Explicit Relevance Models in Intent-Oriented Information Retrieval Diversification

Saúl Vargas, Pablo Castells and David Vallet
Universidad Autónoma de Madrid
Escuela Politécnica Superior, Departamento de Ingeniería Informática
{saul.vargas,pablo.castells,david.vallet}@uam.es

ABSTRACT

The intent-oriented search diversification methods developed in the field so far tend to build on generative views of the retrieval system to be diversified. Core algorithm components –in particular redundancy assessment– are expressed in terms of the probability to observe documents, rather than the probability that the documents be relevant. This has been sometimes described as a view considering the selection of a single document in the underlying task model. In this paper we propose an alternative formulation of aspect-based diversification algorithms which explicitly includes a formal relevance model. We develop means for the effective computation of the new formulation, and we test the resulting algorithm empirically. We report experiments on search and recommendation tasks showing competitive or better performance than the original diversification algorithms. The relevance-based formulation has further interesting properties, such as unifying two well-known state of the art algorithms into a single version. The relevance-based approach opens alternative possibilities for further formal connections and developments as natural extensions of the framework. We illustrate this by modeling tolerance to redundancy as an explicit configurable parameter, which can be set to better suit the characteristics of the IR task, or the evaluation metrics, as we illustrate empirically.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval – *retrieval models*

General Terms

Algorithms, Measurement, Performance, Experimentation, Theory

Keywords

Diversity, relevance models, language models, generative models

1. INTRODUCTION

The value of diversity as a fundamental dimension of information utility has started to be cared for in the Information Retrieval (IR) research community for over a decade [4]. Diversity-enhancement methods have been developed [1,7,21,27,29], diversity evaluation methodologies and metrics have been proposed [1,8,24,29], and a diversity task has been included in the latest TREC campaigns [9]. Theories of IR diversity build on a revision of the classic document independence assumption in IR: the marginal utility of a document indeed highly depends on –in general, decreases with– the relevance of the documents the user has previously seen. Considering this, IR systems' output diversification is posited as an effective strategy to cope with the uncertainty (ambiguity and/or incompleteness) involved in user queries, as imperfect expressions of true user needs. By trading diminishing marginal relevance for increased query aspect coverage, diversification strategies seek to maximize the range of users (the precise potential intents behind the query) who will get at least some relevant document, thereby improving the overall gain [1,7,10,21].

Two trends can be broadly distinguished in the diversification methods reported in the literature, based on whether or not they use an explicit representation of query intents [22]. Whereas intent-oriented methods include an explicit query aspect space in their formulation, implicit diversification schemes typically rely at some level on inter-document similarity, under the assumption that dissimilar documents should cover diverse query aspects. Interestingly, intent-implicit approaches generally build –as far as their formalization goes– on an explicit relevance model, often lending from the Probability Ranking Principle [7,27,29]. Whereas, in contrast, the intent-explicit methods tend to elaborate, in their problem statement, formalization, and algorithmic formulation, on generative views of the retrieval system to be diversified, where relevance is implicit [1,21,25]. Core algorithm components –in particular redundancy assessment– are expressed in such approaches in terms of the probability to observe documents, rather than the probability that the documents be relevant. This has been sometimes described as a view considering the selection of a single document in the underlying task model [28].

In this paper we propose an alternative formulation of aspect-based diversification algorithms which explicitly includes a formal relevance model. Our research has a theoretical motivation in seeking an alternative, nuanced understanding of the intent-explicit diversity problem, and the resulting variants in the formulation and development of aspect-based diversification methods. On the other hand, the explicit relevance approach has advantages of its own. We report experiments in search and recommendation tasks where the approach shows competitive or better performance than its original counterparts. Moreover, the relevance-based formulation opens the way for further extensions and elaborations with models involving an explicit representation of relevance. As a particular case, we show that the framework provides a sound basis for tuning redundancy penalization in a principled way, as a smooth consistent extension of the diversity model.

The rest of the paper is organized as follows. We briefly introduce and discuss intent-oriented diversification schemes in the next section, paying specific attention to their document-oriented gen-

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

erative formal foundation. We introduce our relevance-based revision of such schemes after that in section 3, including details for the development, estimation and computation of the components of our resulting diversification framework. We empirically test the effectiveness of our approach in section 4. Section 5 describes a formal extension of the proposed framework to model and adjust the diversification to different degrees of tolerance to redundancy. We discuss the results and implications of our approach in section 6, and conclude with a summary and final comments in section 7.

2. RELEVANCE MODELS VS. GENERATIVE MODELS

The different views on relevance implied by the Probability Ranking Principle (PRP) [20] and the Language Modeling (LM) approaches to IR [16] raised interest and debate in the research community by the turn of the past decade [23]. The absence of a clear explicit notion of relevance in the early LM formulations has been often considered to involve an underlying assumption of single relevant document selection [3]. Even though –or precisely because– there does not seem to be a unique common understanding on such issues in the field, and a unique view on whether or how LM actually capture relevance, we see theoretical –and potentially practical– interest in exploring new formulations of the diversity problem –and the derivation of algorithms thereupon– which explicitly build on the probability of relevance.

As stated in the introduction, we focus on approaches to diversity which are founded on an explicit representation of query intents. As two prototypical representative instances of this approach, we focus on IA-Select [1] and xQuAD [21] as the algorithms of reference for our study. Both have proved to be quite effective in IR diversity tasks, outperforming prior non-explicitly intent-oriented approaches (see [22] for a comprehensive empirical analysis and comparisons). We start by briefly reviewing the formulation of these diversification schemes, and the fundamental principles on which they build, in order to contrast them later to a relevance-based alternative.

The definition of the IA-Select approach is developed in [1] around a quality component $V(d|q,c)$ broadly defined as the likelihood that a document d satisfies the query q given the user intent c. The diversification problem is stated as finding a subset of documents –of a given size– that maximizes the probability that at least one of them is relevant. This is formulated as finding the set S that minimizes:

$$p(S|q) = \sum_c p(c|q)\left(1 - \prod_{d\in S}\big(1 - V(d|q,c)\big)\right) \qquad (1)$$

While solving this problem is NP-Hard, the authors provide a practical approximation to the optimal solution by greedily re-ranking a baseline set of documents picking one document at a time that maximizes the following objective function:

$$g(d|q,c,S) = \sum_c p(c|q)\, V(d|q,c) \prod_{d'\in S}\big(1 - V(d'|q,c)\big) \quad (2)$$

where S is the set of documents that were selected in the previous iterations. Agrawal et al generically describe $V(d|q,c)$ as a term quantifying to what extent the document d responds to the user need expressed by q when the intended sense of the query is c. The authors do not enforce a strict probabilistic rigor in the development of this component, and in fact omit an explicit specification, thus leaving its implementation somewhat open to different realizations of the expressed principle. For their own experiments, the implementation of $V(d|q,c)$ is hinted as the product of a baseline retrieval scoring function (the system to be diversified) by the probability that d belongs to category c –which may be read as $p(c|d)$. No explicit probability of relevance is introduced in their development and –if we may indulge on a rather informal note– one might find in the $V(d|q,c)$ notation some reminiscence of a probability distribution over documents.

