Aggregated Search Result Diversification

Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis

School of Computing Science
University of Glasgow
G12 8QQ, Glasgow, UK
{rodrygo,craigm,ounis}@dcs.gla.ac.uk

Abstract. Search result diversification has been effectively employed to tackle query ambiguity, particularly in the context of web search. However, ambiguity can manifest differently in different search verticals, with ambiguous queries spanning, e.g., multiple place names, content genres, or time periods. In this paper, we empirically investigate the need for diversity across four different verticals of a commercial search engine, including web, image, news, and product search. As a result, we introduce the problem of aggregated search result diversification as the task of satisfying multiple information needs across multiple search verticals. Moreover, we propose a probabilistic approach to tackle this problem, as a natural extension of state-of-the-art diversification approaches. Finally, we generalise standard diversity metrics, such as ERR-IA and α-nDCG, into a framework for evaluating diversity across multiple search verticals.

1 Introduction

Queries submitted to a web search engine are typically short and often ambiguous [28]. For instance, a user issuing the query '*amazon*' may be looking for the e-commerce company or the rainforest. Likewise, a user issuing a less ambiguous query such as '*amazon.com*' may be still interested in different aspects of this query, e.g., books, electronics, or digital music. To maximise the chance that different users will find at least one relevant search result to their particular information need, an effective approach is to diversify these results [13].

Existing diversification approaches have been deployed mostly in the context of web (e.g., [1, 10, 12, 26, 27]) and newswire (e.g., [6, 9, 32, 33]) search, but there have also been approaches dedicated to diversifying image (e.g., [22, 24]) and product (e.g., [19, 31]) search results. Nevertheless, the nature of ambiguity can drastically vary across different search verticals. For instance, while query ambiguity arguably takes a more topical nature in a context such as web or image search, a news search query (e.g., '*olympics*') may give rise to temporal ambiguity (e.g., 2012? 2016?). In the same vein, a map search query (e.g., '*columbia*') may introduce geographical ambiguity (e.g., Maryland? Missouri? South Carolina?), while a blog search query (e.g., '*politics*') may entail social ambiguity (e.g., left-wing? neutral? right-wing?). Moreover, with the prevalence of aggregated search interfaces in modern web search [18, 23], a search engine may be faced with the task of tackling query ambiguity spanning multiple search verticals.

In this paper, we introduce *aggregated search result diversification* as the problem of satisfying multiple possible information needs across multiple search verticals. To quantify the need for diversity in different search verticals, we investigate the nature of query ambiguity across four verticals of a commercial search engine, namely, web, image, news, and product search. Our investigation, based on queries from the TREC 2009 and 2010 Web tracks [10, 12], shows that the ambiguity of a query varies considerably across different verticals, as do the likelihood of the different aspects underlying this query. Based upon this investigation, we propose a probabilistic approach for aggregated search result diversification. In particular, we extend state-of-the-art diversification approaches into a holistic approach to diversify the search results across multiple search verticals. Finally, we generalise standard diversity metrics into a framework for evaluating aggregated search result diversification. As a result, we extend the notion of whole-page relevance [3] to quantify the diversity of an entire result page.

Our major contributions are four-fold: (1) we introduce the problem of aggregated search result diversification; (2) we motivate this new problem through an empirical investigation using publicly available data from a commercial search engine; (3) we propose a probabilistic approach for tackling this problem; and (4) we introduce a general framework for evaluating approaches to this problem.

The remainder of this paper is organised as follows. Section 2 overviews related work in search result diversification and aggregated search. Section 3 investigates the nature of query ambiguity across four verticals of a commercial search engine, as a motivation for this work. Section 4 formalises the aggregated search result diversification problem. Section 5 describes our probabilistic approach for tackling the introduced problem. Section 6 proposes a framework for evaluating whole-page diversity. Lastly, Section 7 presents our conclusions.

2 Related Work

In this section, we describe diversification approaches that have been deployed in verticals such as web, newswire, image, and product search. We then review related work on aggregating results from multiple search verticals.

2.1 Search Result Diversification

The goal of search result diversification is to produce a ranking with maximum coverage and minimum redundancy with respect to the aspects underlying a query [13]. In recent years, several diversification approaches have been proposed, covering a range of search verticals such as web (e.g., [1, 10, 12, 26, 27]), newswire (e.g., [9, 32, 33]), image (e.g., [22, 24]), and product (e.g., [19, 31]) search.

