Intent-Aware Search Result Diversification

Rodrygo L. T. Santos
rodrygo@dcs.gla.ac.uk

Craig Macdonald
craigm@dcs.gla.ac.uk

Iadh Ounis
ounis@dcs.gla.ac.uk

School of Computing Science
University of Glasgow
G12 8QQ Glasgow, UK

ABSTRACT

Search result diversification has gained momentum as a way to tackle ambiguous queries. An effective approach to this problem is to explicitly model the possible aspects underlying a query, in order to maximise the estimated relevance of the retrieved documents with respect to the different aspects. However, such aspects themselves may represent information needs with rather distinct intents (e.g., informational or navigational). Hence, a diverse ranking could benefit from applying intent-aware retrieval models when estimating the relevance of documents to different aspects. In this paper, we propose to diversify the results retrieved for a given query, by learning the appropriateness of different retrieval models for each of the aspects underlying this query. Thorough experiments within the evaluation framework provided by the diversity task of the TREC 2009 and 2010 Web tracks show that the proposed approach can significantly improve state-of-the-art diversification approaches.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Performance

Keywords

Web Search, Relevance, Diversity

1. INTRODUCTION

User queries often carry some degree of ambiguity [38]. On the one hand, genuinely ambiguous queries (e.g., ‘*zeppelin*’) have multiple *interpretations* (e.g., ‘*airship*’, ‘*band*’). On the other hand, even those queries with a single, clearly defined interpretation might still be underspecified, as it is not clear which *aspects* of this interpretation the user is actually interested in (e.g., ‘*led zeppelin*’... ‘*website*’? ‘*downloads*’? ‘*biography*’? ‘*albums*’? ‘*reunion*’?) [13].

Search result diversification has recently gained attention as a means to tackle query ambiguity. Instead of trying to identify the ‘correct’ interpretation behind a query, the idea is to diversify the search results, in the hope that different users will find at least one of these results to be relevant to their information need [1]. Differently from the traditional assumption of independent document relevance [30], diversification approaches typically consider the relevance of a document in light of the other retrieved documents. For instance, once users have found the information they are seeking, it is questionable whether documents with the same (or very similar) information will still be of any use [13].

An effective approach to diversifying search results is to *explicitly* account for the various aspects[1] underlying an ambiguous query [1, 7, 33, 36]. By doing so, the problem becomes to select a ranking of documents that collectively provide the maximum relevance with respect to these different aspects. In a real scenario, however, the actual aspects of a query are not known, nor is the relevance of each retrieved document to each query aspect determined with certainty. Moreover, the relevance of a document to a query aspect may depend on the *intent*[2] underlying this aspect (e.g., informational or navigational [4, 31]). Additionally, different aspects could feasibly represent information needs with different intents. For instance, ‘*website*’, ‘*downloads*’, and ‘*biography*’ arguably represent navigational, transactional, and informational aspects of the query ‘*led zeppelin*’, respectively. Similarly, ‘*albums*’ could exemplify a typed query aspect, representing an information need for a list of entities. In the same vein, ‘*reunion*’ might denote a question-answering aspect of the query ‘*led zeppelin*’, regarding whether the legendary rock band have any plans of reuniting in the near future.

Queries with different intents have been shown to benefit from different retrieval models [22]. Likewise, we hypothesise that explicit diversification approaches may benefit from accounting for the intents of different query aspects. For instance, relevance estimations with respect to the ‘*website*’ aspect of the query ‘*led zeppelin*’ could be arguably improved by applying a retrieval model suitable for navigational queries. In this paper, we propose a novel diversification approach, aimed at learning the appropriateness of

SIGIR'11, July 24–28, 2011, Beijing, China.
Copyright 2011 ACM 978-1-4503-0757-4/11/07 ...$10.00.

[1] Unless otherwise noted, through the rest of this paper, we refer to query ‘*interpretations*’ and ‘*aspects*’ indistinctly.

[2] Agrawal et al. [1] use ‘*intents*’ in the sense of what we call ‘*interpretations*’. We believe our choice is more appropriate in light of the established nomenclature in the literature.

multiple intent-aware retrieval models for each aspect. As a result, the relevance of a document to multiple aspects—i.e., its diversity—can be estimated more effectively.

We thoroughly evaluate our approach in the context of the diversity task of the TREC 2009 and 2010 Web tracks [11, 12]. In particular, we investigate learning strategies that either select the most appropriate retrieval model or merge multiple retrieval models for each query aspect. The results of our investigations attest the effectiveness of both strategies within our proposed intent-aware approach for diversifying Web search results, with significant improvements on top of state-of-the-art diversification approaches.

The remainder of this paper is organised as follows. Section 2 describes related work on search result diversification and search intents. Section 3 further details our main contributions. Section 4 describes our approach for leveraging intent-aware models for diversification. Sections 5 and 6 detail the experimental setup and the evaluation of our approach. Finally, Section 7 presents our conclusions.

2. RELATED WORK

In this section, we provide background on the search result diversification problem and related approaches to this problem. We then set the grounds for our proposed approach by reviewing related work on the use of search intents.

2.1 Search Result Diversification

Most of the existing diversification approaches are somehow inspired by the work of Carbonell and Goldstein [5]. The basic idea of their Maximal Marginal Relevance (MMR) method is to iteratively select a document with the highest similarity to the query and lowest similarity to the already selected documents, in order to promote novelty. Subsequent implementations of this idea include the approach of Zhai et al. [41] to model relevance and novelty within a risk minimisation framework. In particular, they promote documents with highly divergent language models from those of the already selected documents. Chen and Karger [10] proposed a probabilistic approach to the related problem of finding at least one relevant result for a given query, by choosing documents under the assumption that those already chosen are not relevant to the query. More recently, Wang and Zhu [39] proposed to diversify a document ranking as a means to reduce the risk of overestimating its relevance. In their work, two documents are compared based on the correlation of their relevance scores.

By assuming that similar documents will cover similar aspects, the aforementioned approaches only consider the aspects underlying a query *implicitly*. An alternative approach consists of *explicitly* modelling these aspects [36]. For instance, Agrawal et al. [1] proposed the IA-Select algorithm for search result diversification. It employs a classification taxonomy over queries and documents to iteratively promote documents that share a high number of classes with the query, while demoting those documents with classes already well represented in the ranking. Similarly, Carterette and Chandar [7] proposed a probabilistic approach to maximise the coverage of the retrieved documents with respect to the aspects of a query, by modelling these aspects as topics identified from the top ranked documents. Recently, Santos et al. [33] introduced the xQuAD probabilistic framework for search result diversification, which explicitly represents different query aspects as 'sub-queries'. They defined a diversification objective based on the estimated relevance of documents to multiple sub-queries, as well as on the relative importance of each sub-query in light of the initial query.

Since our goal is to produce intent-aware relevance estimations given an explicit representation of query aspects, our approach is also set in the context of explicit diversification. Accordingly, in Section 6, we use both IA-Select [1] and xQuAD [33] as a basis for evaluating our approach. In particular, these two approaches represent the state-of-the-art in explicit search result diversification.

2.2 Intents in Information Retrieval

Different information retrieval tasks have benefited from taking into account the intent of a query (e.g., informational, navigational, or transactional [4, 31]). These approaches can be generally categorised based on whether or not they rely on the classification of queries into predefined intents.

Query intent detection approaches first classify a query with respect to a predefined set of intents. A retrieval model specifically trained for the predicted intent is then applied to retrieve documents for the query. For instance, Kang and Kim [22] showed that queries of different intents can benefit from the application of intent-specific retrieval models. A major shortcoming of this approach, however, is the limited accuracy of existing intent detection mechanisms [17].