xQuAD does have a quite precise probabilistic formulation [21]. It is stated and developed upon a document generative model formulation, whereby the resulting key terms of the algorithms reflect the probability to observe documents, rather than the explicit probability of the latter to be relevant. Specifically, the xQuAD scheme consists of a greedy algorithm with the same essential structure as IA-Select, but with a differently defined objective function:

$$g(d|q,c,S) = (1-\lambda)\, p(d|q) + \lambda\, p(d,\neg S|q) \qquad (3)$$

That is, the objective is a linear combination of the probability to observe a document given the query (which may be understood as the baseline retrieval function), and the probability to observe the document given the query, assuming –strictly sticking to the probabilistic notation– no previously selected document had be observed given the query (which represents the marginal utility of the document). The marginal utility component is in turn developed by marginalizing over the set of query intents, which after some Bayesian derivations and mild conditional independence assumptions results in:[1]

$$p(d,\neg S|q) \sim \sum_c p(c|q)\, p(d|c,q) \prod_{d'\in S}\big(1 - p(d'|c,q)\big)$$

Thus the xQuAD algorithm is formulated and developed upon (conditional) document distributions, as ultimately embodied in the $p(d|q)$ and $p(d|c,q)$ components.

Therefore neither IA-Select, nor xQuAD, or other explicitly intent-oriented diversification schemes that we are aware of, include relevance as an explicit magnitude or random variable.

3. RELEVANCE-BASED DIVERSIFICATION ALGORITHMS

3.1 Relevance-Based IA-Select

Taking up from a more literal interpretation of Agrawal's initial problem statement, we investigate a revision of the IA-Select formulation by defining $V(d|q,c) \equiv p(r|d,q,c)$, where r is the binary relevance random variable (and r is used as a shorthand for $r=1$). With this formulation, the objective function to be maximized in the problem statement (equation 1) then explicitly becomes:

$$p(S|q) = \sum_c p(c|q)\left(1 - \prod_{d\in S}\big(1 - p(r|d,q,c)\big)\right)$$

[1] The original version of xQuAD [21] would use $p(d|c)$ rather than $p(d|q,c)$. Later publications [22] present however the latter form, which we have also found to work better in our experiments, whereby we favor it here. Our analysis and considerations apply just the same to both variants in any case.

And the marginal utility objective function for the greedy approximation algorithm (equation 2) gets then defined as:

$$g(d|q,c,S) = \sum_c p(c|q)\, p(r|d,q,c) \prod_{d' \in S} \big(1 - p(r|d',q,c)\big) \quad (4)$$

In order to implement this formulation we need a means to estimate $p(r|d,q,c)$. Before addressing that, we formulate a similar revision of the xQuAD algorithm.

3.2 Relevance-Based xQuAD

For this algorithm we actually reconsider the own initial formulation of the approach. Rather than expressing the objective function in terms of document probabilities as in equation 3, we define the objective function on an explicit relevance model, as:

$$g(d|q,c,S) = (1-\lambda)\, p(r_d|q) + \lambda\, p(r_d, \neg r_S|q) \quad (5)$$

where r_d means d is relevant –that is, $p(r_d|q) \stackrel{\text{def}}{=} p(r|d,q)$– and $\neg r_S$ means no document in S is relevant. Taking this starting point, by similar steps to the original xQuAD, we derive:

$$\begin{aligned} p(r_d, \neg r_S|q) &= \sum_c p(c|q)\, p(r_d, \neg r_S|q,c) \\ &= \sum_c p(c|q)\, p(r|d,c,q) \prod_{d' \in S} \big(1 - p(r|d',c,q)\big) \end{aligned}$$

where we have assumed r_d and r_S are conditionally independent given q and c. Substituting all this in the objective function gives:

$$\begin{aligned} g(d|q,c,S) = (1-\lambda)\, p(r|d,q) + \\ + \lambda \sum_c p(c|q)\, p(r|d,q,c) \prod_{d' \in S} \big(1 - p(r|d',q,c)\big) \end{aligned}$$

It is easy to see that the above function becomes the relevance-based version of IA-Select (equation 4) when $\lambda = 1$. We hence notably find that, on a relevance model foundation, xQuAD turns out to be a generalization of IA-Select. We shall thus deal with the two reformulations as a single approach henceforth.

We now turn to the problem of estimating the two models involved in the new objective function, so as to come up to an effectively computable form for the latter.

3.3 Aspectual Relevance Model

Two main components: the aspect model $p(c|q)$, and the relevance model $p(r|d,q,c)$, need to be estimated in order to compute the objective function $g(d|q,c,S)$ of our diversification approach.

First, for the relevance model, by applying Bayes' rule we have:

$$p(r|d,q,c) = \frac{p(c|d,q,r)\, p(r|d,q)}{p(c|d,q)}$$

The $p(r|d,q)$ component can be obtained from the retrieval system, as we shall discuss later in the next subsection.

The marginalization of the conditional aspect distribution with respect to relevance gives $p(c|d,q) = p(c|d,q,r)\, p(r|d,q) + p(c|d,q,\neg r)\, \big(1 - p(r|d,q)\big)$. Therefore by substitution in the numerator above, the relevance model can be rewritten as:

$$p(r|d,q,c) = \frac{p(c|d,q) - p(c|d,q,\neg r)\, \big(1 - p(r|d,q)\big)}{p(c|d,q)} \quad (6)$$

Now we assume that under non-relevance, aspects are independent from documents and queries, that is $p(c|d,q,\neg r) \sim p(c|\neg r)$. In other words, given non-relevance, we consider there is no particular relation the aspect is enforced to meet with respect to the query and the document –other than not forming a relevant tuple. Since most aspect / document / query tuples are not related in practice, as relevance is highly sparse, such negative condition can be considered as negligible. Comparable assumptions can be found in probabilistic developments in the literature (see e.g. [29]) –only with no aspect variable involved. Furthermore, we assume no particular bias between aspects towards relevance or non-relevance, i.e. we approximate $p(c|\neg r) \sim p(c)$. The aspect prior can be estimated from the coverage of aspects in the document collection, if the latter is considered as a fair representative sample space for users' information need intents. Or, in the absence of further information, the prior can be taken as uniform.

Now, $p(c|d,q)$ for equation 6 can be approximated in different ways, depending on the nature of the aspect space and the available observations. When aspects consist of categorical data in such a way that $p(c|d)$ can be directly estimated from the data, we use the approach:

$$p(c|d,q) \sim \frac{p(c|d)\, p(c|q)/p(c)}{\sum_{c'} p(c'|d)\, p(c'|q)/p(c')}$$

using Bayes' rule, marginalization over aspects (in the denominator), and assuming again conditional independence of documents and queries given a query aspect (if a uniform aspect prior is assumed, $p(c)$ and $p(c')$ further cancel out). In our experiments, we take this approximation when ODP is used as the aspect space for search diversification (see section 4.1). We use it in movie and music recommendation tasks as well, where movie genres and artist tags, respectively, are taken as the aspect space (section 4.2).