In the context of web search, Agrawal et al. [1] proposed to diversify the search results with respect to a taxonomy of categories. Their approach focused on promoting search results with a high coverage of categories also covered by the query, but poorly covered by the other results. Rafiei et al. [26] proposed to favour search results that correlate lowly (in terms of content or received clicks) with the other results, so as to promote novelty in the ranking. Santos et al. [27] proposed a probabilistic framework to rank the search results with respect to their coverage and novelty in light of multiple query aspects—represented by different query reformulations—as well as the relative importance of these aspects.

Newswire search result diversification was first tackled by Carbonell and Goldstein [6]. They proposed to rank the search results based on these results' estimated relevance to the query and their dissimilarity to the other results. Zhai et al. [33] extended this idea by comparing the search results in terms of the divergence of their language models, while Wang and Zhu [32] exploited relevance score correlations. Similarly, Chen and Karger [9] proposed to diversify the search results conditioned on the assumed irrelevance of the other results.

In the context of image search, van Leuken et al. [22] proposed to cluster the retrieved images using visual features, so that representative images from different clusters could form a diverse ranking. A similar approach was proposed by Deselaer et al. [15], however mixing both textual and visual features. In a different vein, Paramita et al. [24] proposed to diversify image search results spatially, by leveraging location information associated to every image.

Finally, in the context of product search, Vee et al. [31] proposed to diversify the search results for structured queries. From an initial ranking of products satisfying the query predicates, they devised tree-traversal algorithms to efficiently compare product pairs with respect to their attribute values. Similarly, Gollapudi and Sharma [19] deployed facility dispersion algorithms in order to promote diversity. Their approach compares the products retrieved for a query in terms of their categorical distance according to a given taxonomy.

2.2 Aggregated Search

Commercial web search engines often complement web search results with results from other search verticals, such as images, videos, and news articles [23]. As a modern instantiation of distributed information retrieval (DIR) [5], aggregated search involves the representation, selection, and aggregation of search results from multiple verticals. However, differently from traditional DIR problems, aggregated search deals with highly heterogeneous resources (i.e., search verticals) in a cooperative environment (i.e., verticals are usually run by the same company) and with abundance of usage data (e.g., vertical-specific query logs) [18]. Research in aggregated search has mostly focused on vertical selection, with fewer studies investigating the composition and evaluation of aggregated interfaces.

Vertical selection closely resembles resource selection in DIR [5]. While resource selection approaches typically focus on the contents of different resources (e.g., their size or their estimated number of relevant documents), modern vertical selection approaches leverage a wealth of available evidence from usage data. For instance, Diaz [16] proposed to predict the newsworthiness of web search queries by leveraging implicit user feedback (e.g., clicks and skips). Beitzel et al. [4] proposed a semi-supervised classification approach for automatically labelling queries with respect to 18 different verticals. A supervised approach was proposed by Arguello et al. [2] by exploiting evidence from vertical-specific query logs and Wikipedia-induced vertical samples. Later, Diaz and Arguello [17] proposed to improve classification-based vertical selection by leveraging click data.

The composition of aggregated search interfaces was investigated by Ponnuswami et al. [25]. They improved click-through rates by learning to display results from already selected verticals in light of the displayed web search results. In terms of evaluation, Sushmita et al. [30] conducted a user study on factors affecting click-through rates on aggregated search. They observed a significant bias towards rank positions, but not towards any particular vertical. Finally, Bailey et al. [3] proposed a method for evaluating the relevance of a results page as a whole, as opposed to evaluating the relevance of individual search results.

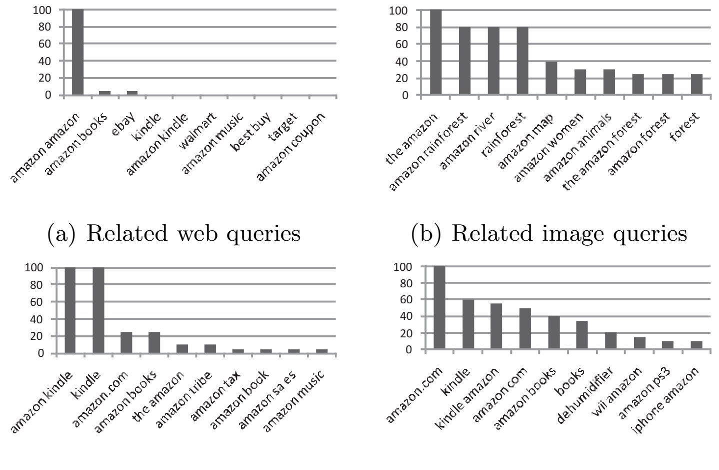
In the next section, we will bridge the gap between the search result diversification and the aggregated search problems, by investigating the nature of query ambiguity across multiple search verticals. This investigation will provide empirical motivation for the aggregated search result diversification problem, and the basis for modelling and evaluating approaches to this problem.

