k^* = \arg \max_{s_k} \sum_t \left(P(t | \mathbf{q}) P(t_k = t | s_k) \sum_{j_1, \dots, j_{n-1}} \prod_{l \in \{j_1, \dots, j_{n-1}\}} P(t_l = t | s_l^*) \right. \\ \left. - P(t | \mathbf{q}) P(t_k = t | s_k) \sum_{j_1, \dots, j_{n-1}} \prod_{l \in \{j_1, \dots, j_{n-1}\}} P(t_l = t | s_l^*) \max_{\substack{i \in [1, k-1] \\ i \notin \{j_1, \dots, j_{n-1}\}}} P(t_i = t | s_i^*) \right). \quad (11)$$

Assuming m of the 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

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

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

$$s_k^* = \arg \max_{s_k} \frac{n}{m+1} \text{Sim}_1(s_k, \mathbf{q}) - \frac{m-n+1}{m+1} \max_{s_i \in S_{k-1}^*} \text{Sim}_2(s_k, s_i, \mathbf{q}) \quad (12)$$

Fortuitously, we note that the $\binom{m+1}{n}$ divisor cancelled with the numerators, yielding this elegant and interpretable result. In practice, we can make one additional simplifying assumption: under the greedy approach of selecting s_k given S_{k-1}^* when $k \geq n+1$, we expect that there are already n relevant documents chosen in the set $S_{k-1}^* = \{s_1^*, \dots, s_{k-1}^*\}$, and hence in expectation $m = n$. Assuming $m = n$ we obtain

$$s_k^* = \arg \max_{s_k} \frac{n}{n+1} \text{Sim}_1(s_k, \mathbf{q}) - \frac{1}{n+1} \max_{s_i \in S_{k-1}^*} \text{Sim}_2(s_k, s_i, \mathbf{q}) \quad (13)$$

leading to $\lambda = \frac{n}{n+1}$ when comparing to MMR in (1), which achieves our goal of formally expressing the relevance vs. diversity tradeoff as a function of n .

As a reality check, we see that the result $\lambda = \frac{n}{n+1}$ coincides with the result of $\lambda = \frac{1}{2}$ for $n = 1$ as derived in the previous section for expected 1-call@ k . Hence, we have achieved one of our key theoretical results and have shown that when optimizing expected n -call@ k w.r.t. a latent subtopic binary relevance model diversification decreases as $n \rightarrow 1$, thus confirming the empirical observations regarding n -call@ k in [Wang and Zhu 2009] (Figure 2c).

3.3. Connections to Maximal Marginal Relevance

The result in (13) is strikingly similar to maximal marginal relevance (MMR) [Carbonell and Goldstein 1998] — 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 $P(t_i = t|s_i)$ with vector indices for each topic $t \in T$. Then the similarity and diversity terms from (13) can be respectively written as

$$\text{Sim}_1(\mathbf{q}, s_k) = \sum_{t \in T} P(t' = t|\mathbf{q})P(t_k = t|s_k) = \langle \mathbf{T}', \mathbf{T}_k \rangle \text{ and} \quad (14)$$

$$\text{Sim}_2(s_k, S_{k-1}) = \sum_{t \in T} P(t|\mathbf{q})P(t_k = t|s_k)\tilde{P}(t|S_{k-1}^*) = \langle \mathbf{T}_k, \mathbf{T}_{S_{k-1}^*} \rangle_{\mathbf{T}'}. \quad (15)$$

Here, we let $\langle \cdot, \cdot \rangle$ denote an inner product of two vectors and $\langle \cdot, \cdot \rangle_{\mathbf{v}}$ a *v-reweighted* inner product, defined as in (15).

While having similarity and diversity terms similar to MMR, Exp- n -call@ k in (13) differs from MMR in two notable ways:

- (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. Hence we can avoid directly answering the question of how subtopics and subtopic probabilities are specified by using the “kernel trick” and kernels with a PPK interpretation like TF or TFIDF to *directly* compute the required optimization quantities in (14) and (15).
- (2) Optimizing Exp- n -call@ k introduces query-specific relevance into the diversification term as shown by the query topic (\mathbf{T}') reweighted diversity function in (15). Indeed this modification to MMR is quite intuitive and may prove advantageous since result set diversification need only be considered over the subtopics that are actually relevant to the query.

