On the Suitability of Intent Spaces for IR Diversification

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

ABSTRACT
Recent developments in Information Retrieval diversity are based on the consideration of a space of information need aspects, a notion which takes different forms in the literature. The choice of a suitable aspect space for diversification is a critical issue when designing an IR diversification strategy, which has not been explicitly addressed to some depth in the literature. This paper aims to identify relevant properties of the aspect space which may help the system designer in making a suitable choice in selecting and configuring this space, and diagnosing malfunctions of the diversification algorithms. In particular, we identify the mutual information between aspects and documents as a meaningful magnitude, in terms of which anomalous cases can be characterized. We further seek to discern favorable cases through a combination of theoretic and empirical analysis.

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

General Terms
Algorithms, Measurement, Performance, Experimentation, Theory.

Keywords
Diversity, query intent, query aspect spaces, mutual information.

1. INTRODUCTION
IR diversity theory, diversification algorithms, and evaluation methodologies are based on the consideration of a gap between a user need expression and the complete, precise actual user need [1,8,12,14,18]. In order to grasp this gap of uncertainty, search enhancement methods and metrics proposed in the area introduce a space of user need subunits or features –for use of a generic word– upon which the search is diversified and evaluated for diversity. These features have taken different forms, motivations, and names in the research literature, such as query interpretations, query intents, query aspects, nuggets, subqueries, subtopics, categories, etc. Common to the different approaches and angles, these information need features are understood to be unobserved variables –as are true information needs and relevance themselves– from the retrieval system point of view.

The models and approaches proposed in this area seek to capture or approximate these unobserved features by some form of proxy space. The nature, source and procedures to obtain these proxies have been diverse in the literature. Agrawal et al [1] use categories from the Open Directory Project (ODP) taxonomy; Santos et al [12] extract subqueries from query reformulations issued by search engines; Rafiei et al [11] test Wikipedia disambiguation entries and ODP as sources for aspect extraction, using a query log to determine the relation between queries and categories; Welch et al [17] also use Wikipedia disambiguation and investigate the use of Wordnet term relationships; He et al [10] use automatically created query clusters; Campannini et al [2] extract query specializations from query logs; query subtopics are provided manually in the TREC diversity task [6]; and so forth. We shall henceforth use the term "aspect" to refer generically to the space of query features for diversification –as far as the different forms this space may take can viewed generically.

In a practical setting, there are thus different alternatives in the nature, source, and approach for the definition, extraction, and handling of the aspect space for diversification, and some may be more effective than others. The comparison of alternatives, their potential advantages and drawbacks, or what makes a suitable choice, has not been systematically addressed per se in the literature. Different prototypical aspect sources for diversification have been tested in empirical studies nonetheless (most notably Santos et al [13] explicitly comparing different aspect sources on TREC data), and we find the analysis of aspect configurations worthy of further research. In this broad context, we focus here on a particular characteristic of aspect extraction approaches, namely the grain size and distributional properties of the resulting aspect spaces, which are typically configurable even within the frame of a specific aspect source, representation, and extraction strategy.

It has been observed that diversification algorithms may degrade to no diversification, or random diversification under certain conditions of the aspect distribution over queries and documents [17]. Several options are often available to the developer when designing a diversification algorithm in practical situations, some of which may easily drive the algorithm towards the "bowling gutters" of either randomness or innocuousness. In this paper we analyze the properties of an aspect space that determine these situations. The broad motivation of our research is the definition of criteria to assess the suitability of aspect spaces for diversification. Within this general goal, we study and characterize distributional properties of the aspect space that determine extreme cases, and present theoretical and empirical observations of the intermediate spectrum.

In our study, we analyze the effect of aspect spaces in terms of their *potential for change* in the diversification of retrieval system results, rather than their specific *potential for improvement*, which is highly domain-dependent. The potential for change that an aspect space enables can be seen as a proxy of the discriminative power of the aspect space for diversification. This does naturally not guarantee the quality of diversified results. Whether a diversification strategy properly takes advantage of this power and room for change is a matter of the quality of the strategy itself, and other properties of the aspect space (beyond the quantitative room for change they enable), which are outside the scope of this study. There is nonetheless a strong link between the room for change and the room for improvement by diversification (which is the ultimate underlying concern), inasmuch as the latter requires the former, which motivates this study.