3 The Nature of Ambiguity

Aggregated search can be regarded as performing a surface-level diversification, in the sense that it tackles content-type ambiguity [14]. For instance, it is unclear whether a user issuing the query '*amazon*' to a web search engine would be satisfied with the e-commerce company's homepage (a standard web search result), its current stock performance (a financial search result), or the latest news about the company's recently announced music storage service (a news search result). Nevertheless, at the deeper level of individual verticals, the nature of the ambiguity of a single query can vary even further. To illustrate this observation, we use Google Insights for Search,[1] a service providing statistics of searchers' interest according to four Google verticals: web, image, news, and product search. For this particular study, we focus on related searches provided by each vertical for an input query, which can be regarded as representing the most likely information needs underlying this query in the context of a particular vertical. As an example, Fig. 1 shows the top 10 queries related to the query '*amazon*' in each of the four considered verticals, along with the normalised search volume associated to each of these queries. To ensure a uniform setting across the four considered verticals, we constrain our analysis to the US market. Additionally, we consider the total search volume generated within the period of January 2008 to March 2011, so as to attenuate seasonal or bursty interest fluctuations.

From Fig. 1(a), we observe that, in the web search vertical, the query '*amazon*' is likely to refer to some aspect of the e-commerce company, or similar companies related to it. However, in the image search vertical (Fig. 1(b)), the same query more likely refers to the Amazon rainforest. Likewise, this query may refer to the launch of new products and services or the discovery of a new tribe in the rainforest in the news vertical (Fig. 1(c)), or to the most popular products offered by the e-commerce company in the product vertical (Fig. 1(d)).

[1] `http://www.google.com/insights/search`



(a) Related web queries

(b) Related image queries

(c) Related news queries

(d) Related product queries

Fig. 1: Search volume for the queries most related to '*amazon*' in four verticals.

To empirically quantify the nature of ambiguity across different verticals, we analyse the statistics provided by Google Insights for Search for all 100 queries from the TREC 2009 and 2010 Web tracks [10, 12]. In particular, this set comprises queries of medium popularity sampled from the logs of a commercial search engine, hence providing a representative set of somewhat ambiguous yet not too popular web search queries. Limited by Google Insights for Search, we obtain up to the 50 most frequent queries related to each of the TREC Web track queries, along with their associated search volume, according to the aforementioned market and period constraints. Of the initial 100 queries, three do not have enough search volume in any of the four considered verticals, and are hence discarded from our analysis. Of the remaining 97 queries, 36 occur in only one vertical, 18 in two, 13 in three, and 30 queries occur in all four considered verticals. In cumulative terms, 61 ($\approx 63\%$) of the 97 considered ambiguous queries yield a non-negligible search volume in more than one vertical, which confirms the need to tackle query ambiguity spanning multiple search verticals.

To provide a consistent cross-vertical comparison, we further analyse the ambiguity of the 30 queries that occur in all four considered verticals. In particular, for each query q, let X be a categorical random variable with sample space $\mathcal{A}(q) = \{a_1, \ldots, a_k\}$, i.e., the set of all aspects underlying q, with each aspect represented by a query related to q, as obtained from all verticals. Likewise, let Y be a discrete random variable with sample space $\mathcal{V}(q) = \{v_1, \ldots, v_m\}$, i.e., the set of all verticals available for q. Lastly, let Z_v be a real-valued random variable with values $Z_v(a) = f_{X|Y}(X = a|Y = v)$, i.e., the frequency with which the aspect a is observed given that the vertical v was selected, as given by total search volume reported for the aspect a by the vertical v. We propose three metrics to contrast query ambiguity across the four considered verticals:

Ambiguity. The ambiguity of a query q quantifies the range of possible information needs underlying this query in light of a particular vertical v. We define it as the number of unique aspects related to q according to v:

$$\mathit{ambiguity}(q, v) = |\{a \in \mathcal{A}(q) : Z_v(a) > 0\}|. \tag{1}$$

Dominance. The definition of dominance complements our basic notion of ambiguity by showing how the interest for different information needs underlying a query is distributed. In particular, dominance quantifies the bias towards one or a few highly likely aspects of a query q in light of a particular vertical v. It is defined as the sample skewness g_1 [20] of the frequency distribution of the aspects related to q according to v:

$$\mathit{dominance}(q, v) = g_1(Z_v) = \frac{\sum_{a \in \mathcal{A}(q)} (Z_v(a) - \bar{Z}_v)^3}{(k-1)^3}, \tag{2}$$

where $\bar{Z}_v$ denotes the mean frequency over all $a \in \mathcal{A}(q)$ given v. A positive dominance indicates a bias towards frequent aspects, while a dominance approaching zero reveals a normal distribution around the mean frequency.

Agreement. The agreement of a pair of verticals with respect to a query quantifies the similarity of the distribution of information needs underlying this query across the two verticals. In other words, it measures the extent to which these verticals generate the same interest for the possible information needs underlying the query. We define the agreement between verticals v_i and v_j for a query q as the Czekanowski index[2] C [20] between the frequency distribution of aspects related to q according to v_i and v_j:

$$\mathit{agreement}(q, v_i, v_j) = C(Z_{v_i}, Z_{v_j}) = \frac{2 \sum_{a \in \mathcal{A}(q)} \min(Z_{v_i}(a), Z_{v_j}(a))}{\sum_{a \in \mathcal{A}(q)} (Z_{v_i}(a) + Z_{v_j}(a))}, \tag{3}$$

where the denominator performs a normalisation to enable the direct comparison of the agreement of different pairs of verticals. A value of one denotes total agreement, while a value of zero denotes total disagreement.

Table 1 shows the mean ambiguity, dominance, and agreement for the 30 TREC 2009 and 2010 Web track queries occurring in all four verticals, along with 95% confidence intervals for the means. From the table, we first note that, compared to news and product search, web and image search queries yield a significantly higher ambiguity. While the reason for this observation may be trivial (i.e., web and image search arguably receive a higher traffic), an important consequence is that diversification approaches should be aware of the varying number of possible aspects underlying the same query submitted to different verticals. In terms of dominance, all verticals show positive values, which indicate a moderate bias towards a few highly likely information needs. As an illustration (not shown in Table 1) of how a few aspects dominate the interest of the user population,

[2] For binary distributions, the Czekanowski index is equivalent to the Dice index.

Table 1: Mean ambiguity, dominance, and agreement across 30 TREC 2009 and 2010 Web track queries occurring in four Google verticals. A 95% confidence interval for the means according to the Student's t-distribution is also shown.

		web	image	news	product
ambiguity		45.567 ±3.083	43.167 ±4.198	21.233 ±6.830	30.433 ±7.114
dominance		6.295 ±0.925	5.350 ±0.720	7.741 ±1.406	7.406 ±1.276
agreement		web	image	news	product
	web	1.000 ±0.000	0.372 ±0.054	0.282 ±0.060	0.285 ±0.066
	image	–	1.000 ±0.000	0.223 ±0.065	0.204 ±0.055
	news	–	–	1.000 ±0.000	0.120 ±0.041
	product	–	–	–	1.000 ±0.000

to account for 70% of the search interest around all aspects of an ambiguous query, the web, image, news, and product verticals require, on average, only the top 35±3, 39±3, 48±10, and 46±8% most frequent aspects of this query, respectively. In absolute terms, the news search vertical shows the highest dominance, although not significantly higher than that of the other verticals. Finally, in terms of cross-vertical agreement, the highest non-trivial value observed is 0.372, when the web and image search verticals are compared. This observation quantitatively corroborates the illustrative example in Fig. 1, by showing that different verticals produce very dissimilar distributions of query aspects.

Overall, the results in this section highlight the specificities of query ambiguity in different search verticals, and the practical issues that must be considered when tackling it. Motivated by this investigation, in the next section, we formalise the problem of aggregated search result diversification.