Despite these minor differences, however, we note the strong similarity between the optimization of expected n -call@ k and MMR where we formally derived that $\lambda = \frac{n}{n+1}$.

4. RELATION TO OTHER DIVERSIFICATION APPROACHES

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. We note that the results as derived in the last section formally motivate much of MMR and now we discuss further connections between optimizing expected $n\text{-call}@k$ and other diversification approaches proposed in the literature.

4.1. Greedy Submodular Optimization

[Agrawal et al. 2009] proposes a set-based objective function to answer ambiguous web queries in a setting where there exists a predefined taxonomy of information, and that both queries and documents may belong to more than one category according to this taxonomy. The proposed set-based objective function aims at maximizing the probability that the average user finds at least one useful resulting document retrieved within the top k results. Mathematically, this objective function is defined below:

$$P(S|\mathbf{q}) = \sum_c P(c|\mathbf{q}) \left(1 - \prod_{s \in S} (1 - V(s|q, c)) \right) \quad (16)$$

where S is a set of documents, and $V(s|q, c)$ broadly defines the likelihood that a document s satisfies the query \mathbf{q} given the taxonomy (or category) c of the query and the document.³ The authors point out that one particular advantage of this set-based objective function is that it is submodular, hence permitting a $(1 - \frac{1}{e})$ approximation if optimized greedily.

Now we show that our expected $1\text{-call}@k$ objective in Equation (4) is equivalent to the objective function in Equation (16) by simply writing out the mathematical expectation in terms of the sum over all possible topics (e.g., taxonomy in [Agrawal et al. 2009]) the weighted relevance where weights are the topic distributions below

$$\begin{aligned} \text{Exp-1-Call}@k(S_k, \mathbf{q}) &= \mathbb{E} \left[\bigvee_{i=1}^k r_i = 1 \mid s_1, \dots, s_k, \mathbf{q} \right], \\ &= \sum_{t \in T} P(t|\mathbf{q}) \left(1 - \prod_{i=1}^k (1 - p(t_i = t|q, s_i)) \right) \end{aligned}$$

Clearly our proposed objective is equivalent to the objective Equation (16) proposed in [Agrawal et al. 2009] when one replaces the likelihood function $V(s|\mathbf{q}, c)$ by $p(t_i = t|\mathbf{q}, s_i)$. More recently, [Vargas et al. 2012] propose variants of IA-Select by introducing several interesting formal probabilistic relevance models to instantiate $V(s|q, c)$, which are more appropriate in modeling the relevance in a probabilistic framework. Furthermore, [Vallet and Castells 2012] propose to introduce a user as an explicit random variable in state of the art diversification methods, thus developing a generalized framework for personalized diversification.

Another recent interesting instantiation of $V(s|q, c)$ is proposed in [Zuccon et al. 2012] motivated by the facilitation location problem [Gonzalez 2007] taken from Operations Research: for a set of customer “locations” D , one aims at choosing a subset S in D to open k “facilities” that optimize a graph-theoretic objective that depends on

³We have slightly adapted notation in the above equation from the original for consistency with our previous definitions in this article.

the cost of opening a facility at each location and also the distance between each pair of locations. However all of the three described methods do not derive their objective functions from the expected 1-call@ k objective as we have achieved.

4.2. 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. Note that diversification based on portfolio theory is extended in [Shi et al. 2012] by introducing latent factors for collaborative filtering tasks. One major difference in the framework [Wang and Zhu 2009] 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.

4.3. 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 (8) 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 $\{s_1^*, \dots, s_{k-1}^*\}$. 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 $\{s_1^*, \dots, s_{k-1}^*\}$:

$$\begin{aligned} & \prod_{i=1}^{k-1} (1 - P(t_i = t|s_i^*)) \\ &= 1 - \left[\sum_{i=1}^{k-1} P(t_i = t|s_i^*) - \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} P(t_i = t|s_i^*) P(t_j = t|s_j^*) + \dots - (-1)^{k-1} \prod_{i=1}^{k-1} P(t_i = t|s_i^*) \right] \end{aligned} \quad (17)$$

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. (17) 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 17 is illustrated in Figure 2. In words, this term is calculating the total topic probability coverage of t by all selected items $\{s_1^*, \dots, s_{k-1}^*\}$ by properly applying the inclusion-exclusion principle to ensure that overlapping probability coverage is not double counted. Then referring back to Equation 7, we note that s_k is chosen by max-