Instead of classifying a query into a predefined target intent, an alternative is to identify similar queries from a training set, and then to apply a retrieval model appropriate for this set. This approach has an advantage over a classification of queries based on a fixed set of intents, as queries of the same intent often benefit from different retrieval models [17]. For example, Geng et al. [20] proposed an instance-based learning approach using k-nearest neighbour (k-NN) classification to improve Web search effectiveness. In their approach, a k-NN classifier is used to identify training queries similar to an unseen query. A retrieval model is then learned based on the identified queries and applied to the unseen query. A more general approach was proposed by Peng et al. [27]. In their work, multiple ranking functions are chosen from a pool of candidate functions, based on their performance on training queries similar to an unseen query.

Our approach is similar in spirit to the approaches of Kang and Kim [22], Geng et al. [20], and Peng et al. [27]. However, while these approaches focused on inferring the intent of a *query*, we target the problem of inferring the intent underlying different *aspects* of this query. Besides this difference in granularity, our intent-aware approach tackles a different search scenario, namely, search result diversification.

In a similar vein, Santos et al. [34] proposed a selective diversification approach, aimed at tailoring a diversification strategy to the ambiguity level of different queries. In particular, given an unseen query, their approach learns a trade-off between relevance and diversity, based on optimal trade-offs observed for similar training queries. As a result, their approach effectively determines when to diversify the results for an unseen query, and also by how much. Our proposed approach also differs from the approach of Santos et al. [34], in that ours focuses on selecting appropriate retrieval models for different query *aspects*, as opposed to the query itself. More importantly, their approach is orthogonal to ours. In essence, instead of determining *when* to diversify the results for a given query, we tackle the problem of *how* to diversify these results given the identified aspects of this query.

3. CONTRIBUTIONS OF THIS PAPER

The major contributions of this paper are:

1. A novel aspect intent-aware diversification approach, aimed at predicting, for each identified query aspect, the appropriateness of different retrieval models.

2. A thorough evaluation of the proposed approach within the standard experimentation paradigm of the diversity task of the TREC 2009 and 2010 Web tracks.

4. INTENT-AWARE SEARCH RESULT DIVERSIFICATION

As discussed in Sections 1 and 2, different aspects of an ambiguous query can have rather different intents. To illustrate this, consider topic #1 from the diversity task of the TREC 2009 Web track [11], as shown in Figure 1. In this example, different aspects of the query '*obama family tree*' are represented as a set of sub-topics, identified from the query log of a commercial search engine. Moreover, these sub-topics represent aspects with an informational ('inf') or a navigational ('nav') intent, as judged by TREC assessors.

```
<topic number="1" type="faceted">
  <query>obama family tree</query>
  <description>
    Find information on President Barack Obama's family history,
    including genealogy, national origins, places and dates of
    birth, etc.
  </description>
  <subtopic number="1" type="nav">
    Find the TIME magazine photo essay "Barack Obama's Family
    Tree".
  </subtopic>
  <subtopic number="2" type="inf">
    Where did Barack Obama's parents and grandparents come from?
  </subtopic>
  <subtopic number="3" type="inf">
    Find biographical information on Barack Obama's mother.
  </subtopic>
</topic>
```

Figure 1: TREC 2009 Web track, topic #1, along with its corresponding sub-topics.

Inspired by related work in Section 2.2, we hypothesise that diversification approaches can benefit from retrieval models targeted to the intents of different query aspects. For instance, for the query exemplified in Figure 1, a diversification approach could leverage a *navigational intent-aware model* for the first query aspect, and an *informational intent-aware model* for the second and third aspects.

In this work, we propose a supervised learning approach for estimating the appropriateness of multiple intent-aware retrieval models for each query aspect. Given a query q, our goal is to maximise the diversity of the retrieved documents with respect to the aspects underlying this query. Without loss of generality, following an explicit diversification strategy, we can quantify the diversity of a document d given a query q and the other retrieved documents $\mathcal{S}$ as the expected relevance of d with respect to the aspects of q, denoted $\mathcal{A}(q)$:

$$\mathrm{P}(d|\mathcal{S},q) = \sum_{a\in\mathcal{A}(q)} \mathrm{P}(a|q)\,\mathrm{P}(d|\mathcal{S},q,a), \tag{1}$$

where $\mathrm{P}(a|q)$ captures the relative importance of each aspect a given the query q, and $\mathrm{P}(d|\mathcal{S},q,a)$ denotes the probability of the document d being relevant to this aspect, given how well the documents in $\mathcal{S}$ already satisfy this aspect.

Equation (1) can be seen as a canonical formulation of the objective functions deployed by different explicit diversification approaches in the literature [34]. In particular, these approaches differ primarily in how they represent the set of aspects associated with a query, and in how they estimate the relevance of each document to every identified query aspect. For instance, Agrawal et al. [1] rely on the top-level categories from the Open Directory Project (ODP)[3] for representing query and document classes, and integrate relevance and classification scores for ranking documents. Santos et al. [33] exploit query reformulations from commercial search engines as representations of the different aspects of a query, and directly estimate the relevance of documents to these aspects. Our proposed approach is agnostic to any particular mechanism for generating explicit query aspect representations. Indeed, our only assumption is that different aspects may convey different user intents.

In the interest of keeping the description of our approach general, in the remainder of this section, we adopt an abstract view of aspects and intents.[4] In particular, to formalise our approach, we further derive Equation (1), by marginalising $\mathrm{P}(d|\mathcal{S},q,a)$ over a target set of intents $\mathcal{I}$:

$$\mathrm{P}(d|q,\mathcal{S}) = \sum_{a\in\mathcal{A}(q)} \mathrm{P}(a|q) \sum_{i\in\mathcal{I}} \mathrm{P}(i|a)\,\mathrm{P}(d|\mathcal{S},q,a,i), \tag{2}$$

where $\mathrm{P}(i|a)$ denotes the probability that the aspect a of the initial query q conveys the intent i. Accordingly, $\mathrm{P}(d|\mathcal{S},q,a,i)$ denotes the relevance of the document d in light of the other retrieved documents $\mathcal{S}$, the query q, the aspect a, and the intent i. Once again, without loss of generality, assuming that different aspects are equally probable (i.e., $\mathrm{P}(a|q) = \frac{1}{|\mathcal{A}(q)|}, \forall a \in \mathcal{A}(q)$),[5] our task becomes two-fold:

1. To infer the probability $\mathrm{P}(i|a)$ of each intent $i \in \mathcal{I}$ given a query aspect $a \in \mathcal{A}(q)$;

2. To learn an appropriate retrieval model $\mathrm{P}(d|\mathcal{S},q,a,i)$ for each predicted intent $i \in \mathcal{I}$.

In Section 4.1, we propose a classification approach for the first task. For the second task, as described in Section 4.2, we resort to learning-to-rank.

4.1 Inferring the Query Aspect Intents

In order to infer the probability of different intents for a query aspect, we propose a linear classification approach. In particular, given a query aspect a, our goal is to estimate the probability of an intent $i \in \mathcal{I}$ as:

$$\mathrm{P}(i|a) = f(\mathbf{w}\cdot\mathbf{x}_a), \tag{3}$$

where $\mathbf{x}_a$ is a feature vector representing the aspect a, and $\mathbf{w}$ is a weight vector, which is learned from labelled training data. The function f maps the dot product of the weight and feature vectors into the desired prediction outcome. Alternative regimes for instantiating the function f are described in Section 4.1.1. Section 4.1.2 describes our choices for labelling training data. Lastly, Section 4.1.3 defines the classification space considered in this paper, and describes the query aspect features leveraged for this classification task.