4 Problem Formulation

Let $\mathcal{V}(q)$ denote the set of verticals v selected for a query q. Moreover, let $\mathcal{R}(q)$ denote the union of all search results r retrieved from these verticals. Finally, let $\mathcal{Q}(\cdot)$ denote the set of relevant aspects for a given input (a query or a result). For a rank cutoff $\tau > 0$, the goal of the aggregated search result diversification problem is to find a subset $\mathcal{S}(q) \subseteq \mathcal{R}(q)$, such that:

$$\mathcal{S}(q) = \mathop{\arg\max}_{\mathcal{S}_i(q)\subseteq\mathcal{R}(q)} \left| \cup_{\substack{v\in\mathcal{V}(q)\\ r\in\mathcal{S}_i(q)}} \mathcal{Q}(q|v)\cap\mathcal{Q}(r)\right|, \text{ s.t. } |\mathcal{S}_i(q)| \le \tau. \tag{4}$$

From a search result diversification perspective, this formulation extends the diversification problem to account for query ambiguity across multiple search verticals. The key difference here is that the relevant aspects for a given query now depend on each individual vertical (i.e., $\mathcal{Q}(q|v)$), as motivated by our investigation in Section 3. From an aggregated search perspective, this formulation impacts the representation and selection of search verticals, which may benefit from accounting for the estimated diversity of the results provided by each vertical. Moreover, as we will show in Sections 5 and 6, this formulation impacts the

```
**Diversify**$^{agg}(q, \mathcal{R}(q), \mathcal{V}(q), \tau)$
1 $\mathcal{S}(q) \leftarrow \emptyset$
2 **while** $|\mathcal{S}(q)| < \tau$ **do**
3     $r^* \leftarrow \arg\max_{r \in \mathcal{R}(q) \setminus \mathcal{S}(q)} \; f(r|q, \mathcal{S}(q), \mathcal{V}(q))$
4     $\mathcal{R}(q) \leftarrow \mathcal{R}(q) \setminus \{r^*\}$
5     $\mathcal{S}(q) \leftarrow \mathcal{S}(q) \cup \{r^*\}$
6 **end while**
7 **return** $\mathcal{S}(q)$
```

Alg. 1: Greedy aggregated diversification.

criteria adopted for aggregating results from multiple verticals and for evaluating this aggregation, as these results should ideally satisfy different information needs—as opposed to a single, precisely defined need—from different verticals.

5 Modelling Aggregated Diversification

By directly extending the basic diversification problem, the aggregated diversification problem also inherits its complexity. Indeed, both problems are instances of the maximum coverage problem, which is NP-hard [1]. Fortunately, there is a well-known greedy algorithm for this problem, which achieves a $(1 - 1/e)$-approximation. This is also the best possible worst-case approximation ratio achievable in polynomial time, unless NP $\subseteq$ DTIME$(n^{O(\log \log n)})$ [21].

In this section, we instantiate this greedy algorithm to tackle the aggregated diversification problem. In particular, Alg. 1 takes as input a query q, an initial ranking $\mathcal{R}(q)$ with $n = |\mathcal{R}(q)|$, a set of search verticals $\mathcal{V}(q)$, and an integer τ, with $0 < \tau \leq n$. It then iteratively constructs a re-ranking $\mathcal{S}(q)$, with $|\mathcal{S}(q)| \leq \tau$, by selecting, at each iteration, a search result $r \in \mathcal{R}(q) \setminus \mathcal{S}(q)$ that maximises the objective function f (line 3 in Alg. 1). This function evaluates a search result r given the query q, the results in $\mathcal{S}(q)$, selected in the previous iterations of the algorithm, and the considered verticals $\mathcal{V}(q)$. In this paper, we propose a probabilistic interpretation for the function f:

$$f(r|q, \mathcal{S}(q), \mathcal{V}(q)) \equiv \mathrm{P}(r|\mathcal{S}(q), q). \tag{5}$$

This formulation defines the diversity of a single result r as the probability of observing r conditioned on the observation of the query q and the already selected results in $\mathcal{S}(q)$.[3] In order to account for the available verticals, we marginalise the above probability over $\mathcal{V}(q)$ as a latent variable:

$$f \equiv \mathrm{P}(r|\mathcal{S}(q), q) = \sum_{v \in \mathcal{V}(q)} \mathrm{P}(v|q)\, \mathrm{P}(r|\mathcal{S}(q), q, v), \tag{6}$$

[3] Conditioning on $\mathcal{S}(q)$ is a typical mechanism for promoting novel results [9]—i.e., results different from those (assumed irrelevant) already in $\mathcal{S}(q)$.