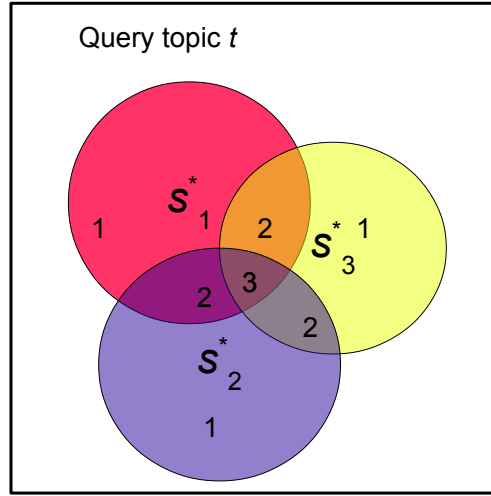


Fig. 2. Inclusion-exclusion principle. The sets represent candidate items s for a query, and the area covered by each set is the “information” covered by that item for query topic t . Numbers on different areas indicates the number of sets that share these areas.

imizing 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 $\{s_1^*, \dots, s_{k-1}^*\}$ to naturally encourage diversity. We note that this is a soft probabilistic version of the “in or out” topic coverage approach of WSL.

4.4. 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 .

4.5. 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 formalize MMR and instead use naïve Bayes to directly evaluate (9). 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.

4.6. 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- n -

$\text{call}@k$ 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 $\text{Exp-}n\text{-call}@k$ result set.

5. CONCLUSION

In this article, we made fundamental contributions to the theoretical understanding of expected $n\text{-call}@k$ (guided by simpler derivations for $n = 1, 2$) and its relationship with diversity. In contrast to a variety of diverse retrieval algorithms derived from alternate rank-based relevance criteria such as average precision and reciprocal rank, we provided a complementary theoretical perspective on the emergence of diversity via optimization of $n\text{-call}@k$ in a latent subtopic model of relevance that focused on the subtopic motivation for diverse retrieval. Further, we precisely formalized the mathematical relationship between $n\text{-call}@k$ and diversity and confirmed empirical observations in the literature. Altogether, these theoretical results have provided deep connections to the popular MMR diversification algorithm and both motivated and contrasted with other diversification approaches, hence providing a new theoretical basis for the investigation of diverse retrieval algorithms.

APPENDIX

A. DIRECT DERIVATION FOR EXPECTED 2-CALL@K

The following derivation serves as a specific example to help understanding the derivation of optimizing expected $n\text{-call}@k$ for general $n \geq 1$. This is a straightforward extension to optimizing expected $1\text{-call}@k$, utilizing a similar derivational approach based on a logical expansion of the optimization objective:

$$\begin{aligned}
 s_k^* &= \arg \max_{s_k} \mathbb{E} [R_k \geq 2 \mid S_{k-1}^*, s_k, \mathbf{q}] \\
 &= \arg \max_{s_k} \mathbb{E} \left[(r_1 = 1 \wedge r_2 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 1) \vee (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 1) \vee \right. \\
 &\quad (r_1 = 0 \wedge r_2 = 0 \wedge r_3 = 1 \wedge r_4 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 0 \wedge r_4 = 1) \vee \\
 &\quad (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 0 \wedge r_4 = 1) \vee \dots \vee \\
 &\quad (r_1 = 0 \wedge \dots \wedge r_{k-2} = 0 \wedge r_{k-1} = 1 \wedge r_k = 1) \vee \\
 &\quad (r_1 = 0 \wedge \dots \wedge r_{k-3} = 0 \wedge r_{k-2} = 1 \wedge r_{k-1} = 0 \wedge r_k = 1) \vee \dots \vee \\
 &\quad \left. (r_1 = 1 \wedge r_2 = 0 \wedge \dots \wedge r_{k-1} = 0 \wedge r_k = 1) \mid S_{k-1}^*, s_k, \mathbf{q} \right]
 \end{aligned}$$

Next we reorganize our formula to isolate those terms mentioning r_k :

$$\begin{aligned}
 s_k^* &= \arg \max_{s_k} \mathbb{E} \left[(r_1 = 1 \wedge r_2 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 1) \vee (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 1) \vee \right. \\
 &\quad (r_1 = 0 \wedge r_2 = 0 \wedge r_3 = 1 \wedge r_4 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 0 \wedge r_4 = 1) \vee \\
 &\quad (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 0 \wedge r_4 = 1) \vee \dots \vee \\
 &\quad \left. \bigvee_{j=1}^{k-1} \left(r_k = 1 \wedge \bigwedge_{\substack{i=1 \\ i \neq j}}^{k-1} r_i = 0 \wedge r_j = 1 \right) \mid S_{k-1}^*, s_k, \mathbf{q} \right]
 \end{aligned}$$