When aspects belong to the query space, $p(d|c)$ can be estimated from the baseline retrieval function (as suggested e.g. in [21,22]), in a similar way as $p(d|q)$ is estimated for queries (see equation 8 below). In that case, we take the approach:

$$p(c|d,q) \sim \frac{p(d|c)\, p(c|q)}{p(d|q)} \quad (7)$$

using Bayes' rule and assuming conditional independence of documents and queries given a query aspect. Note that this independence assumption is rather mild in this context, since the estimation of $p(d|c)$ applies to documents sampled from the result set for q –i.e. $p(d|c)$ in practice is meaning $p\big(d\big|c,R_q\big)$, which is a form of blind relevance feedback approximation to $p(d|c,q)$, R_q being the set of documents returned for q that are to be diversified. In our experiments, we shall use this approximation when TREC subtopics (manually provided by human assessors for each query) are taken as the aspect space by the diversification algorithm (see section 4.1).

The generative model $p(d|q)$ in equation 7 can be either estimated (more or less) directly from the baseline search system –when the retrieval function implements a language model– or it can be approximated from a relevance model as (see e.g. [3] for this type of derivation):

$$p(d|q) \sim \frac{p(r|d,q)}{\sum_{d' \in R_q} p(r|d',q)} \quad (8)$$

Finally, as to the estimation of $p(c|q)$, we consider two alternatives. For categorical aspects, we marginalize over documents in the result set:

$$p(c|q) \sim \sum_{d \in R_q} p(c|d)\, p(d|q)$$

where we assume conditional independence of query aspects and queries given a document in R_q, and we use the estimate for $p(d|q)$ discussed above (e.g. equation 8). For manually-provided query subtopics (TREC data), we assume a uniform distribution $p(c|q) \sim 1/n_q$, where n_q is the number of subtopics of query q.

3.4 Relevance Model Estimation

Estimating the explicit probability of relevance of a document for a query has been researched in the IR literature from different points of view, distinguished from each other by the source from which relevance distributions are estimated.

Retrieval systems based on the PRP rank documents by decreasing probability of relevance $p(r|d,q)$ given a query [20]. However, rather than directly estimating this probability, these systems compute a function that is equivalent in rank to $p(r|d,q)$, and easier to estimate. The retrieval function is obtained by a series of monotonous operations on the probability of relevance, which involve lossy transformations, in the sense that it is not possible to recover $p(r|d,q)$ back from the final expression. There is thus not in general an analytic approach to obtain an explicit relevance model from a PRP retrieval system [3].

On the other hand, a number of studies have researched the relation between retrieval function score distributions and relevance [2], and some have devised approaches to estimate the probability of relevance from score values [18]. In essence, these approaches analyze the distribution of the scoring function, and their correspondence with known relevance information, thus training a model of relevance given system scores, by regression on the available data [18].

We propose a similar but quite easier approach to estimate the probability of relevance, which just requires the availability of either precision estimates, or click statistics, for the retrieval system being diversified. Rather than mapping scores to probabilities of relevance, we use a common relevance estimate for all queries, based only on the position of documents. Namely, we estimate the probability of relevance by:

$$p(r|d,q) \sim p_s(r|\tau_s(d,q),q)$$

where $\tau_s(d,q)$ is the rank position of document d in the result list returned by the baseline retrieval s in response to query q. Note that this step involves, in a way, a form of rank-based retrieval system output normalization where (similarly to e.g. rank-sim normalization [17]) we only use the document order τ_s returned by the system, regardless of the score values. After this step, estimating the probability of relevance of a document for a query amounts to estimating the probability of relevance at each rank position $p_s(r|k,q)$ for the retrieval system s. To simplify the notation, we shall henceforth drop the 's' subscript from $p_s(r|k,q)$, but it should be implicitly understood, as this distribution estimate is system-dependent.

3.4.1 Estimate on Relevance Judgments

If relevance judgments are available to train the relevance model or, equivalently, we have an estimate of precision at k of the retrieval system (for k ranging up to the size of the document set to be diversified), the positional relevance probability can be estimated as:

$$p(r|k,q) \sim k\,P_q@k - (k-1)\,P_q@(k-1) \qquad (9)$$

where $P_q@k$ denotes the precision at k of the system on query q. In our experiments we estimate precision by splitting the set of test queries (with their relevance judgments) into two disjoints subsets of equal size for 2-fold cross-validation. For the queries in one subset, we approximate $p(r|k,q) \sim p(r|k)$ using an average precision estimate $P_q@k \sim P@k$, computed in the complementary query subset.

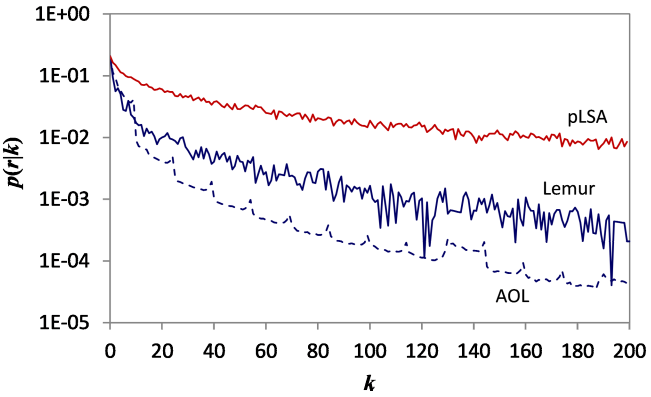

Figure 1. Relevance distribution estimate on the top 200 positions for different retrieval systems on different datasets (y axis displayed in log scale). The estimates for Lemur and pLSA are based on $P@k$ measurements, whereas the AOL curve is derived from click statistics from the AOL query log. The Lemur curve is computed on TREC 2009/10 diversity task data. The relevance curve of pLSA is derived from a recommendation task on MovieLens data (see section 4.2).

As an illustrative example, Figure 1 shows the average relevance distribution estimate resulting for the Lemur Indri search system and the pLSA recommender –which we use as baselines in our experiments in section 4. The precision estimates are taken from the TREC 2009/10 diversity task data for Lemur, and from the MovieLens[2] dataset for pLSA (more details in section 4.2). For Lemur, the distribution decreases from $p(r|1) \sim 0.21$ to around 10^{-4} by $k = 200$. pLSA displays a higher relevance probability due to the nature of the recommendation task on this dataset. Compared to more involved methods such as the regression techniques described in [18], this approach is equivalent to simply using the raw relevance data (a histogram of positional relevance) without regularization, rather than a parameterized fit (e.g. by a logistic function).