DDR'12, February 12, 2012, Seattle, Washington, USA.
ACM 978-1-4503-0747-5/12/02

Table 1. Decomposition of the objective function of three state of the art diversification algorithms into a diversity component $\eta(d,S)$ and an aggregative function $\varphi_q(x,y)$. In the latter, x is the relevance seeking component (which is defined by the baseline retrieval function), and y is the diversity component.

Diversification algorithm	Diversity component $\eta(d,S)$	Aggregative function $\varphi_q(x,y)$
MMR [3]	$\min_{d' \in S} dist(d,d')$	$\lambda\, x + (1-\lambda) y$
IA-Select [1]	$(c,x) \rightarrow p(c\|d) \prod_{d' \in S} (1 - p(c\|d')\, x)$	$\sum_{c \in \mathcal{A}} p(c\|q)\, x\, y(c,x)$
xQuAD [12]	$c \rightarrow p(d\|c) \prod_{d' \in S} \big(1 - p(d'\|c)\big)$	$\lambda\, x + (1-\lambda) \sum_{c \in \mathcal{A}} p(c\|q)\, y(c)$

2. DIVERSIFICATION ALGORITHMS

Most approaches to IR diversification in the literature state the diversification goal as the maximization of some objective function that reflects the degree of diversity of a set of retrieved documents [1,3,5,12,19,20]. The maximization is found to be an NP-hard problem [4], the solution to which is commonly approximated by a greedy algorithm. The algorithm uses itself an objective function (not necessarily the same as the initial one), and incrementally builds a reranked version S of the baseline document set R by iteratively picking one document at a time which maximizes this objective function. This function, which we shall denote as φ, generally takes a document and a set of documents as input, that is. $\varphi: \mathcal{D} \times \mathcal{P}(\mathcal{D}) \rightarrow \mathbb{R}$, where $\mathcal{D}$ denotes the document collection, and the diversification procedure can be generically described as shown in Algorithm 1.

Algorithm 1 Generic greedy diversification

```
Diversify (R)
    S ← ∅
    while R ≠ ∅
        d ← argmax_{d∈R} φ(d,S)
        S ← S ∪ {d}
        R ← R − {d}
    return S
```

Two components can be commonly identified within the greedy objective function: the baseline retrieval function (or some monotonic derivation thereof), which we shall denote as $f_q: \mathcal{D} \rightarrow \mathbb{R}$, and a diversity (or perhaps more precisely, novelty) component $\eta: \mathcal{D} \times \mathcal{P}(\mathcal{D}) \rightarrow \mathbb{R}$, where $\eta(d,S)$ measures the lack of redundancy of a document d with respect to a set of documents S. In terms of these two components, φ can be expressed as $\varphi(d,S) = \varphi_q\left(f_q(d), \eta(d,S)\right)$, with φ_q being monotonically increasing with respect to its two inputs. A convenient property for the diversity component is that $\eta(d,\emptyset)$ be constant for all d, whereby $\varphi(d,\emptyset) \propto f_q(d)$, and so the top document in the baseline remains at the top in the diversified ranking.

When diversity is defined in terms of query aspects, η can be seen as returning a function on the set of user need aspects. Furthermore, the diversity component may depend on the baseline retrieval function (as in IA-Select [1]). In such case, for maximum generality, we may consider $\eta: \mathcal{D} \times \mathcal{P}(\mathcal{D}) \rightarrow [\mathcal{A} \times \mathbb{R} \rightarrow \mathbb{R}]$, where $\mathcal{A}$ is the set of all aspects, and $[\mathcal{A} \times \mathbb{R} \rightarrow \mathbb{R}]$ denotes the set of all functions from $\mathcal{A} \times \mathbb{R}$ to $\mathbb{R}$. The first input of such functions is an information need aspect $c \in \mathcal{A}$, and the second is expected to be a baseline retrieval score $x \in \mathbb{R}$,where $x = f_q(d)$ for some d. Table 1 shows example instantiations of this scheme for diversification algorithms in the literature.