where $\mathrm{P}(v|q)$ is the probability of selecting the vertical v for the query q, and $\mathrm{P}(r|\mathcal{S}(q), q, v)$ denotes the probability of the search result r being relevant given the already selected results in $\mathcal{S}(q)$, the query q, and the vertical v. The latter probability is (in some form) at the core of most of the diversification approaches in the literature. In this work, we follow the state-of-the-art approaches [1, 27], in order to explicitly account for the possible aspects underlying the query q in light of the vertical v. To do so, we further marginalise the probability $\mathrm{P}(r|\mathcal{S}(q), q, v)$ in Equation (6) over the set of aspects $\mathcal{A}(q|v)$ identified for q given v, as follows:

$$f \equiv \mathrm{P}(r|\mathcal{S}(q), q) = \sum_{v \in \mathcal{V}(q)} \mathrm{P}(v|q) \sum_{a \in \mathcal{A}(q|v)} \mathrm{P}(a|q, v)\, \mathrm{P}(r|\mathcal{S}(q), q, v, a), \qquad (7)$$

where $\mathrm{P}(a|q, v)$ denotes the likelihood of the aspect a given the query q and the vertical v, and $\mathrm{P}(r|\mathcal{S}(q), q, v, a)$ is the probability of the search result r being relevant given the already selected results in $\mathcal{S}(q)$, q, v, and a.

The problem is now reduced to estimating the various components in Equation 7. In particular, the set of verticals $\mathcal{V}(q)$ available for a query is normally fixed, while the probability $\mathrm{P}(v|q)$ can be estimated using any standard vertical selection approach, such as those described in Section 2.2. As for the set $\mathcal{A}(q|v)$ of query aspects identified from each vertical, as well as their likelihood $\mathrm{P}(a|q, v)$, one can deploy query log mining techniques to vertical-specific usage logs [27]. Alternatively, a taxonomy of categories appropriate to each individual vertical could be considered [1]. Finally, provided that the relevance estimation mechanism used by each considered vertical is available—which is the case in a typically cooperative aggregated search scenario—the probability $\mathrm{P}(r|\mathcal{S}(q), q, v, a)$ can be directly estimated by the state-of-the-art approaches in the literature [1, 27].

Given the lack of a shared test collection for aggregated search evaluation, we leave the empirical validation of our proposed approach for the future. Such a collection could be constructed as part of a formal evaluation campaign (e.g., within the auspices of TREC) or as an individual or group effort (e.g., via crowdsourcing). Nevertheless, in the next section, we prepare the grounds for such an evaluation by proposing a suitable framework for this purpose.

6 Evaluating Aggregated Diversification

Traditional information retrieval evaluation metrics assume that the query unambiguously defines a user's information need. However, this assumption may not hold true in a real search scenario, when there is uncertainty regarding which aspect of the query the user is interested in [29]. Cascade metrics, such as ERR [8] and α-DCG [13], partially address this problem, by modelling a user who stops inspecting the result list as soon as a relevant result is found, hence rewarding novelty. To ensure that a high coverage of the possible aspects underlying the query is also rewarded, a possible solution is to extend existing metrics and compute their expected value given the likelihood of different aspects. This is precisely the idea behind the so-called *intent-aware* metrics for diversity evaluation [1]. In particular, given a ranking of documents $\mathcal{R}(q)$ and a set of relevant aspects $\mathcal{Q}(q)$ for a query q, a traditional evaluation metric $\mathit{Eval}(\mathcal{R}(q))$ can be cast into an intent-aware metric according to:

$$\mathit{Eval}\text{-IA} \equiv \sum_{a \in \mathcal{Q}(q)} \mathrm{P}^*(a|q)\, \mathit{Eval}(\mathcal{R}(q)|a), \tag{8}$$

where $\mathrm{P}^*(a|q)$ is the 'true' probability of observing a relevant aspect $a \in \mathcal{Q}(q)$, while $\mathit{Eval}(\mathcal{R}(q)|a)$ evaluates the ranking $\mathcal{R}(q)$ with respect to this aspect.