[3] http://www.dmoz.org

[4] A concrete instantiation of query aspects and their possible intents is discussed in Section 4.1.3.

[5] For alternatives on how to estimate the likelihood of different query aspects, we refer the reader to [33, 36].

4.1.1 Classification Regimes

We propose two alternative regimes for instantiating the function f in Equation (3): *model selection* and *model merging*. The model selection regime employs a hard classification approach [40]. In particular, this approach treats different intents as mutually exclusive, hence assigning each aspect a single (i.e., the most likely) intent. For instance, for a target set of intents $\mathcal{I} = \{i_1, i_2, i_3\}$, a possible selection outcome could be: $\mathrm{P}(i_1|a) = 1, \mathrm{P}(i_2|a) = 0, \mathrm{P}(i_3|a) = 0$. In this example, the aspect a would be associated with its most likely intent, i_1, and only the retrieval model $\mathrm{P}(d|\mathcal{S}, q, a, i_1)$ would have an impact on the estimated relevance of document d to the aspect a. This classification regime resembles the selective retrieval approaches described in Section 2.2, except that the most appropriate model is selected at the aspect level (as opposed to the query level).

Our second regime, model merging, provides a relaxed alternative to model selection. In particular, it deploys a soft classification approach, in order to obtain a full probability distribution over the considered intents [40]. For the above example, a possible outcome of the model merging classification could be $\mathrm{P}(i_1|a) = 0.6, \mathrm{P}(i_2|a) = 0.3, \mathrm{P}(i_3|a) = 0.1$. In this case, the estimated relevance of a document d to the aspect a would be determined by a linear combination:

$$\begin{aligned}\mathrm{P}(d|\mathcal{S}, q, a) = {} & 0.6 \times \mathrm{P}(d|\mathcal{S}, q, a, i_1) \\ & + 0.3 \times \mathrm{P}(d|\mathcal{S}, q, a, i_2) \\ & + 0.1 \times \mathrm{P}(d|\mathcal{S}, q, a, i_3).\end{aligned}$$

Different classifiers can be deployed to implement both the model selection and model merging regimes. Further details about the specific classifiers that enable both regimes in our investigations are provided in Section 5.3.

4.1.2 Classification Labels

In order to determine the ground-truth intent for different query aspects, we investigate two alternatives. The first one is based on the direct judgement by humans, who base their assessment solely on the observed aspects. However, the differences between query aspects may go beyond their apparent characteristics. For instance, aspects with the same judged intent could still benefit from leveraging different retrieval models [17]. Additionally, judging the intent of different aspects may be costly for large training datasets.

To overcome these limitations, we propose a second alternative for automatically labelling training aspects. Given a training query q with aspects $\mathcal{A}(q)$ and a set of target intents $\mathcal{I}$, with $|\mathcal{A}| = k$ and $|\mathcal{I}| = p$, we devise an oracle selection mechanism. In particular, this oracle mechanism always chooses the best out of the p^k possible selections for the k aspects of q, according to a diversity evaluation metric (e.g., ERR-IA [9], or any of the metrics described in Section 5.4). Although estimating this oracle may be infeasible for large values of k, it can be easily estimated for most practical settings. For instance, the maximum number of aspects per query in the TREC 2009 and 2010 Web tracks is $k = 8$. Moreover, if many more aspects were available for a particular query, less plausible aspects could be discarded without much loss. Indeed, this is precisely what leading Web search engines do when displaying only the top suggestions for a user query, which have been shown to deliver an effective diversification performance [33]. Finally, it is worth noting that this entire labelling process is conducted offline.

4.1.3 Classification Features

So far, we have intentionally described our approach without a strict definition of the classification space. This abstract view of classification instances as query aspects demonstrates the generality of our proposed approach, and its applicability to different explicit diversification approaches in the literature (e.g., [1, 7, 33, 36]). In particular, our proposed approach is not bound to any particular query aspect representation. In fact, any aspect representation that portrays the multitude of information needs underlying an ambiguous query [35] is potentially applicable, as different information needs can convey different user intents.

Nonetheless, to enable our investigations in Section 6, we follow Santos et al. [33, 36] and adopt a concrete representation of query aspects as '*sub-queries*'. In particular, a sub-query can be seen as a keyword-based representation of the information need expressed by a query aspect. In our experiments, we consider two mechanisms for generating sub-queries, as described in Section 5.1. Additionally, limited by the TREC Web test collection used in our experiments [11, 12], we restrict the space of target intents to navigational and informational ones. Based on this representation of aspects and intents, and inspired by research on related query analysis tasks, we devise a large feature set for sub-query intent classification. In particular, these include features computed from the words in the sub-query itself, as well as from the top documents retrieved for this sub-query. In total, we devise 838 features, based on 21 different feature classes. These features are described on the left side of Table 1, and organised into three groups:

Query log features (LOG). Query logs provide valuable evidence for discriminating between informational and navigational intents. To exploit such evidence, we compute several sub-query features based on the 15-million query MSN Search 2006 Query Log. For instance, we count the raw frequency of sub-queries, as navigational sub-queries are generally more popular than informational ones. Likewise, informational sub-queries intuitively require more effort from the users while inspecting the retrieved results. We quantify this in terms of the number of examined results and the time spent in doing so, as well as the click entropy [15].

Query performance predictors (QPP). The intent of a sub-query may be reflected not only on the sub-query itself, but also on the documents retrieved for this sub-query. For instance, a low coherence of the top-retrieved documents could indicate a sub-query with an informational intent. This, in turn, can reflect on the performance of this sub-query when used in a retrieval system. To exploit this intuition, we build upon a large body of research on query performance prediction [6] and leverage both pre- and post-retrieval predictors as sub-query features. In particular, the former are solely based on statistics of the sub-query terms, while the latter also leverage information from the documents retrieved for the sub-query.

Taxonomy-based features (TAX). Informational needs are intuitively broader than navigational ones, in terms of the concepts they cover. To quantify this intuition, we devise different features based on concepts from two different taxonomies derived from Wikipedia: categories and named entities. For the latter, we consider entities of four types: people, organisations, products, and locations. In partic-