Since in our study we are interested in the effect of aspect spaces, we focus on diversity approaches that are based on such notion. As far as schemes such as MMR [3] do not use query aspects, our analysis does not apply to them (note however that the distance function in MMR could be defined in terms of aspects, but we will not explore that direction here).

As to the aspect-based approaches, one may observe that the respective diversity components $\eta(d,S)$ of IA-Select and xQuAD share some common characteristics. They are both based on the conditional distribution between documents and aspects (or vice-versa), and a product of probability complements (with an additional multiplying parameter in IA-Select). Despite the differences between them, both formulations may exhibit similar behavior patterns with respect to probabilistic relations between aspects and documents, as we analyze in the next section.

The implications of the analysis that follows may exceed these two specific diversification algorithms, and would apply to other variations that are based on a similar probabilistic assessment of the redundancy between the documents to be reranked and the partial greedy ranking.

3. ASPECT SPACE INFORMATIVENESS

We study the suitability of aspect spaces from the point of view of their informational properties. As we have reviewed in the previous section, conditional distributions between aspects and documents lie at the core of the analyzed diversification methods of interest for our study. We therefore analyze the informational properties of this dependence, as a potential major criterion for the suitability of an aspect space. More specifically, we investigate whether and how the strength of this dependence may affect the resulting behavior of the diversifiers.

In Information Theory, the strength with which two variables depend on each other is measured by their mutual information. In the case of documents and aspects, this is defined as:

$$I(\mathcal{A};\mathcal{D}) = \sum_{c \in \mathcal{A}, d \in \mathcal{D}} p(c,d) \log \frac{p(c,d)}{p(c)p(d)}$$

where $\mathcal{A}$ and $\mathcal{D}$ above denote random variables ranging over aspects and documents respectively –as a shorthand we abuse notation by using the same symbol for the random variables and the set where they take values. A mutual information zero indicates that the two variables are independent, and higher mutual information reflects progressively stronger degrees of dependence.

3.1 Minimum mutual information

As a general trend, the higher $I(\mathcal{A};\mathcal{D})$ is, the larger are the differences in $p(d|c)$ and $p(c|d)$ between documents for a fixed aspect c. It is natural to figure out that a high mutual information is desirable, since it helps better discriminate between documents in terms of their covered aspects and hence their mutual diversity. This is true to some extent, but the issue is somewhat more complex, as we see next.

Indeed, if $I(\mathcal{A};\mathcal{D}) = 0$, we have $p(d|c) = p(d)$, whereby in xQuAD the diversity component does not depend on c and becomes $\eta(d,S) = p(d)\prod_{d' \in S}(1 - p(d'))$. If we assume a uniform document prior, this term becomes constant for all documents $d \in R$, the objective function becomes $\varphi_q\left(f_q(d), \eta(d,S)\right) \propto f_q(d)$, and the diversification algorithm thus degrades to the original ranking. That is, the diversifier has no effect and leaves the baseline ranking unchanged. Similarly, with $I(\mathcal{A};\mathcal{D}) = 0$ we have $p(c|d) = p(c)$, whereby in IA-Select the diversity component $\eta(d,S)$ does not depend on d and is therefore constant for all documents $d \in R$. It is easy to see that in this situation we also have $\varphi_q\left(f_q(d), \eta(d,S)\right) \propto f_q(d)$, thus degrading again to the baseline with no diversification.

In other words, with zero mutual information, the observation of a document says nothing about the information need aspects that are being covered. Any aspect is covered to the same extent by all documents. Given a document, any aspect is as probable as any other, and the aspect space is therefore useless for diversification.

3.2 Maximum mutual information

After the preceding analysis, one might hypothesize that a maximal mutual information might then be a desirable situation, but this is not quite the case. The maximum informativeness is reached when for any document d there is a unique aspect with $p(c_d|d) = 1$, and $p(c|d) = 0$ for any other $c \neq c_d$. We also have that inversely, for any aspect c there is a single document d_c with $p(d_c|c) = 1$, all other documents $d \neq d_c$ having $p(d|c) = 0$. That is, with maximum mutual information, observing a document is equivalent to observing its single aspect, and vice versa. When this is the case, it can be seen that $I(\mathcal{A};\mathcal{D}) = H(\mathcal{A}) = H(\mathcal{D})$ and this is equal to $\log|\mathcal{D}|$ if the prior document distribution is uniform.