3.4.2 Estimate on Click Statistics

A positional relevance model $p(r|k)$ can also be built from simple click statistics for the baseline retrieval system [19]. Assuming a cascade browsing model [15], and the simplifying assumption that a document is clicked if and only if it is relevant, we may consider the approximation:

$$p(click|k) \sim p(\neg stop|k-1)\,p(r|k) \qquad (10)$$

where $p(click|k)$ is the probability that a document at position k is clicked, and $p(\neg stop|k-1)$ denotes the event that the user continues (does not stop) browsing after position $k-1$. This term can in turn be decomposed by marginalization into the relevance and non-relevance cases:

[2] http://www.grouplens.org/node/73

$$p(\neg stop|k) = p(\neg stop|k,r)\,p(r|k) + \\ + p(\neg stop|k,\neg r)\,\big(1 - p(r|k)\big) \qquad (11)$$

Assuming the probability to stop is independent from the position of the document given its relevance, and combining equations 10 and 11 we get:

$$p(r|k) \sim \frac{p(click|k)}{p(\neg stop|r)p(r|k-1) + p(\neg stop|\neg r)\big(1 - p(r|k-1)\big)}$$

Starting from $p(r|1) \sim p(click|1)$, the above equation provides a recursive means to estimate the probability of relevance at each position in the ranking by just having some statistics (e.g. from a query log) of the ratio of clicks at each position, i.e. a background estimate of $p(click|k)$. The two remaining parameters represent a user model (as has been well-studied before, see e.g. [15]), where $p(\neg stop|\neg r)$ reflects the user's "patience", and $p(\neg stop|r)$ is related to how many relevant documents the user is willing to get. This user model allows to account for the position bias in click statistics, and can be estimated in different ways. In the absence of specific data or criteria about this, a simplification such as $p(stop|r) = 1$ –a single relevant document is enough– and $p(stop|\neg r) = 0$ –the user never gives up until finding a relevant document– has been found to be acceptable for many purposes [8,15], which in our case just yields:

$$p(r|k) \sim \frac{p(click|k)}{1 - p(r|k-1)} \qquad (12)$$

If click data became too sparse below a certain position, $p(click|k)$ might be estimated by regression techniques from this point on, since we need an estimate of $p(r|k)$ for as many positions as the diversification system is intended to rerank.

As an illustrative example, Figure 1 shows the relevance curve derived from the AOL query log dataset using this approach, which we shall also test in our experiments. The curve is not significantly far below Indri on TREC 2009/10 (in fact it is above it in the top ten). Since AOL can be assumed to be a well optimized engine, this (small) difference can be attributed to a divergence in user behavior beyond the first web search results page (i.e. the behavior on the first and subsequent pages do not fit the same model), and/or the imprecision of the model. On the other hand, the AOL curve reflects practical (and personal) utility for real Web users, which is probably more demanding than topical (less subjective) relevance as applied by TREC assessors.

The estimation of relevance models continues to be a research topic in the IR field, and our approach might benefit from any improvement on this point –the more accurate the relevance model, the better the diversification algorithm can be expected to perform. Interestingly however, even with rough relevance probability estimates like the ones described here, the relevance-based diversification method seems to achieve good performance. Testing our framework with simple approaches as defined in equations 9 and 12, we already observe comparable or better results than the formulations based on generative models, as we report in the next section.

4. EXPERIMENTAL RESULTS

We have tested our relevance-based diversification framework on two IR domains: ad-hoc search, as defined in the TREC Web track diversity task [9], and movie recommendation [24].

4.1 Search Diversity

In order to test our framework on search diversity, we use the data from the TREC diversity task, namely, the ClueWeb collection category B, the topic data, and the relevance assessments from the TREC 2009 and 2010 editions [9]. We use the Lemur Indri retrieval model[3] (version 5.2) as the baseline search engine.

We rerank the top 100 documents returned by the search system by the original a) IA-Select and b) xQuAD algorithms, and c) by our relevance-based reformulation. We test two query aspect spaces for diversification: 1) the Open Directory Project (ODP) categories as in [1], and 2) the subtopics manually provided in the TREC 2009/10 diversity task, as a reference for comparison. We do not apply an exhaustive optimization or tuning of the diversifiers, since the goal of the experiment is the comparison of alternatives, rather than reaching the maximum performance possible. To this respect, our results are roughly comparable in range to those reported e.g. in [21].

With ODP categories, we use the Textwise[4] classification service to estimate $p(c|d)$, by normalization of the score returned by the classifier. For TREC subtopics, we derive an estimate by submitting them as queries to the baseline search system, as discussed earlier in section 3.3. We use a simple uniform aspect prior estimate, and compute the rest of components as described in that section. We test both equations 9 and 12 to estimate the rank-relevance model $p(r|k)$, using relevance judgments from the TREC dataset and AOL click statistics, respectively. TREC relevance information is used in a 2-fold cross-validation, where the relevance judgments of TREC 2009 are used to estimate $P@k$ and derive $p(r|k)$ in TREC 2010 topics, and vice-versa.

Figure 2 compares the performance of the original xQuAD algorithm and the relevance-based reformulation (RxQuAD, using relevance judgments –equation 9– for the $p(r|k)$ estimate), measured by ERR-IA, for λ ranging from 0 to 1. It can be seen that the relevance-based approach consistently outperforms the original xQuAD version. The plots also give an idea of the sensitiveness of the algorithms to the choice of the λ parameter. It is interesting to notice that using TREC subtopics, the best result is reached for $\lambda = 1$, that is with maximum diversity. This makes sense, as the diversifiers use a "relevance-safe" aspect space, inasmuch as

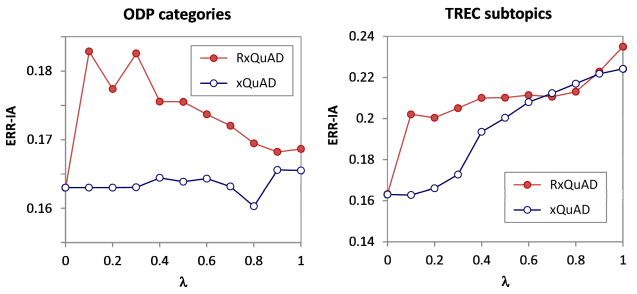

Figure 2. Comparative performance (measured by ERR-IA) of relevance-based diversification (RxQuAD) and xQuAD, ranking over values of λ. The algorithms are tested on TREC 2009/10 data for ClueWeb category B, using ODP (left) and TREC subtopics (right) as the query aspect space, and TREC relevance judgments for the $p(r|k)$ estimate.

[3] http://lemurproject.org/indri

[4] http://textwise.com

Table 1. Comparative evaluation on TREC 2009/10 data for ClueWeb category B, using ODP categories (above) and TREC subtopics (below). The best result for each metric (in the ODP and TREC blocks respectively) is highlighted in bold. For our RxQuAD scheme, we show the results with $p(r|k)$ estimates from the TREC relevance judgments (Qrels), and click statistics from AOL (Clicks). The value of λ that is used for xQuAD and RxQuAD is indicated on the corresponding row. Values marked with ▲ and △ indicate, respectively, significant and non-significant improvements over IA-Select and xQuAD (in this order). Similar convention with ▼ and ▽ indicates values below xQuAD or IA-Select. Statistical significance is established by Wilcoxon p < 0.05 in all cases.