Sub-Query Features				Document Features			
Group	Feature Class	Description	Total	Group	Feature Class	Description	Total
LOG	ClickCount	No. of clicks	120	WM	BM25	BM25 score [25]	5
LOG	ClickEntropy	URL-level click entropy [15]	3	WM	DPH	DPH score [25]	5
LOG	HostEntropy	Host-level click entropy	2	WM	LM	LM score (Dirichlet) [25]	5
LOG	QueryFrequency	No. of occurrences	4	WM	PL2	PL2 score [25]	5
LOG	ResultCount	No. of examined results	3	WM	MQT	No. of matching query terms	5
LOG	SessionDuration	Session duration (in sec.)	3	FM	BM25F	BM25F score [25]	1
QPP	AvICTF	Pre-retrieval predictor [21]	1	FM	PL2F	PL2F score [25]	1
QPP	AvIDF	Pre-retrieval predictor [21]	1	DM	MRF	MRF proximity score [25]	8
QPP	AvPMI	Pre-retrieval predictor [21]	1	DM	pBiL	DFR pBiL proximity score [25]	8
QPP	ClarityScore	Post-retrieval predictor [19]	30	LA	Absorbing	Absorbing Model score [28]	1
QPP	EnIDF	Pre-retrieval predictor [21]	1	LA	Edgerecip	No. of reciprocal links [3]	1
QPP	Gamma	Pre-retrieval predictor [21]	2	LA	Inlinks	No. of inlinks	1
QPP	QueryDifficulty	Post-retrieval predictor [2]	30	LA	Outlinks	No. of outlinks	1
QPP	QueryFeedback	Post-retrieval predictor [42]	30	LA	InvPageRank	PageRank transposed score	2
QPP	QueryScope	Pre-retrieval predictor [21]	1	LA	PageRank	PageRank score [26]	1
QPP	TermCount	No. of terms	1	SP	SpamFusion	Spam likelihood [16]	1
QPP	TokenCount	No. of tokens	1	URL	URLDigits	No. of digits in domain and host	2
TAX	ConceptCosine	Cosine over concepts [37]	4	URL	URLComponents	No. of host/path/query components	3
TAX	ConceptCount	No. of concepts [34]	360	URL	URLLength	Length of host/path/query string	3
TAX	ConceptEntropy	Entropy over concepts [37]	120	URL	URLType	Root, subroot, path, file	1
TAX	DisambSenses	No. of disamb. senses [32]	120	URL	URLWiki	Whether URL is from Wikipedia	1
GRAND TOTAL			838				61

Table 1: All features used in this work. Sub-query features (left side) are used for inferring the intents of different query aspects. Document features (right side) are used to produce intent-aware learned models.

ular, we represent the documents retrieved for each sub-query in the space of the concepts from these different taxonomies.[6] Based on this representation, we compute various distributional features, such as the average number of retrieved concepts, the average distance between pairs of documents, and the concept entropy of the entire retrieved list. Additionally, we also quantify the number of ambiguous entities among the top documents retrieved for a sub-query. Our intuition is that the presence of such entities further indicates the broadness of the sub-query [37].

Most of these features are extracted in multiple variants. For instance, retrieval-based features are computed based on five different approaches, as implemented by the Terrier IR platform [25]: Okapi BM25, the Divergence From Randomness (DFR) DPH and PL2 models, a language modelling (LM) approach with Dirichlet smoothing, and a count of the number of matching query terms. Additionally, these features are estimated at six rank cutoffs: 1, 3, 5, 10, 50, and 100. Finally, distributional features (e.g., the number of concepts across the retrieved documents) are summarised using up to four different statistics: mean, standard deviation, median, and maximum. Altogether, these amount to the grand total of 838 features.

4.2 Learning Intent-Aware Retrieval Models

In Section 4.1.1, we proposed two regimes for inferring an intent distribution $\mathrm{P}(i|a)$ for each aspect a. In this section, we propose a learning-to-rank approach for producing suitable intent-aware retrieval models for each intent of a.

4.2.1 Model Learning

In order to produce an intent-aware model $\mathrm{P}(d|\mathcal{S}, q, a, i)$ for each intent i underlying the aspect a, we once again resort to machine learning. In particular, we deploy a large set of document features, and leave it to a learning-to-rank algorithm to generate retrieval models optimised for different intents. To achieve this goal, each model is learned using the entire feature set, but with a different training set of queries for each target intent. Given the intents considered in our investigation (i.e., informational and navigational), we use two intent-targeted query sets from the TREC 2009 Million Query track [8]. The first set contains 70 informational queries and the second set contains 70 navigational queries, as judged by TREC assessors. As a learning algorithm, we use Metzler's Automatic Feature Selection (AFS) [24]. In particular, AFS learns effective ranking models by directly optimising an IR evaluation metric. In our experiments, it is deployed to optimise mean average precision (MAP).

4.2.2 Document Features

To enable the generation of effective intent-aware retrieval models, we deploy a total of 61 document features, summarised on the right portion of Table 1, and organised into six groups: standard weighting models (WM), field-based models (FM), term dependence models (DM), link analysis (LA), spam (SP), and URL features. As these are all standard features traditionally used in the learning-to-rank literature [29],[7] we refer the interested reader to the descriptions and pointers provided in the table. In particular, each feature is computed for a sample of 5000 documents retrieved by the DFR DPH weighting model for each query. Standard weighting models and term dependence models are deployed with their commonly suggested parameter settings in the literature. Field-based models are trained through simulated annealing [23]. The remaining (query-independent) features are optimised using FLOE [18]. Finally, all feature scores are normalised to lie between 0 and 1 for each query.

Table 2 lists the top 10 features selected by AFS for each of our produced intent-aware models. For each feature, we also show its attained performance in terms of MAP when com-

[6] Category features are computed from documents retrieved from Wikipedia for each sub-query, while entity features are based on documents retrieved from the target collection.

[7] We leave the investigation of features that exploit the dependencies between d and the documents in $\mathcal{S}$ for the future.

	Informational		Navigational	
	Feature	MAP	Feature	MAP
1	DPH	0.2614	DPH	0.2110
2	URLDigits	0.2752	MRF (body)	0.2273
3	PL2 (title)	0.2819	BM25 (title)	0.2408
4	BM25F	0.2915	URLWiki	0.2517
5	pBiL (body)	0.2963	MQT	0.2592
6	pBiL (anchor)	0.2985	URLLength	0.2629
7	Edgerecip	0.3001	Absorbing	0.2666
8	LM (title)	0.3010	InvPageRank	0.2695
9	MQT (body)	0.3017	Inlinks	0.2718
10	MQT	0.3026	pBiL (body)	0.2738

Table 2: Top 10 selected features in the two intent-aware retrieval models used in this paper.

bined with the features selected before it. From the table, we observe that the top selected features are generally intuitive. For instance, DPH (which is used to generate the initial sample of documents for learning) is the top feature for both models. Likewise, as expected, various URL and link analysis features (e.g., URLWiki, URLLength, Absorbing, InvPageRank, Inlinks) are ranked high in the navigational model. Besides producing intuitive intent-aware models, we believe that our data-driven approach based on a large set of features provides a more robust alternative to hand-picking features traditionally associated with a particular intent.

4.3 Summary

In this section, we have introduced a novel supervised learning approach for diversifying the search results in light of the intents of different aspects of an ambiguous query. To enable our investigations in Section 6, we have instantiated our intent-aware diversification approach in light of two target aspect intents, namely, informational and navigational. Given these intents, we have described large feature sets for both inferring the intent distribution of different aspects (Section 4.1), as well as for learning the corresponding intent-aware retrieval models (Section 4.2). Although the choice of appropriate feature sets naturally depends on how learning instances (i.e., aspects) and labels (i.e., intents) are represented [40], it is worth reiterating that our approach is agnostic to these representations. While instantiating it for a different aspect representation or a different set of intents may require devising different features, no modification to the approach itself would be necessary. Moreover, although motivated by the learning tasks at hand, both feature sets in Table 1 comprise features deployed for a variety of different purposes in the literature. As a result, we believe they might be useful for deploying our approach with target intents beyond the two considered in our current investigations.

5. EXPERIMENTAL SETUP

In the next section, we investigate our intent-aware diversification approach proposed in Section 4. In particular, we aim to answer two main research questions:

1. Can we improve diversification performance with our intent-aware *model selection* regime?

2. Can we improve diversification performance with our intent-aware *model merging* regime?

These questions are investigated in Sections 6.1 and 6.2, respectively. In the remainder of this section, we detail the experimental setup that supports these investigations.