In this extreme, we may see that in IA-Select we have $\eta(d,S)(c) = p(c|d)$ which equals 1 for the unique aspect c_d that is covered by d, and 0 for all other document-aspect pairs. Hence, $\varphi_q\left(f_q(d), \eta(d,S)\right) = p(c_d|q)\, f_q(d) = p(d|q) f_q(d)$. Inasmuch as $p(d|q)$ and $f_q(d)$ are monotonically related (e.g. when f_q is a retrieval function based on statistical language models), the result again degrades to the baseline with no diversification –even if the baseline retrieval function diverged from $p(d|q)$, the effect would not be that of diversification, but a mix of retrieval strategies both seeking the maximization of total returned relevant documents, regardless of their diversity or the mutual dependency of their covered relevance aspects.

In xQuAD, we have a similar situation with $\eta(d,S)(c) = p(d|c)$ being 1 for the aspect c_d uniquely covered by d, and 0 for all other aspects. This results in $\varphi_q\left(f_q(d), \eta(d,S)\right) = \lambda\, f_q(d) + (1-\lambda)p(d|q)$, where again, $f_q(d)$ and $p(d|q)$ are equivalent or push in the same direction –that of null diversification.

These effects match the intuition: each aspect is exclusively covered by a unique document, therefore any set of documents covers as many different aspects as documents it contains. There is total absence of redundancy between documents, and any set has maximal diversity, which cannot be improved any further (which is naturally an illusion by effect of a poor aspect space choice).

3.3 Practical considerations

We have thus seen that indeed the degree of dependence between documents and aspects can be identified as a major factor in the choice of an aspect space for diversification. The minimum and maximum extremes may seem too obvious to fall into, but they are easier to get close to than it might seem. For instance, if the aspect space is taken from available document features or classification schemes, it is not unusual to find long-tailed distributions of classes among documents. Such distributions may in practice approach minimum mutual information for the few most frequent classes, and maximum in the vast majority of long tail ones. This becomes still more extreme by the class subsampling involved in working with the small set of top *n* documents to be diversified.

Just to mention an anecdotic but representative example, we run into this type of situation when considering, for instance, movie directors as the aspect space to diversify movie recommendation in the MovieLens dataset.[1] In the small version of this collection (about 1,600 movies), most directors appear in a unique movie, and only 3% have more than 5 films –and the chances to find a few repeated directors within the top *n* recommended movies is even lower. Mutual information is excessive, so that diversification algorithms just do not work. Features such as the movie language or country approach the opposite extreme, where a single most frequent feature value (English and USA respectively) accounts for more than half the films (over 70% and 60% respectively), resulting in insufficient mutual information, with less than 5% of the remaining feature values having any use for diversification. In contrast, movie genre is an example of an effective, more balanced space for diversification in this dataset (as shown e.g. in [15,16]).

Long-tailed distributions are also typical of collaborative tagging environments, which are being increasingly used as a large scale document labeling resource for IR techniques (see e.g. [9] as just one example). As a general trend, aspect distributions emerging from spontaneous (social, etc.) phenomena often exhibit a power law or long tail structure. In contrast, editorial labeling (ODP, Wikipedia disambiguation pages, etc.) tends to display a more balanced structure, often by intentional design –since balance is typically part of classification scheme design guidelines.

The distributional considerations can also be seen from a query-specific point of view. It is not necessary that $I(\mathcal{A};\mathcal{D})$ reaches extremes on the whole collection and the whole aspect space for diversification to be ineffective. Given a query q, it suffices that $I(\mathcal{A}_q;R_q)$ be extreme for all the above analysis to hold, where $\mathcal{A}_q = \{c \in \mathcal{A} | p(c|q) > 0\}$ is the set of possible aspects of q, and R_q is the set of relevant documents for q. In this perspective, $I(\mathcal{A}_q;R_q) = 0$ means q has minimum ambiguity (all relevant documents correspond to a single aspect), and $I(\mathcal{A}_q;R_q) = 1$ means the opposite (each relevant document covers a unique aspect). This formalizes the rationale that meaningful diversification is not possible for extremely ambiguous or extremely specific queries.