Now we rewrite disjoint events as a sum of probabilities, ignoring constant probabilities independent of r_k that do not change the $\arg \max_{s_k}$ results:

$$s_k^* = \arg \max_{s_k} \sum_{j=1}^{k-1} P \left(r_k = 1 \wedge \bigwedge_{\substack{i=1 \\ i \neq j}}^{k-1} r_i = 0 \wedge r_j = 1 \middle| S_{k-1}^*, s_k, \mathbf{q} \right)$$

Next we formally evaluate the above query w.r.t. our latent subtopic binary relevance model in Figure 1 and simplify:

$$\begin{aligned} s_k^* &= \arg \max_{s_k} \sum_{j=1}^{k-1} \left(\sum_{t_1, \dots, t_k, t} P(t|\mathbf{q}) P(t_k|s_k) \mathbb{I}[t_k = t] P(t_j|s_j^*) \mathbb{I}[t_j = t] \prod_{\substack{i=1 \\ i \neq j}}^{k-1} P(t_i|s_i^*) \mathbb{I}[t_i \neq t] \right) \\ &= \arg \max_{s_k} \sum_t P(t|\mathbf{q}) P(t_k = t|s_k) \sum_{j=1}^{k-1} \left(P(t_j = t|s_j^*) \prod_{\substack{i=1 \\ i \neq j}}^{k-1} (1 - P(t_i = t|s_i^*)) \right) \end{aligned}$$

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

$$s_k^* = \arg \max_{s_k} \sum_t P(t|\mathbf{q}) P(t_k = t|s_k) \sum_{j=1}^{k-1} \left(P(t_j = t|s_j^*) \prod_{\substack{i=1 \\ i \neq j}}^{k-1} (1 - P(t_i = t|s_i^*)) \right)$$

Exploiting the same inversion idea used to replace \square in (7):

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

Distributing $P(t_j = t|s_j^*)$, the relevance and diversity components begin to emerge:

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

Finally, making the deterministic subtopic probability assumption that $P(t_k = t|s_k) \in \{0, 1\}$ and $P(t_j = t|s_j^*) \in \{0, 1\}$, we obtain the final result:

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

We remark that this is of the same form as (11) in Section 3.2 when $n = 2$, albeit derived in a much simpler fashion under this more restrictive case. The derivation for 2-call@k from this point proceeds identically to that already provided following (11).

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.
- 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.
- 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.
- GOFFMAN, W. 1964. On relevance as a measure. *Information Storage and Retrieval* 2, 3, 201–203.
- GONZALEZ, T. F. 2007. *Handbook of Approximation Algorithms and Metaheuristics (Chapman & Hall/Crc Computer & Information Science Series)*. Chapman & Hall/CRC.
- 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.
- LIM, K. W., SANNER, S., AND GUO, S. 2012. On the mathematical relationship between expected n-call@k and the relevance vs. diversity trade-off. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '12. ACM, New York, NY, USA, 1117–1118.
- RADLINSKI, F., BENNETT, P. N., CARTERETTE, B., , AND JOACHIMS, T. 2009. Redundancy, diversity and interdependent document relevance. *SIGIR Forum* 43, 2, 46–52.
- SALTON, G. AND MCGILL, M. 1983. *Introduction to modern information retrieval*. McGraw-Hill.
- 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.
- SHI, Y., ZHAO, X., WANG, J., LARSON, M., AND HANJALIC, A. 2012. Adaptive diversification of recommendation results via latent factor portfolio. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '12. ACM, New York, NY, USA, 175–184.
- VALLET, D. AND CASTELLS, P. 2012. Personalized diversification of search results. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA.
- 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.
- 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.
- ZUCCON, G., AZZOPARDI, L., ZHANG, D., AND WANG, J. 2012. Top-k retrieval using facility location analysis. In *Proceedings of the 34th European conference on Advances in Information Retrieval*. ECIR'12. Springer-Verlag, Berlin, Heidelberg, 305–316.