5.1 Collection, Queries, and Sub-Queries

Our analysis is conducted within the standard experimentation paradigm provided by the diversity task of the TREC 2009 and 2010 Web tracks [11, 12]—henceforth denoted WT09 and WT10 tasks, respectively. As a document collection, we consider the category-B ClueWeb09 dataset, as used in these tasks. This collection comprises 50 million English documents, aimed to represent the first tier of a commercial search engine index. In our experiments, we index this collection using Terrier [25], after applying Porter's English weak stemmer and without removing stopwords.

The WT09 and WT10 tasks comprise 50 and 48 queries, respectively. As mentioned in Section 4.1.3, for each of these 98 queries, we generate two sets of sub-queries, in order to provide alternative aspect representations for our investigations. The first sub-query set is based on the official sub-topics identified by TREC assessors for each of these queries. In particular, each WT09 query has an average of 3.54 informational and 1.32 navigational aspects, as judged by TREC assessors. For the WT10 queries, these numbers become 2.84 and 1.50, respectively. As TREC only provides a natural language description for each sub-topic, we obtain a shorter, keyword-like version using Amazon's Mechanical Turk. This step was necessary to make these sub-topics better resemble real Web search requests, so as to enable their matching in our query log. Note that this procedure by no means interfere with our conclusions, as these keyword-like sub-topics are uniformly deployed for all tested approaches.

Using the official TREC Web track sub-topics as a sub-query set has two main advantages. Firstly, as discussed in Section 4.1.2, they provide judged intent labels for each sub-query, which can be contrasted to our proposed performance-oriented labelling of training data. Secondly and most important, they provide a controlled environment for evaluating the effectiveness of our approach while isolating the impact of any particular aspect representation. In addition to this 'ground-truth' sub-query set, we also evaluate our approach using an alternative sub-query set. Following Santos et al. [33], for each of the 98 queries, we obtain up to 13 query suggestions from a commercial search engine.

5.2 Diversification Approaches

In Section 6, we apply our intent-aware model selection and model merging regimes to two diversification approaches: IA-Select [1] and xQuAD [33]. In particular, both approaches instantiate the general explicit diversification objective described in Equation (1), and hence can directly leverage our intent-aware aspect relevance estimations. Additionally, as discussed in Section 2.1, these approaches are representative of the state-of-the-art in search result diversification. Indeed, a variant of xQuAD was among the top performing approaches in the diversity task of both TREC 2009 and 2010 [11, 12]. In our investigations, both IA-Select and xQuAD diversify the top 1000 documents retrieved by DPH.

5.3 Classification Approaches

In Section 4.1, we introduced two regimes for leveraging the inferred intents of different aspects: model selection and model merging. The model selection regime builds upon a hard classification of intents. To enable this regime, we deploy two alternative classifiers. Firstly, we train a support vector machine (SVM) classifier with a polynomial kernel through a sequential minimal optimisation [40]. Our second

		Regime	ERR-IA@20	α-nDCG@20	NRBP	MAP-IA
	DPH		0.1607	0.2097	0.1318	0.0442
WT09 sub-topics	+IA-Select	Uni(inf)	0.2020	0.2472	0.1733	0.0634
	+IA-Select	Uni(nav)	0.2155	0.2634	0.1885	0.0652
	+IA-Select	Sel(ora,judg)	**0.2166** (7.2,0.5)	**0.2623** (6.1,-0.4)	**0.1859** (7.3,-1.4)	0.0641 (1.1,-1.7)
	+IA-Select	Sel(log,judg)	0.2090 (3.5,-3.0)	0.2548 (3.1,-3.3)	0.1799 (3.8,-4.6)	0.0657 (3.6,0.8)
	+IA-Select	Sel(svm,judg)	0.2126 (5.2,-1.3)	0.2594 (4.9,-1.5)	0.1834 (5.8,-2.7)	**0.0664** (4.7,1.8)
	+IA-Select	Sel(ora,perf)	**0.2782** (37.7▲,29.1▲)	**0.3134** (26.8▲,19.0▲)	**0.2562** (47.8▲,35.9▲)	**0.0724** (14.2▲,11.0▲)
	+IA-Select	Sel(log,perf)	0.2394 (18.5△,11.1△)	0.2780 (12.5△,5.5)	0.2155 (24.4△,14.3△)	0.0688 (8.5▲,5.5)
	+IA-Select	Sel(svm,perf)	0.2289 (13.3,6.2)	0.2713 (9.7,3.0)	0.2036 (17.5,8.0)	0.0677 (6.8△,3.8)
	+xQuAD	Uni(inf)	0.1945	0.2423	0.1664	0.0565
	+xQuAD	Uni(nav)	0.2219	0.2661	0.1954	0.0665
	+xQuAD	Sel(ora,judg)	0.2131 (9.6,-4.0)	**0.2602** (7.4△,-2.2)	0.1826 (9.7,-6.6)	0.0643 (13.8△,-3.3)
	+xQuAD	Sel(log,judg)	0.2094 (7.7,-5.6)	0.2542 (4.9,-4.5)	0.1801 (8.2,-7.8)	0.0631 (11.7,-5.1)
	+xQuAD	Sel(svm,judg)	**0.2153** (10.7,-3.0)	0.2594 (7.1,-2.5)	**0.1868** (12.3,-4.4)	**0.0671** (18.8,0.9)
	+xQuAD	Sel(ora,perf)	**0.2650** (36.2▲,19.4▲)	**0.3025** (24.8▲,13.7▲)	**0.2413** (45.0▲,23.5▲)	0.0688 (21.8▲,3.5▲)
	+xQuAD	Sel(log,perf)	0.2395 (23.1▲,7.9△)	0.2784 (14.9▲,4.6△)	0.2153 (29.4▲,10.2△)	0.0689 (21.9△,3.6)
	+xQuAD	Sel(svm,perf)	0.2370 (21.9▲,6.8)	0.2754 (13.7△,3.5)	0.2125 (27.7▲,8.8)	**0.0694** (22.8▲,4.4)
Query suggestions	+IA-Select	Uni(inf)	0.1944	0.2378	0.1676	0.0553
	+IA-Select	Uni(nav)	0.1978	0.2503	0.1662	0.0610
	+IA-Select	Sel(ora,perf)	**0.2736** (40.7▲,38.3▲)	**0.3083** (29.6▲,23.2▲)	**0.2518** (50.2▲,51.5▲)	**0.0687** (24.2▲,12.6▲)
	+IA-Select	Sel(log,perf)	0.2063 (6.1,4.3)	0.2548 (7.1,1.8)	0.1756 (4.8,5.7)	0.0631 (14.1▲,3.4△)
	+IA-Select	Sel(svm,perf)	0.2085 (7.3,5.4)	0.2560 (7.7,2.3)	0.1790 (6.8,7.7)	0.0625 (13.0,2.5△)
	+xQuAD	Uni(inf)	0.1766	0.2292	0.1436	0.0545
	+xQuAD	Uni(nav)	0.2003	0.2485	0.1723	0.0626
	+xQuAD	Sel(ora,perf)	**0.2620** (48.4▲,30.8▲)	**0.2985** (30.2▲,20.1▲)	**0.2384** (66.0▲,38.4▲)	**0.0687** (26.1▲,9.7▲)
	+xQuAD	Sel(log,perf)	0.2051 (16.1▲,2.4)	0.2526 (10.2△,1.6)	0.1754 (22.1△,1.8)	0.0636 (16.7▲,1.6)
	+xQuAD	Sel(svm,perf)	0.2042 (15.6△,1.9)	0.2526 (10.2,1.6)	0.1743 (21.4△,1.2)	0.0628 (15.2△,0.3)
	WT09 best (uogTrDYCcsB) [11]		0.1922	0.3081	0.1617	0.0592