			λ	α-nDCG @20	ERR-IA @20	nDCG-IA@20	S-precision @*r*	S-recall @20
	Lemur		-	0.2587	0.1630	0.2396	0.0788	0.4636
ODP categories	IA-Select		-	0.2651	0.1681	0.2423	0.0935	0.4483
	xQuAD		0.9	0.2675	0.1656	0.2451	0.0893	0.4864
	RxQuAD	Qrels	0.1	**0.2858**$^{\triangle\blacktriangle}$	0.1828$^{\triangle\blacktriangle}$	**0.2655**$^{\triangle\blacktriangle}$	**0.1045**$^{\triangle\blacktriangle}$	**0.4898**$^{\blacktriangle\triangle}$
		Clicks	0.4	0.2841$^{\blacktriangle\triangle}$	**0.1831**$^{\triangle\triangle}$	0.2605$^{\triangle\blacktriangle}$	0.1036$^{\triangle\blacktriangle}$	0.4830$^{\blacktriangle\triangledown}$
TREC subtopics	IA-Select		-	0.3541	0.2346	**0.3213**	**0.1300**	**0.5787**
	xQuAD		1.0	0.3445	0.2241	0.3127	0.1149	0.5704
	RxQuAD	Qrels	1.0	**0.3543**$^{\triangle\triangle}$	**0.2349**$^{\triangle\triangle}$	0.3192$^{\triangledown\triangle}$	0.1205$^{\blacktriangledown\triangle}$	0.5782$^{\triangledown\triangle}$
		Clicks	1.0	0.3512$^{\triangledown\triangle}$	0.2320$^{\triangledown\triangle}$	0.3166$^{\triangledown\triangle}$	0.1185$^{\blacktriangledown\triangle}$	0.5748$^{\triangledown\triangle}$

subtopics are tightly related to the query, and therefore diversifying for them does not result in such a potential relevance loss tradeoff as with ODP categories.

Table 1 shows the comparative results on a more complete set of diversity metrics, computed with the evaluation script from the TREC diversity task. We also show subtopic precision at *r* (S-precision@*r*) [29] where *r* is the subtopic recall (S-recall) in the set of documents being diversified. In other words, S-precision@*r* is the number of subtopics in the reranked set, divided by the first position where all the subtopics have been covered –it thus complements S-recall@20 with a measure of how early in the ranking all possible aspects are covered. Note that α-nDCG, ERR-IA, and nDCG-IA account for both diversity and relevance, whereas S-recall and S-precision@*r* are pure diversity metrics. For xQuAD and our approach, the results correspond to the best λ in terms of ERR-IA, selected manually with a precision of 0.1.

It can be seen that the relevance-oriented approach improves the two original diversifiers with ODP categories on all metrics (except on subtopic recall for the click-based configuration against xQuAD). With TREC subtopics, IA-Select is considerably effective –which we believe is due to the mostly single-valued aspect coverage per document, a situation where IA-Select seems to work particularly well–, and has similar effectiveness to our approach, with some advantage on S-precision@*r*. Our approach performs better than xQuAD on this aspect space in all cases though. In general, the best results for our framework are obtained with the baseline-specific estimate of $p(r|k)$ (equation 9) on TREC relevance judgments ('Qrels' rows in Table 1), as one might expect. We may however observe that the results obtained with the AOL click statistics ('Clicks' rows) are almost as good as the results with the relevance judgments. This suggests that our approach is not particularly demanding or sensitive to the nature of the required relevance information to estimate $p(r|k)$.

4.2 Recommendation Diversity

In order to test our approach on a different domain other than ad-hoc search, we conduct additional experiments on recommender system tasks. For this purpose we use two well-known datasets: the 1M version of the MovieLens collection, which includes one million ratings (on a 1-5 scale) by 6,040 users for 3,900 items; and an extract from Last.fm provided by Ò. Celma [6], including the full listening history of 992 users up till May 2009. The Last.fm data involves 176,948 artists and a total of 19,150,868 user accesses to music tracks. We use two recommender system baselines: a probabilistic Latent Semantic Analysis (pLSA) recommender [14], which is among the top performing on this data; and a simple non-personalized recommender which recommends movies by popularity (i.e. by the number of ratings), which has shown to be competitive in these settings [11].

We follow the adaptation of the diversity problem to recommendation scenarios proposed in [25], by which items (here movies and music artists) are taken as the equivalent of documents, and users play the part of queries. Movie genre is used as the equivalent of query aspects in MovieLens, and social tags (assigned to artists by Last.fm users) are used for the same purpose in the music dataset. MovieLens includes 19 genres, whereas in Last.fm we use a total of 123,819 different tags.

The user-item interaction data (user ratings in MovieLens, artist playcounts in Last.fm) is split into training and test sets, following common experimental practice in the Recommender Systems field [13], where the test data are used as the equivalent of relevance judgments. In our experiments we take five random 80-20% splits of the MovieLens data and repeat for 5-fold cross-validation. In Last.fm we do a temporal 60-40% data split of the song access records based on their timestamp (see [6] for further details on the evaluation methodology we adhere to).

Since the aspect space is categorical, and aspects are associated to items in a binary way (movies either belong or do not belong to a genre, and similarly for artist tags), we take the simple approximation $p(c|d) \sim [d \in c]/\sum_{c'}[d \in c']$, where $[d \in c] = 1$ when d "has" the aspect (genre or tag) c, and zero otherwise. In other words, we assume a uniform conditional aspect distribution among the set of aspects covered by each item. Different from the search setting, we estimate the background aspect prior from the overall distribution of aspects in the set of items, given that the item-aspect association is explicit, manual, reliable, and therefore can be considered fairly informative. We estimate $p(r|k)$ by equation 9, using test ratings as relevance judgments for the computation of $P@k$, with a random 2-fold split of test users. We use very slight variations of the derivations in section 3.3, better suit-

Table 2. Comparative evaluation on MovieLens (left) and Last.fm (right) data, for diversification over a pLSA recommender baseline (top block), and recommendation by item popularity (bottom). The best result of each metric is highlighted in bold for each block. The improvement of the baselines is statistically significant in all cases. Improvements with respect to IA-Select and xQuAD (in this order) are marked with ▲ when statistically significant, and △ otherwise. Values with ▼ indicate a significant decrease respect to them. Statistical significance is established by Wilcoxon p < 0.001 in all cases. The value of λ that is used for xQuAD and RxQuAD is indicated next to each row and dataset block.

		MovieLens						Last.fm					
	λ	α-nDCG @20	ERR-IA @20	nDCG-IA @20	S-precision @r	S-recall @20	λ	α-nDCG @20	ERR-IA @20	nDCG-IA @20	S-precision @r	S-recall @20
pLSA	-	0.3108	0.1880	0.2189	0.0658	0.6448	-	0.1209	0.0684	0.1566	0.2710	0.0738
IA-Select	-	0.3079	0.1885	0.2195	0.0684	0.6567	-	0.1198	0.0778	0.1795	0.2669	0.0727
xQuAD	0.9	**0.3564**	0.2053	**0.2415**	0.0959	0.7453	1.0	0.1264	0.0794	0.1857	0.2866	0.0765
RxQuAD	0.5	0.3467▲▼	**0.2116**▲△	0.2294▲▼	**0.1195**▲▲	**0.7806**▲▲	1.0	**0.1428**▲▲	**0.1030**▲▲	**0.1914**▲△	**0.3543**▲▲	**0.0906**▲▲
Item popularity	-	0.1944	0.1013	0.1101	0.0677	0.7205	-	0.1329	0.0849	0.2022	0.1468	0.0604
IA-Select	-	0.2239	0.1448	0.1377	0.0666	0.7045	-	0.1250	0.0910	0.2139	0.1471	0.0602
xQuAD	0.9	0.2315	0.1243	0.1343	0.1010	0.7893	1.0	0.1345	0.0947	**0.2284**	0.1587	0.0641
RxQuAD	0.9	**0.2413**▲▲	**0.1494**▲▲	**0.1429**▲▲	**0.1947**▲▲	**0.8413**▲▲	1.0	**0.1421**▲▲	**0.1058**▲▲	0.2136▲▼	**0.2096**▲▲	**0.0790**▲▲

ed for model estimation on recommendation input data. We also found it effective to normalize the top-level diversity component in the xQuAD schemes before its linear combination with $p(r|d,q)$ (equations 3 and 5). We use a distribution-based normalization technique [12] for both versions (the original xQuAD and our variation) which showed to be effective in our experiments.