4. THE INTERMEDIATE SPECTRUM

Having analyzed the extremes of the dependence strength between documents and aspects, the question remains: what is the ideal balance in mutual information (in the terms stated in this study, i.e. the potential for change)?

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

The question is complex to answer. Once departing from the extremes, the degree of aspect-document dependence strength influences the resulting amount of change in different interrelated ways, which are difficult to capture and describe formally. An increase in dependence strength might, for instance, cause further changes in rank position, but shorter in distance, thereby balancing the global amount of change in the ranking –as measured by some rank distance measure. It is possible though to analyze and observe specific aspects in this direction, as we discuss next.

4.1 Aspect distribution skewness

Multinomial and long-tail aspect distributions are two prototypical cases of the distributions that are frequently found in practical application domains. For a fixed collection size and a fixed number of aspects, assuming only one aspect per document (that is, $p(c|d) = 1_{c_d}(c)$ for some $c_d \in \mathcal{A}$ –where c_d might also be covered by other documents besides d) and a uniform prior document distribution, it can be seen that a uniform aspect distribution maximizes the mutual information of aspects and documents. Indeed, under these assumptions, we have $I(\mathcal{A};\mathcal{D}) = H(\mathcal{A})$ which is maximum with $H(\mathcal{A}) = \log_2|\mathcal{A}|$ when the prior aspect distribution is uniform, i.e. each aspect is covered by the same number of documents. The uniform aspect distribution (i.e. an even number of covering documents per aspect in the collection) is thus an upper bound of mutual information for a fixed aspect space size.

As the distribution moves away from uniform, the mutual information decreases monotonically. We may illustrate the evolution towards increasingly long-tailed distributions as, for instance, an adjusted power law defined by $|c_k| \sim 1 + C\, k^{-\alpha}$, where $|c_k| = |\{d \in \mathcal{D} | p(c_k|d) > 0\}| > 0$. This distribution is uniform for $\alpha = 0$ and its skewness increases with α. Let us denote by $\mathcal{A}_\alpha$ an aspect distribution following a power law of exponent α. It can be seen –we omit the details here– that the entropy of the power law is monotonically decreasing, from $I(\mathcal{A}_0;\mathcal{D}) = H(\mathcal{A}_0) = \log_2|\mathcal{A}_0|$ to $I(\mathcal{A}_\alpha;\mathcal{D}) = H(\mathcal{A}_\alpha) \to \log_2|\mathcal{D}| - \frac{|\mathcal{D}|-|\mathcal{A}|+1}{|\mathcal{D}|}\log_2(|\mathcal{D}| - |\mathcal{A}| + 1)$ as $\alpha \to \infty$, an extreme at which all aspects but one are covered by a single document, and one aspect is covered by all the rest of documents. It can be seen that this expression is $I(\mathcal{A}_\infty;\mathcal{D}) \sim \frac{|\mathcal{A}|}{|\mathcal{D}|}\log_2|\mathcal{D}|$ with a negligible error. This situation is actually reached at some finite $\alpha_\infty < \infty$.

Let us denote by $\left|\mathcal{A}_\alpha^{max}\right|$ the number of aspects that maximizes the distance for a power law aspect distribution of exponent α. It is easy to see that the number of aspects needed for the extreme $\mathcal{A}_\infty$ to reach the same mutual information as the distance-maximizing uniform $\mathcal{A}_0$ is $\left|\mathcal{A}_\infty^{max}\right| \sim \log_2\left|\mathcal{A}_0^{max}\right| |\mathcal{D}|/\log_2|\mathcal{D}| \gg \left|\mathcal{A}_0^{max}\right|$ if $\left|\mathcal{A}_0^{max}\right| \ll |\mathcal{D}|$, which means a considerably higher number of aspects are needed in the skewed distribution to level up with a uniform distribution. For instance, if $\left|\mathcal{A}_0^{max}\right| = 100$ and $|\mathcal{D}| = 100{,}000$, we have $\left|\mathcal{A}_\infty^{max}\right| \sim 40{,}000$, almost half as many aspects as documents in the collection (an unrealistically high and impractical number of aspects). Intermediate distribution skewness $0 < \alpha < \alpha_\infty$ results in intermediate situations $\left|\mathcal{A}_0^{max}\right| < \left|\mathcal{A}_\alpha^{max}\right| < \left|\mathcal{A}_\infty^{max}\right|$ between these extremes.