Table 3: Diversification performance of IA-Select and xQuAD using informational or navigational models uniformly (UNI) or selectively (SEL) according to the WT09 topics, with the WT10 topics used for training.

classifier performs a multinomial logistic regression with a ridge estimator [40]. In both cases, the single most likely intent is chosen for each aspect, in a typical selective fashion. In order to enable our second regime, model merging, we fit the output of the SVM classifier to a logistic regression model, hence obtaining a full probability distribution over intents for each aspect underlying the query [40]. In order to cope with the high dimensionality of our sub-query feature set, classification is performed after a dimensionality reduction via principal component analysis [40]. All classification tasks are performed using the Weka suite.[8]

5.4 Evaluation and Training Procedure

We report our results based on the official evaluation metrics in the diversity task of the TREC 2010 Web track [12]: ERR-IA [9], α-nDCG [13], NRBP [14], and MAP-IA [1]. The first three metrics implement a cascade user model, which penalises redundancy by assuming an increasing probability that users will stop inspecting the results as they find their desired information. The fourth metric is based on a simpler model, which rewards a high coverage of query aspects, without directly penalising redundancy.

Our evaluation ensures a complete separation between training and test settings. In particular, we use the WT09 and WT10 queries interchangeably as training and test, in a cross-year evaluation fashion (i.e., we train on WT09 and test on WT10, and vice versa). This training procedure renders our results on the WT10 queries directly comparable to those of participant systems in TREC 2010. For the reported results on the WT09 queries, however, we note that TREC 2009 participant systems naturally did not have access to WT10 queries for training. This training procedure is used for the classification approaches described in Section 5.3, as well as for xQuAD's diversification trade-off parameter λ [33]. As for IA-Select, it is a parameter-free diversification approach, and hence requires no training.

6. EXPERIMENTAL EVALUATION

In this section, we evaluate our intent-aware diversification approach, in order to answer the two research questions stated in Section 5. In particular, Section 6.1 investigates the effectiveness of our *model selection* regime, while Section 6.2 analyses the effectiveness of the *model merging* regime. Both regimes were described in Section 4.1.1.

6.1 Intent-Aware Model Selection

Our primary goal in this experiment is to assess the effectiveness of our model selection regime for search result diversification. As described in Section 4.1.1, this regime selects the most likely between an informational and a navigational intent-aware retrieval model for each aspect. As a baseline, we consider a simple regime that uniformly deploys one of the informational or navigational models for all aspects, regardless of the intents of these aspects. To validate our findings, as described in Section 5.2, we test both our model selection regime as well as the baseline uniform regime applied to two diversification approaches: IA-Select and xQuAD. Additionally, these diversification approaches are deployed using two different aspect representations: the official TREC Web track sub-topics and query suggestions from a search engine, as discussed in Section 5.1. Moreover, we test variants of our model selection regime. Each variant is denoted Sel($\mathcal{C}$,$\mathcal{L}$), where $\mathcal{C}$ and $\mathcal{L}$ denote a classifier and a set of classification training labels, respectively. In particular, $\mathcal{C}$ can be one of three classifiers: an oracle (ora), which simulates a perfect classification accuracy, and the logistic regression (log) and support vector machine (svm)

[8] http://www.cs.waikato.ac.nz/ml/weka/

		Regime	ERR-IA@20	α-nDCG@20	NRBP	MAP-IA
	DPH		0.1952	0.2620	0.1509	0.0469
WT10 sub-topics	+IA-Select	Uni(inf)	0.2485	0.3261	0.2011	0.0717
	+IA-Select	Uni(nav)	0.2866	0.3465	0.2490	0.0729
	+IA-Select	Sel(ora,judg)	0.2829 (13.8,-1.3)	0.3496 (7.2,0.9)	0.2428 (20.7,-2.5)	**0.0763** (6.4,4.7)
	+IA-Select	Sel(log,judg)	0.2797 (12.6,-2.4)	0.3442 (5.6,-0.7)	0.2396 (19.1,-3.8)	0.0730 (1.8,0.1)
	+IA-Select	Sel(svm,judg)	**0.2897** (16.6△,1.1)	**0.3535** (8.4,2.0)	**0.2485** (23.6△,-0.2)	0.0750 (4.6,2.9)
	+IA-Select	Sel(ora,perf)	**0.3791** (52.6▲,32.3▲)	**0.4228** (29.7▲,22.0▲)	**0.3491** (73.6▲,40.2▲)	**0.0859** (19.8▲,17.8▲)
	+IA-Select	Sel(log,perf)	0.3117 (25.4▲,8.8▲)	0.3710 (13.8▲,7.1▲)	0.2734 (36.0▲,9.8△)	0.0773 (7.8,6.0△)
	+IA-Select	Sel(svm,perf)	0.3044 (22.5▲,6.2)	0.3638 (11.6△,5.0)	0.2667 (32.6▲,7.1)	0.0765 (6.7,4.9)
	+xQuAD	Uni(inf)	0.2472	0.3241	0.2007	0.0715
	+xQuAD	Uni(nav)	0.2905	0.3479	0.2535	0.0754
	+xQuAD	Sel(ora,judg)	0.2699 (9.2,-7.1)	0.3408 (5.2,-2.0)	0.2245 (11.9,-11.4)	0.0782 (9.4,3.7)
	+xQuAD	Sel(log,judg)	0.2708 (9.5,-6.8)	0.3346 (3.2,-3.8)	0.2333 (16.2,-8.0)	0.0743 (3.9,-1.5)
	+xQuAD	Sel(svm,judg)	**0.2913** (17.8△,0.3)	**0.3512** (8.4,0.9)	**0.2546** (26.9△,0.4)	**0.0793** (10.9,5.2)
	+xQuAD	Sel(ora,perf)	**0.3616** (46.3▲,24.5▲)	**0.4119** (27.1▲,18.4▲)	**0.3294** (64.1▲,29.9▲)	**0.0864** (20.8▲,14.6▲)
	+xQuAD	Sel(log,perf)	0.3090 (25.0▲,6.4)	0.3664 (13.1△,5.3)	0.2726 (35.8▲,7.5)	0.0791 (10.6,4.9)
	+xQuAD	Sel(svm,perf)	0.3098 (25.3▲,6.6)	0.3680 (13.5△,5.8)	0.2707 (34.9▲,6.8)	0.0798 (11.6△,5.8)
Query suggestions	+IA-Select	Uni(inf)	0.2468	0.3135	0.2053	0.0652
	+IA-Select	Uni(nav)	0.2826	0.3419	0.2454	0.0677
	+IA-Select	Sel(ora,perf)	**0.3677** (49.0▲,30.1▲)	**0.4174** (33.1▲,22.1▲)	**0.3343** (62.8▲,36.2▲)	**0.0774** (18.7▲,14.3▲)
	+IA-Select	Sel(log,perf)	0.2942 (19.2▲,4.1)	0.3523 (12.4△,3.0)	0.2575 (25.4△,4.9)	0.0715 (9.7▲,5.6△)
	+IA-Select	Sel(svm,perf)	0.2945 (19.3▲,4.2)	0.3530 (12.6▲,3.2)	0.2586 (26.0▲,5.4)	0.0722 (10.7△,6.6△)
	+xQuAD	Uni(inf)	0.2456	0.3110	0.2025	0.0600
	+xQuAD	Uni(nav)	0.2579	0.3211	0.2174	0.0597
	+xQuAD	Sel(ora,perf)	**0.3554** (44.7▲,37.8▲)	**0.4078** (31.1▲,27.0▲)	**0.3210** (58.5▲,47.7▲)	**0.0768** (28.0▲,28.6▲)
	+xQuAD	Sel(log,perf)	0.2743 (11.7△,6.4)	0.3375 (8.5△,5.1)	0.2349 (16.0△,8.0)	0.0716 (19.3▲,19.9▲)
	+xQuAD	Sel(svm,perf)	0.2805 (14.2▲,8.8△)	0.3435 (10.5△,7.0)	0.2418 (19.4▲,11.2)	0.0730 (21.7▲,22.3▲)
	WT10 best (uogTrB67xS) [12]		0.2981	0.4178	0.2616	0.0737