Figure 3 shows the performance of our relevance-based algorithm (RxQuAD) compared to xQuAD on MovieLens (left) and Last.fm (right), in terms of ERR-IA, for λ ranging from 0 to 1. The two collections are quite different both in their volumetric statistics (size, etc.), the nature of the user-item interaction data (movie ratings vs. music track playcounts), and the nature and distribution of item aspects (editorial genres vs. community-contributed tags), which accounts for the different behavior of the algorithms with respect to the λ parameter. In particular, the performance of our approach shows a drastic drop from λ = 0.9 to 1 in MovieLens, whereas it improves consistently with λ on Last.fm, peaking at λ = 1 (which corresponds, as pointed out in section 3.2, to the relevance-based version of IA-Select). We attribute this difference to the fact that the baseline performance on Last.fm is quite low, whereby diversifying involves a lower relevance loss (hence the optimum improvement with maximum diversification). MovieLens allows for better baseline performance, whereby moderating the diversity degree is more appropriate, and an extreme diversification results in a drastic accuracy loss. Overall the improvement respect to xQuAD is clear.

Table 2 shows results on further metrics, showing also the diversification of the popularity-based recommender baseline, in addition to pLSA. As in the experiments in search diversity, the λ parameter in xQuAD and RxQuAD is chosen to optimize for ERR-IA on each dataset. We see that our approach is consistently better in most cases. Only over pLSA in MovieLens we observe mixed results, with xQuAD producing better values on α-nDCG and nDCG-IA respectively, while RxQuAD is best on ERR-IA, and pure diversity –as measured by S-precision@r and S-recall. RxQuAD achieves clearer improvements on the popularity baseline. So it does on Last.fm for both baselines –except on nDCG-IA against popularity. This suggests that RxQuAD finds more room for improvement over the original algorithms on weak baselines (popularity recommendation vs. pLSA) and/or difficult datasets (Last.fm compared to MovieLens), with a low baseline effectiveness, whereas on a strong baseline run, there is no clear winner. We also observe that the IA-Select algorithm is not always effective on these recommendation tasks. We attribute this to the strong redundancy penalization of IA-Select (as we shall discuss later), which may involve a loss of relevant documents, particularly over a strong baseline like pLSA. The ineffectiveness is mostly observed in terms of α-nDCG, a metric which (by a default α = 0.5) applies a softer redundancy discount which IA-Select may mismatch. Furthermore, because of the $p(c|d)$ term, items with multiple aspects (which abound in Last.fm) are demoted by IA-Select, which explains some low subtopic recall values.

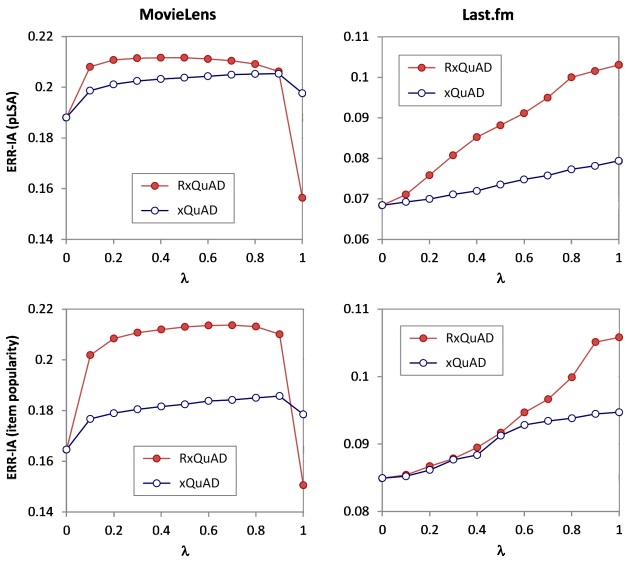

Figure 3. Comparative performance (measured by ERR-IA) of relevance-based diversification (RxQuAD) and xQuAD, ranging over values of λ. The algorithms are tested on the MovieLens 1M dataset (left), using movie genre as the query aspect space; and Last.fm (right) using social tags as information need aspects. pLSA is the baseline recommender system diversified in the two top graphics, and the popularity-based item recommender in the bottom ones.

5. RELEVANCE-BASED REDUNDANCY MANAGEMENT

Beyond the interest and potential advantages of the relevance-based diversification model as a stand-alone development, an explicit relevance model provides the basis for the introduction and derivation of further extensions on a formal probabilistic basis. We show this by extending our framework with an explicit model of the tolerance to redundancy: different tasks, or different users, introduce different conditions on how redundancy should be handled and penalized. We show next how this can be accounted for by a smooth generalization of our framework.

5.1 Tolerance to Redundancy

Let $stop$ denote, as in section 3.4, a binary random variable that is true when a user, in some retrieved document list browsing context, stops reading documents. And let $stop_S$ denote the fact that a user stops browsing at some point after reading some documents in a set S. We may refine the xQuAD diversity component as $p(r_d, \neg stop_S|q)$, where the marginal utility of the document d is defined in terms of the user stopping before reaching d. This results into a nuanced reformulation of the objective function:

$$g(d|q,c,S) = (1-\lambda)\, p(r|d,q) + \\ + \lambda \sum_c p(c|q)\, p(r|d,c,q) \prod_{d' \in S} \big(1 - p(stop|d',c,q)\big)$$

This form of the objective generalizes the original one by abstracting from the reasons why a document d –in the context of a particular ranking– would not add value to the effective utility of the result list.

Now, as we did in section 3.4 (equation 11, but here with further conditioning variables), we may marginalize the stopping probability with respect to relevance:

$$p(stop|d',c,q) = p(stop|d',c,q,r)\, p(r|d',c,q) + \\ + p(stop|d',c,q,\neg r)\, p(\neg r|d',c,q)$$

where again different simplifications can be considered. First, within the objective function for greedy document selection, we should consider $p(stop|d',c,q,\neg r) = 0$ for $d' \in S$, as the utility of the next document (which the objective function means to assess) would not be an issue if the user had stopped browsing already somewhere in S. Another reasonable simplification is to assume the user's decision to stop at a specific document only depends on finding relevance, i.e. $p(stop|d',c,q,r) \sim p(stop|r)$, whereby the model reduces to:

$$p(stop|d',c,q) \sim p(r|d',c,q)\, p(stop|r)$$

This way the original diversification algorithm is generalized to a form where an additional parameter $p(stop|r)$ represents the user tolerance to redundancy –or in some sense, how many documents it takes for the user to be satisfied:

$$g(d|q,c,S) = (1-\lambda)\, p(r|d,q) + \\ + \lambda \sum_c p(c|q)\, p(r|d,c,q) \prod_{d' \in S} \big(1 - p(r|d',c,q)\, p(stop|r)\big)$$

The introduction of this additional parameter allows to better match this characteristic of users and/or retrieval tasks. It allows to control (raise or soften) the penalization that should be applied to documents possessing aspects that are already covered in the original formulation of xQuAD. The latter implicitly assumes $p(stop|r) = 1$, that is, the user stops as soon as he finds a relevant document (zero tolerance to redundancy), which reflects again an implicit assumption that users are willing to select a single document –which is often not the case.