Despite being well established and validated [11], diversity metrics assume that a ranking of homogeneous search results is used as input. To cope with the presence of heterogeneous results (e.g., documents, images, videos, maps) in the increasingly prevalent aggregated search interfaces of modern search engines, we propose to generalise intent-aware metrics into a framework for evaluating diversity across multiple search verticals. In particular, we define an *aggregated intent-aware* (AIA) metric as the expected value of the corresponding intent-aware metric across multiple verticals, according to:

$$\mathit{Eval}\text{-AIA} \equiv \sum_{v \in \mathcal{V}(q)} \mathrm{P}^*(v|q) \sum_{a \in \mathcal{Q}(q|v)} \mathrm{P}^*(a|q,v)\, \mathit{Eval}(\mathcal{R}(q)|a,v), \tag{9}$$

where $\mathrm{P}^*(v|q)$ is the 'true' probability of observing the vertical v given the query q, $\mathrm{P}^*(a|q,v)$ is the 'true' probability of observing a relevant aspect $a \in \mathcal{Q}(q|v)$, and $\mathit{Eval}(\mathcal{R}(q)|a,v)$ now evaluates the ranking $\mathcal{R}(q)$ with respect to each vertical v and each aspect a identified in light of v. This formulation provides a framework for leveraging a wealth of existing evaluation metrics in order to synthesise the relevance and diversity of a whole page of results [3]. As concrete instantiations of Equation (9), we introduce aggregated intent-aware versions of the most widely used metrics for diversity evaluation, namely, ERR-IA [8] and α-DCG [13]:

$$\text{ERR-AIA} \equiv \sum_{v \in \mathcal{V}(q)} \mathrm{P}^*(v|q) \sum_{a \in \mathcal{Q}(q|v)} \mathrm{P}^*(a|q,v)\, \text{ERR}(\mathcal{R}(q)|a,v), \tag{10}$$

$$\alpha\text{-DCG-AIA} \equiv \sum_{v \in \mathcal{V}(q)} \mathrm{P}^*(v|q) \sum_{a \in \mathcal{Q}(q|v)} \mathrm{P}^*(a|q,v)\, \alpha\text{-DCG}(\mathcal{R}(q)|a,v). \tag{11}$$

Note that the normalised version of both ERR-AIA and α-DCG-AIA (i.e., nERR-AIA and α-nDCG-AIA, respectively) requires producing an optimal re-ranking of $\mathcal{R}(q)$, which is an NP-hard problem, as discussed in Section 5. Nevertheless, the greedy approach in Alg. 1 can be used for this purpose, without noticeable loss in practice [7]. Also note that Equations (10) and (11) only penalise redundancy within each individual vertical, but not across multiple verticals. In practice, we assume that similar results of the same type (e.g., two videos about the same event) may be redundant, but similar results of different types (e.g., a video and a news story covering the same event) may be actually complementary.

In order to produce a realistic test collection for aggregated search result diversification, 'true' estimations of the likelihood of verticals could be derived from a large sample of the query logs of an aggregated search engine, while the likelihood of different aspects could be estimated from the logs of individual verticals. Lastly, the relevance of a search result could be judged with respect to the 'true' aspects identified from the vertical providing this result. Besides enabling the evaluation of this newly proposed problem, such a collection would benefit ongoing research on both search result diversification and aggregated search.

7 Conclusions

We have proposed the aggregated search result diversification problem, with the aim of satisfying multiple information needs across multiple search verticals. To empirically motivate this new problem, we have analysed the ambiguity of real search queries submitted to four different verticals of a commercial search engine. Our results support the need for aggregated diversification, by showing that the nature of query ambiguity varies considerably across different verticals. Moreover, we have proposed a probabilistic approach for aggregated diversification, by extending current state-of-the-art diversification approaches to tackle query ambiguity in multiple search verticals. Lastly, we have proposed an evaluation framework for this new problem, by generalising existing metrics.