Table 4: Diversification performance of IA-Select and xQuAD using informational or navigational models uniformly (UNI) or selectively (SEL) according to the WT10 topics, with the WT09 topics used for training.

classifiers described in Section 5.3. As for the classification labels $\mathcal{L}$, as described in Section 4.1.2, we consider both human judgements (judg) as well as the selection with best diversification performance (perf) on the training data.

Tables 3 and 4 compare the aforementioned variants of our model selection regime to the baseline uniform regime on the WT09 and WT10 queries, respectively, in terms of the four metrics described in Section 5.4: ERR-IA@20, α-nDCG@20, NRBP, and MAP-IA. In parentheses, we show the percent improvement of each variant of our model selection regime compared to the uniform regime that uses the informational (Uni(inf)) or the navigational (Uni(nav)) model, respectively. Significant improvements are measured by the Wilcoxon signed-rank test. In particular, the symbols ▲ (▼) and △ (▽) denote a significant increase (decrease) at the $p < 0.01$ and $p < 0.05$ levels, respectively. Lastly, the bottom row in Tables 3 and 4 shows the performance of the top performing category-B system in the WT09 and WT10 tasks [11, 12], respectively, hence providing a further reference value for evaluating our intent-aware approach.

From Tables 3 and 4, we first note that the uniform application of an informational or a navigational model provides a strong baseline performance. The uniform application of the navigational model, in particular, performs at least comparably to the best performing TREC system in both the WT09 and WT10 tasks. To see whether our model selection regime can improve upon this strong baseline, we first look at the performance of this regime using an oracle classifier. The results show that a massive improvement can be attained by selecting the most appropriate model for each aspect (as opposed to uniformly using a single model for all aspects) for both IA-Select and xQuAD, when the performance-oriented labels (perf) are used for training. On the WT09 queries (Table 3), compared to the strongest uniform setting (i.e., Uni(nav)) in terms of ERR-IA@20, improvements can be as high as 29.1% for IA-Select and 19.4% for xQuAD using the Web track sub-topics as sub-queries, and are always significant. When using query suggestions, the potential improvements are 38.3% and 30.8%, respectively. On the WT10 queries (Table 4), similar figures are observed: 32.3% and 24.5% gain for IA-Select and xQuAD, respectively, using the Web track sub-topics; 30.1% and 37.8% using query suggestions. Once again, all improvements are statistically significant. Human judgements, in contrast, provide a supotimal labelling criterion, as denoted by the lower performance attained when using the judg labels. Indeed, even an oracle classifier, which always choses the correct intent according to these judgements (i.e., the Sel(ora,judg) regime), cannot improve over applying the navigational model uniformly. As discussed in Section 4.1.2, this further confirms our intuition that the appropriateness of an intent-aware retrieval model for a given aspect cannot be effectively judged purely on the basis of the apparent characteristics of this aspect.

Besides showing a strong potential for improving diversification performance, as demonstrated using an oracle classifier (i.e., the Sel(ora,perf) regime), our intent-aware approach is also effective in a practical deployment based on standard classifiers. Indeed, Tables 3 and 4 show that our model selection regime using both logistic regression (log) and support vector machine (svm) classifiers with perf labels always improves compared to a uniform regime, often significantly. For instance, on the WT09 queries (Table 3) and considering the Web track sub-topics as aspect representations, compared to the stronger Uni(nav) baseline, improvements in terms of ERR-IA@20 are as high as 11.1% for IA-Select (Sel(log,perf)) and 7.9% for xQuAD (Sel(log,perf)). On the WT10 queries (Table 4), improvements are as high as 8.8% for IA-Select (Sel(log,perf))

		Regime	ERR-IA@20	α-nDCG@20	NRBP	MAP-IA
WT09	DPH		0.1607	0.2097	0.1318	0.0442
	+IA-Select	Sel(svm,judg)	0.2126	0.2594	0.1834	0.0664
	+IA-Select	Mrg(svm,judg)	**0.2224** (4.6)	**0.2661** (2.6)	**0.1930** (5.2)	**0.0670** (0.9)
	+IA-Select	Sel(svm,perf)	**0.2289**	**0.2713**	**0.2036**	0.0677
	+IA-Select	Mrg(svm,perf)	0.2206 (-3.6)	0.2644 (-2.5)	0.1922 (-5.6)	**0.0681** (0.6)
	+xQuAD	Sel(svm,judg)	0.2153	0.2594	0.1868	0.0671
	+xQuAD	Mrg(svm,judg)	**0.2212** (2.7)	**0.2638** (1.7)	**0.1936** (3.6)	**0.0676** (0.7)
	+xQuAD	Sel(svm,perf)	0.2370	0.2754	**0.2125**	**0.0694**
	+xQuAD	Mrg(svm,perf)	**0.2371** (0.0)	**0.2759** (0.2)	0.2117 (-0.4)	0.0692 (-0.3)
WT10	DPH		0.1952	0.2620	0.1509	0.0469
	+IA-Select	Sel(svm,judg)	0.2897	0.3535	0.2485	0.0750
	+IA-Select	Mrg(svm,judg)	**0.2946** (1.7△)	**0.3579** (1.2△)	**0.2537** (2.1△)	**0.0755** (0.7)
	+IA-Select	Sel(svm,perf)	0.3044	0.3638	0.2667	0.0765
	+IA-Select	Mrg(svm,perf)	**0.3069** (0.8)	**0.3663** (0.7)	**0.2690** (0.9)	**0.0769** (0.5)
	+xQuAD	Sel(svm,judg)	0.2913	0.3512	0.2546	0.0793
	+xQuAD	Mrg(svm,judg)	**0.2960** (1.6)	**0.3544** (0.9)	**0.2594** (1.9)	**0.0798** (0.6)
	+xQuAD	Sel(svm,perf)	**0.3098**	**0.3680**	**0.2707**	0.0798
	+xQuAD	Mrg(svm,perf)	0.2997 (-3.3)	0.3592 (-2.4)	0.2622 (-3.1)	**0.0819** (2.6)

Table 5: Diversification performance of IA-Select and xQuAD using informational or navigational models selectively (SEL) or through merging (MRG). WT09 and WT10 results are shown on the top and bottom halves, respectively. As in Tables 3 and 4, WT09 results are trained on WT10 topics, and vice versa.

and 6.6% for xQuAD (Sel(svm,perf)). Similar improvements across the other reported metrics are consistently observed. When query suggestions are used as aspect representations, although improvements are less pronounced, they are consistent and can still be significant.