An equivalent parameter might be inserted in the original xQuAD formulation to soften redundancy penalization, but it would lack the formal justification that the relevance-based approach enables. Furthermore, the xQuAD redundancy penalization is already rather mild compared to RxQuAD, since the discounting term of the novelty component is based on document probabilities $p(d|c)$, which tend to range on much lower values (since they should sum to 1 over all documents covering an aspect) compared to a Bernoulli relevance distribution $p(r|d,c,q)$. The addition of a tolerance parameter to xQuAD would only make this worse –unless it ranged beyond [0,1], which would bring the scheme even farther from a formal probabilistic basis.

On the other hand, tolerance to redundancy has also been explicitly modeled and introduced in the context of metric formalization upon user models [5,8,15,24]. Therefore the use of this parameter in our diversification algorithm has the potential of a better optimization for such metrics by bringing the diversification model closer to the principles and assumptions which are built into the metrics.

5.2 Empirical Observation

In order to illustrate the effect of adjustable redundancy, we display as a heat map in Figure 4 the performance values of the generalized RxQuAD with different values of $p(stop|r)$, measured by α-nDCG with different values of α (also reflecting different degrees of redundancy tolerance). For this test, we select the TREC subtopics in the search task (with $p(r|k)$ estimated on relevance judgments), and the MovieLens dataset for the recommendation task. We keep the same values for λ as were selected in the previous experiments, and the pLSA baseline in the recommendation task. It can be observed that the redundancy penalization effect of $p(stop|r)$ is consistent with the equivalent parameter in the metric, i.e. the values evolve on a diagonal pattern: higher $p(stop|r)$ values in the algorithm perform better for higher α in the metric, and vice versa. The MovieLens graphic is smoother as the results are averaged over about 6,000 users (vs. 100 topics in TREC), and averaged again over 5 folds.

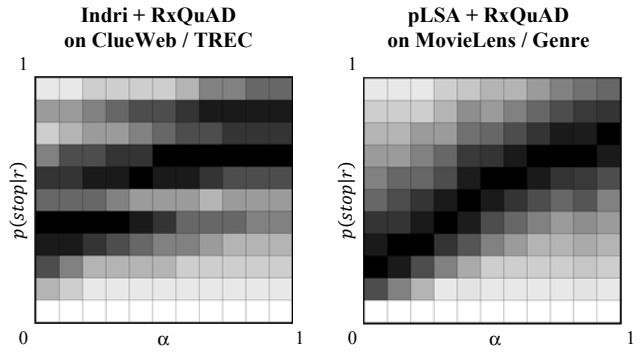

Figure 4. Parameterized tolerance to redundancy in the RxQuAD diversification algorithm by $p(stop|r)$, evaluated with corresponding metric configurations (α parameter in α-nDCG), by increments of 0.1. The values are displayed as a heat map where the darker colors (rank-normalized per column) represent higher α-nDCG values. The left map shows results for search diversity (over Indri) on ClueWeb with TREC subtopics, and the right map shows results recommendation diversity (over pLSA) on MovieLens with genres. A diagonal trend can be observed in the relative metric values.

6. DISCUSSION

In practical terms, using a relevance model in the redundancy component results in a higher redundancy penalization than is applied with a document generation model in xQuAD, as discussed in the previous section. Adding to one over all documents, the variations of $p(d|c)$ are quite small compared to $p(c|q)$, whereby subtopic coverage may tend to overdo redundancy in the overall effect of the algorithm. In contrast, the conditional Bernoulli relevance distribution $p(r|d,c,q)$ does not add to one in general over documents (unless the model assumes a unique relevant document), and may thus range over a significantly higher scale. Our relevance-oriented formulation hence enables a stronger redundancy penalization. This strength can be softened if needed, by the generalization for redundancy adaptation described in the previous section.

Compared to the original xQuAD, IA-Select goes to the opposite extreme in redundancy penalization: according to the approach described by Agrawal et al [1], redundancy is penalized by $p(c|d)$, which may have values close to 1 when documents cover a single aspect with high probability. The effect is that once an aspect is covered by some document in S, any other document covering this aspect is considered to add near zero marginal utility [28]. As a particular consequence, there is little discrimination between degrees of redundancy (i.e. an aspect that has been covered just once vs. further times before), and once all aspects have been covered, the diversified ranking degrades to the original order defined by the baseline retrieval system (see [26] for further analysis). Our relevance-oriented formulation does not result in such extremes either, as far as the probability of relevance, even at the high rank positions does usually not get as close to 1 as $p(c|d)$ may get. In practice, IA-Select may also soften the penalization by multiplying $p(c|d)$ by the baseline retrieval function. The penalization is also milder when documents cover several aspects, in which case $p(c|d)$ ranges over lower values. The algorithm implementation designer has no explicit control over these factors though –in contrast with our proposed extension for redundancy adjustment–, which may also account for the instability of IA-Select across heterogeneous experiments observed in section 4.

Besides these considerations, a relevance-based redundancy assessment better matches the structure of redundancy-sensitive metrics such has ERR-IA and α-nDCG, which are formalized in terms of probabilities of relevance [8]. This may partly account for the observed improved performance. The introduction of an explicit redundancy control parameter allows further gauging how aggressive the novelty-seeking component should be, enabling a finer adjustment to the role of redundancy in specific IR tasks and metrics.

Modeling tolerance to redundancy illustrates the potential advantages that an explicit relevance model brings about. Tolerance to redundancy could be introduced in the original formulations of IA-Select and xQuAD as well e.g. as a scalar parameter in the redundancy penalization component, but it is not clear how this might be given a principled –not simply heuristic– justification. As noted by Welch et al [28], "*common to a majority of prior research* [on search diversification] *is the single relevant document assumption*", which makes it difficult to explicitly formalize variable degrees of acceptable redundancy. The lack of this limitation might be credited as a virtue of an explicit relevance model. Though it has also been argued that a document generative model does not intrinsically negate the selection of several relevant documents [16], it is not clear how this multiple selection could be explicitly reflected upon a generative model. Welch et al [28] actually address this by explicitly modeling the number of relevant documents that the user is seeking to get. The alternative we show here has a considerably simpler development and does not require the introduction of this additional, somewhat artificial variable (the number of sought documents). Instead, our approach builds on models of user behavior, which are being extensively researched in the field (e.g. [8,15] among many other works).