By laying the foundations of aggregated search result diversification, we have bridged current research in the vigorous fields of search result diversification and aggregated search. Nevertheless, we believe we have only scratched the surface of a very promising new field. With the availability of suitable shared test collections, this work can be further expanded in several directions, encompassing alternatives for modelling and evaluating approaches to this new problem.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: WSDM. pp. 5–14 (2009)
2. Arguello, J., Diaz, F., Callan, J., Crespo, J.F.: Sources of evidence for vertical selection. In: SIGIR. pp. 315–322 (2009)
3. Bailey, P., Craswell, N., White, R.W., Chen, L., Satyanarayana, A., Tahaghoghi, S.: Evaluating whole-page relevance. In: SIGIR. pp. 767–768 (2010)
4. Beitzel, S.M., Jensen, E.C., Lewis, D.D., Chowdhury, A., Frieder, O.: Automatic classification of web queries using very large unlabeled query logs. ACM Trans. Inf. Syst. 25(9) (2007)
5. Callan, J.: Distributed information retrieval. In: Croft, W.B. (ed.) Advances in Information Retrieval, chap. 5, pp. 127–150. Kluwer Academic Publishers (2000)
6. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR. pp. 335–336 (1998)
7. Carterette, B.: An analysis of NP-completeness in novelty and diversity ranking. In: ICTIR. pp. 200–211 (2009)
8. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: CIKM. pp. 621–630 (2009)
9. Chen, H., Karger, D.R.: Less is more: probabilistic models for retrieving fewer relevant documents. In: SIGIR. pp. 429–436 (2006)

10. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 Web track. In: TREC (2009)
11. Clarke, C.L.A., Craswell, N., Soboroff, I., Ashkan, A.: A comparative analysis of cascade measures for novelty and diversity. In: WSDM. pp. 75–84 (2011)
12. Clarke, C.L.A., Craswell, N., Soboroff, I., Cormack, G.V.: Overview of the TREC 2010 Web track. In: TREC (2010)
13. Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: SIGIR. pp. 659–666 (2008)
14. Damak, F., Kopliku, A., Pinel-Sauvagnat, K., Boughanem, M.: A user study to evaluate the utility of verticality and diversity in aggregated search. Tech. Rep. 2, IRIT (2010)
15. Deselaers, T., Gass, T., Dreuw, P., Ney, H.: Jointly optimising relevance and diversity in image retrieval. In: CIVR. pp. 1–8 (2009)
16. Diaz, F.: Integration of news content into web results. In: WSDM. pp. 182–191 (2009)
17. Diaz, F., Arguello, J.: Adaptation of offline vertical selection predictions in the presence of user feedback. In: SIGIR. pp. 323–330 (2009)
18. Diaz, F., Lalmas, M., Shokouhi, M.: From federated to aggregated search. In: SIGIR. p. 910 (2010)
19. Gollapudi, S., Sharma, A.: An axiomatic approach for result diversification. In: WWW. pp. 381–390 (2009)
20. Hand, D.J., Smyth, P., Mannila, H.: Principles of data mining. MIT Press (2001)
21. Khuller, S., Moss, A., Naor, J.S.: The budgeted maximum coverage problem. Inf. Proc. Lett. 70(1), 39–45 (1999)
22. van Leuken, R.H., Garcia, L., Olivares, X., van Zwol, R.: Visual diversification of image search results. In: WWW. pp. 341–350 (2009)
23. Murdock, V., Lalmas, M.: Workshop on aggregated search. SIGIR Forum 42, 80–83 (2008)
24. Paramita, M.L., Tang, J., Sanderson, M.: Generic and spatial approaches to image search results diversification. In: ECIR. pp. 603–610 (2009)
25. Ponnuswami, A.K., Pattabiraman, K., Wu, Q., Gilad-Bachrach, R., Kanungo, T.: On composition of a federated web search result page: using online users to provide pairwise preference for heterogeneous verticals. In: WSDM. pp. 715–724 (2011)
26. Rafiei, D., Bharat, K., Shukla, A.: Diversifying Web search results. In: WWW. pp. 781–790 (2010)
27. Santos, R.L.T., Macdonald, C., Ounis, I.: Exploiting query reformulations for Web search result diversification. In: WWW. pp. 881–890 (2010)
28. Song, R., Luo, Z., Nie, J.Y., Yu, Y., Hon, H.W.: Identification of ambiguous queries in Web search. Inf. Process. Manage. 45(2), 216–229 (2009)
29. Spärck-Jones, K., Robertson, S.E., Sanderson, M.: Ambiguous requests: implications for retrieval tests, systems and theories. SIGIR Forum 41(2), 8–17 (2007)
30. Sushmita, S., Joho, H., Lalmas, M., Villa, R.: Factors affecting click-through behavior in aggregated search interfaces. In: CIKM. pp. 519–528 (2010)
31. Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., Yahia, S.A.: Efficient computation of diverse query results. In: ICDE. pp. 228–236 (2008)
32. Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: SIGIR. pp. 115–122 (2009)
33. Zhai, C., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: SIGIR. pp. 10–17 (2003)