Overall, the results in this section answer our first research question, by showing that diversification performance can be significantly improved by leveraging the most appropriate intent-aware retrieval model for each query aspect. Our model selection regime using performance-oriented classification labels is particularly effective, significantly improving upon a uniform regime comparable to the top performing systems of the TREC 2009 and 2010 Web tracks [11, 12]. Furthermore, the consistency of our observations for two state-of-the-art diversification approaches and according to multiple evaluation metrics attests the robustness of the model selection regime. In the next section, we contrast this regime against the alternative model merging regime.

6.2 Intent-Aware Model Merging

After demonstrating the effectiveness of selecting a single model for each query aspect, in this experiment, we investigate whether deploying a model merging regime could bring further improvements. For this investigation, we focus our attention to the TREC Web track sub-topics as an aspect representation. As discussed in Section 5.1, this representation allows for assessing the effectiveness of our merging regime across the two proposed training labelling alternatives, judg and perf. The results based on query suggestions using perf labels lead to identical conclusions and are hence omitted for brevity. In particular, Table 5 shows the diversification performance of IA-Select and xQuAD under the model merging regime (Mrg), in contrast to their performance under the model selection regime (Sel), which serves as our baseline in this investigation. Similarly to Tables 3 and 4, percent differences between these two regimes are shown in parentheses, alongside one of the aforementioned symbols to denote the significance (or lack thereof) of such differences. As discussed in Section 5.3, both regimes are based on predictions made by an SVM classifier. In particular, the model merging regime is enabled by fitting the SVM predictions to a logistic regression model.

From Table 5, we observe that the model merging regime can improve upon the model selection regime in most cases, particularly on the WT10 queries (bottom half of Table 5). However, the merging regime can also underperform compared to the selection regime, when perf labels are used for IA-Select and xQuAD, on WT09 and WT10, respectively. Nevertheless, significant differences are only observed when IA-Select is deployed under the Mrg(svm,judg) regime on the WT10 queries. These results answer our second research question, by showing that merging multiple intent-aware models can be at least as effective as selecting the single most likely model. Moreover, we believe that the merging regime can offer additional benefits for an intent-aware diversification. For one, it can help attenuate the harm of selecting the wrong model for a particular sub-query. Additionally, it provides a natural upper-bound for the selection regime. Indeed, model selection is a special instance of model merging, with a mutually exclusive probability distribution.

7. CONCLUSIONS

In this paper, we have introduced a novel intent-aware approach for search result diversification. Given the possible intents underlying the aspects of a query, our approach learns the appropriateness of retrieval models targeted to each of these intents. These models are then leveraged selectively or combined in a merging fashion in order to refine the estimation of the relevance of the retrieved documents with respect to each query aspect. In particular, our approach builds upon a general explicit diversification model, which makes it seamlessly deployable by existing approaches in the literature. Indeed, thorough experiments in the context of the TREC 2009 and 2010 Web tracks demonstrate that our approach is general and significantly improves the effectiveness of two state-of-the-art diversification approaches.

Our data-driven approach for learning both intent-aware retrieval models and their appropriateness for a given aspect opens up promising directions. In particular, the full potential of our approach could be further exploited by building upon a larger pool of intent-aware retrieval models (e.g., with, say, a transactional model [4]), as well as features capturing dependencies between the retrieved documents.

8. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.

[2] G. Amati, C. Carpineto, G. Romano, and F. U. Bordoni. Query difficulty, robustness and selective application of query expansion. In *ECIR*, pages 127–137, 2004.

[3] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Link-based characterization and detection of Web spam. In *AIRWeb*, 2006.

[4] A. Broder. A taxonomy of Web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336, 1998.

[6] D. Carmel and E. Yom-Tov. Estimating the query difficulty for information retrieval. In *SIGIR*, page 911, 2010.

[7] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *CIKM*, pages 1287–1296, 2009.

[8] B. Carterette, V. Pavluz, H. Fangx, and E. Kanoulas. Million Query track 2009 overview. In *TREC*, 2009.

[9] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, pages 621–630, 2009.

[10] H. Chen and D. R. Karger. Less is more: Probabilistic models for retrieving fewer relevant documents. In *SIGIR*, pages 429–436, 2006.

[11] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *TREC*, 2009.

[12] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 Web track. In *TREC*, 2010.

[13] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pages 659–666, 2008.

[14] C. L. A. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *ICTIR*, pages 188–199, 2009.

[15] P. Clough, M. Sanderson, M. Abouammoh, S. Navarro, and M. Paramita. Multiple approaches to analysing query diversity. In *SIGIR*, pages 734–735, 2009.

[16] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large Web datasets. *Inf. Retr.*, 2011.

[17] N. Craswell and D. Hawking. Overview of the TREC 2004 Web track. In *TREC*, 2004.

[18] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *SIGIR*, pages 416–423, 2005.

[19] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR*, pages 299–306, 2002.

[20] X. Geng, T.-Y. Liu, T. Qin, A. Arnold, H. Li, and H.-Y. Shum. Query dependent ranking using k-nearest neighbor. In *SIGIR*, pages 115–122, 2008.

[21] B. He and I. Ounis. Query performance prediction. *Inf. Syst.*, 31(7):585–594, 2006.

[22] I.-H. Kang and G. Kim. Query type classification for Web document retrieval. In *SIGIR*, pages 64–71, 2003.

[23] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[24] D. A. Metzler. Automatic feature selection in the Markov random field model for information retrieval. In *CIKM*, pages 253–262, 2007.

[25] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *SIGIR, OSIR Workshop*, 2006.

[26] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66, Stanford, 1999.

[27] J. Peng, C. Macdonald, and I. Ounis. Learning to select a ranking function. In *ECIR*, pages 114–126, 2010.

[28] V. Plachouras, I. Ounis, and G. Amati. The static absorbing model for the Web. *J. Web Eng.*, 4(2):165–186, 2005.

[29] T. Qin, T.-Y. Liu, J. Xu, and H. Li. LETOR: A benchmark collection for research on learning to rank for information retrieval. *Inf. Retr.*, 13(4):346–374, 2010.

[30] S. E. Robertson. The probability ranking principle in IR. *J. Doc.*, 33(4):294–304, 1977.

[31] D. E. Rose and D. Levinson. Understanding user goals in Web search. In *WWW*, pages 13–19, 2004.

[32] M. Sanderson. Ambiguous queries: Test collections need more sense. In *SIGIR*, pages 499–506, 2008.

[33] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for Web search result diversification. In *WWW*, pages 881–890, 2010.

[34] R. L. T. Santos, C. Macdonald, and I. Ounis. Selectively diversifying Web search results. In *CIKM*, pages 1179–1188, 2010.

[35] R. L. T. Santos and I. Ounis. Diversifying for multiple information needs. In *ECIR, DDR Workshop*, pages 37–41, 2011.

[36] R. L. T. Santos, J. Peng, C. Macdonald, and I. Ounis. Explicit search result diversification through sub-queries. In *ECIR*, pages 87–99, 2010.

[37] R. Song, Z. Luo, J.-Y. Nie, Y. Yu, and H.-W. Hon. Identification of ambiguous queries in Web search. *Inf. Process. Manage.*, 45(2):216–229, 2009.

[38] K. Spärck-Jones, S. E. Robertson, and M. Sanderson. Ambiguous requests: Implications for retrieval tests, systems and theories. *SIGIR Forum*, 41(2):8–17, 2007.

[39] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR*, pages 115–122, 2009.

[40] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools*. 2005.

[41] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *SIGIR*, pages 10–17, 2003.

[42] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *SIGIR*, pages 543–550, 2007.