A tradeoff of the relevance model is that it needs to be trained on relevance information –a tradeoff which is shared with PRP approaches in contrast to language modeling to build information retrieval systems. We have shown however that the need for training in our framework is not demanding in terms of the involved complexity, the data requirements, or the required accuracy, since a simple approach with few data proves to be good enough to produce quite competitive results. Furthermore, a rough rank-relevance estimate from a commercial search engine click log showed to be sufficient to obtain almost as good results as with expensive editorial relevance judgments.

7. CONCLUSION

IR diversification approaches proposed so far in the field use either an explicit representation of information need aspects, or an explicit relevance model, but not both. We have proposed and developed a revision of prior intent-oriented diversification schemes with the introduction of an explicit relevance model in the formulation of the approach. We observe that an explicit relevance model results in comparable or even better performance than prior approaches in terms of diversity evaluation metrics, in different application domains (search and recommendation) on different datasets. The approach thus favorably compares to its original alternatives, and might open new lines for effectiveness improvements.

From a theoretical standpoint, the relevance-oriented formulation provides an alternative –perhaps more direct– description of the diversity problem, whereupon the algorithmic scheme can be directly derived. The relevance-based foundation may be better suited for the description of diversification processes and their underlying principles: marginal utility, diminishing returns, relative value of documents, and so forth. Concepts such as relevance and utility find clear, unambiguous and more direct reflection in the framework itself. Furthermore, the formulation provides a more direct match of metric schemes in which relevance models underlie [5,8,15,24], therefore potentially providing for a better optimization against such metrics.

An additional side-effect of the relevance-oriented formulation is the unification of the IA-Select and xQuAD approaches into a common scheme. The proposed framework opens moreover new directions for further formal developments where relevance is an intrinsic variable. As a particular case, we show the formal extension of our framework to describe and adjust the algorithm to different degrees of tolerance to redundancy, the consistency of which is empirically validated.

8. ACKNOWLEDGMENTS

This work was supported by the national Spanish projects TIN2011-28538-C02-01 and S2009TIC-1542.

9. REFERENCES

[1] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. Diversifying search results. 2nd ACM International Conference on Web Search and Data Mining (WSDM 2009). Barcelona, Spain, February 2009, 5-14.

[2] Arampatzis, A. and Robertson, S. Modeling score distributions in information retrieval. Information Retrieval 14(1), 2011, 26-46.

[3] Bache, R., Baillie, M., and Crestani, F. Language models, probability of relevance and relevance likelihood. ACM International Conference on Information and Knowledge Management (CIKM 2007). Lisbon, Portugal, 2007, 853-856.

[4] Carbonell, J. G. and Goldstein, J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 1998). Melbourne, Australia, August 2998, 335-336.

[5] Carterette, B. An analysis of NP-completeness in novelty and diversity ranking. Information Retrieval 14(1), February 2011, 89-106.

[6] Celma, Ò. and Herrera, P. A New Approach to Evaluating Novel Recommendations. 2nd ACM International Conference on Recommender Systems (RecSys 2008). Lausanne, Switzerland, 2008, 179-186.

[7] Chen, H. and Karger, D. R. Less is More. 29th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 2006). Seattle, WA, USA, August 2006, 429-436.

[8] Clarke, C. L. A., Craswell, N., Soboroff, I., and Ashkan, A. A Comparative Analysis of Cascade Measures for Novelty and Diversity. 4th ACM International Conference on Web Search and Data Mining (WSDM 2011). Hong-Kong, China, February 2011, 75-84.

[9] Clarke, C. L. A., Craswell, N., Soboroff, I, and Cormack, G. V. Overview of the TREC 2010 Web Track. TREC 2010, Gaithersburg, MD, USA.

[10] Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. Novelty and diversity in information retrieval evaluation. 31st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 2008). Singapore, July 2008, 659-666.

[11] Cremonesi, P., Koren, Y., Turrin, R. Performance of recommender algorithms on top-n recommendation tasks. 4th ACM International Conference on Recommender Systems (RecSys 2010). Barcelona, Spain, September 2010, 39-46.

[12] Fernández, M., Vallet, D., and Castells, P. Using Historical Data to Enhance Rank Aggregation. 29th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 2006). Seattle, WA, USA, August 2006, pp. 643-644.

[13] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22, 1, 2004, 5-53.

[14] Hofmann, T. Latent semantic models for collaborative filtering. ACM Transactions on Information Systems 22(1), January 2004, 89-115.

[15] Hu, B., Zhang, Y., Chen, W., Wang, G., and Yang, Q. Characterizing Search Intent Diversity into Click Models. 20th International Conference on World Wide Web (WWW 2011). Hyderabad, India, March 2011, 17-26.

[16] Lafferty, L., Zhai, C. Probabalistic Relevance Models Based on Document and Query Generation. In Croft W.B. and Lafferty J. (Eds.), Language Modeling for Information Retrieval. Kluwer Academic Publishers, Dordrecht, Netherlands, 2003.

[17] Lee, J. H. Analyses of multiple evidence combination. 20th ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 1997). Philadelphia, PA, USA, July 1997, 267-276.

[18] Nottelmann, H. and Fuhr, N. From Retrieval Status Values to Probabilities of Relevance for Advanced IR Applications. Information Retrieval 6(3-4), September 2003, 363-388.

[19] Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., and Granka, L.A. In Google We Trust: Users' Decisions on Rank, Position, and Relevance. Journal of Computer-Mediated Communication 12(3), June 2007, 801-823.

[20] Robertson, S. E. The Probability Ranking Principle in IR. Journal of Documentation 33(4), 1977, 294-304.

[21] Santos, R. L. T., Macdonald, C., and Ounis, I. Exploiting query reformulations for web search result diversification. 19th International Conference on World Wide Web (WWW 2010). Raleigh, NC, USA, April 2010, 881-890.

[22] Santos, R. L. T., Macdonald, C., and Ounis, I. On the Role of Novelty for Search Result Diversification. Information Retrieval. In Press.

[23] Spark-Jones, K., Robertson, S. E., Hiemstra, D., Zaragoza, H. Language Modelling and Relevance. In Croft W.B. and Lafferty J. (Eds.), Language Modeling for Information Retrieval. Kluwer Academic Publishers, Dordrecht, Netherlands, 2003.

[24] Vargas, S. and Castells, P. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. 5th ACM Int. Conf. on Recommender Systems (RecSys 2011). Chicago, IL, October 2011, 109-116.

[25] Vargas, S., Castells, P., and Vallet, D. Intent-Oriented Diversity in Recommender Systems. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011). Beijing, China, August 2011, 1211-1212.

[26] Vargas. S., Castells, P., and Vallet, D. On the Suitability of Intent Spaces for IR Diversification. International Workshop on Diversity in Document Retrieval (DDR 2012) at the 5th ACM International Conference on Web Search and Data Mining (WSDM 2012). Seattle, WA, USA, February 2012.

[27] Wang, J. and Zhu, J. Portfolio theory of information retrieval. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009). Boston, MA, USA, July 2009, 115-122.

[28] Welch, M. J., Cho, J., and Olston, C. Search result diversity for informational queries. 20th International Conference on World Wide Web (WWW 2011). Hyderabad, India, March 2011, 237-246.

[29] Zhai, C., Cohen, W. W., and Lafferty, J. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003). Toronto, Canada, July 2003, 10-17.