We may thus consider that for a fixed number of aspects, a uniform aspect distribution is a safe option, and so are close deviations around that. In the next section we shall therefore focus on this case as a reasonable representative of a suitable option, yet conveniently simple and tractable for analysis.

4.2 Aspect space size

Taking the simple uniform case described in the previous section, we address the assessment of aspect spaces in terms of their grain size (which is the only parameter of a uniform distribution). Assuming a uniform aspect prior, with a single aspect per document for simplicity, which implies $I(\mathcal{A};\mathcal{D}) = \log|\mathcal{A}|$, we know that both extremes $|\mathcal{A}| = 1$ (zero mutual information) and $|\mathcal{A}| = |\mathcal{D}|$ (maximum mutual information) result in diversifiers degrading to no changes in the baseline ranking.

It is not trivial to analytically describe what happens in between, and it is therefore not obvious what point may be optimal. However, we may observe the evolution of the diversification behavior empirically by running diversifications on simulated aspect spaces. Three variables describe our simulation setting for diversification:

- The total number of aspects $|\mathcal{A}|$ covered by the collection.
- The collection size $|\mathcal{D}|$.
- The top n depth for diversification of the baseline ranking (i.e. the rank position at which diversification stops).

Given a combination of values for these parameters, our simulation procedure consists of randomly assigning aspects (based on the distribution induced by the number of aspects and the collection size) to n ranking positions, occupied by simulated documents (which just consist of an integer ID). Then we run the diversification algorithms using a constant baseline ranking function $f_q(d) = 1$. This isolates the diversification effect from any particular baseline retrieval system, in order to focus on the behavior of the diversity component only. Finally, we measure the Pearson rank correlation between the diversified ranking and the original baseline –the latter being thus equivalent to a random diversification. And we repeat the simulation across variations in the respective axes of the three variables. In all the simulations, the ranking distances were averaged for smoothness by a ten-fold cross-validation.

Figure 1 shows the distance produced by IA-Select and xQuAD for different aspect grain sizes (1 to 300), with a fixed $n = 100$ and $|\mathcal{D}| = 100{,}000$. It can be observed how the diversification distance starts at zero for a single aspect covered by all docu-

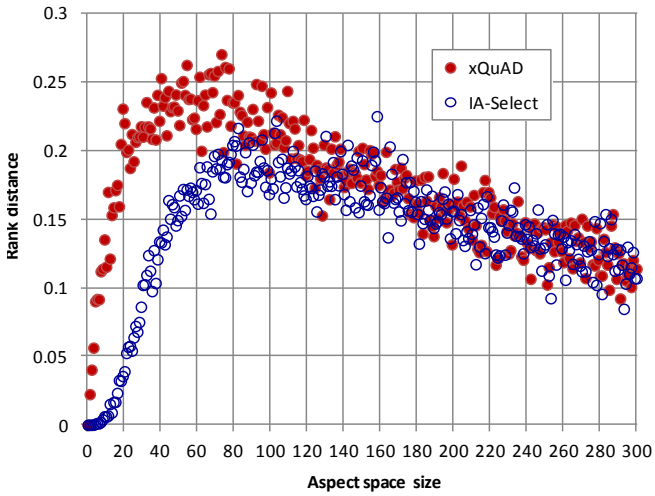

Figure 1. Distance between initial and diversified ranking resulting from different aspect space sizes, for two state of the art diversification algorithms. The distance between ranked lists is measured as $d(R_1, R_2) = 1 - sim(R_1, R_2)$, where sim is the Spearman rank correlation. The collection size is fixed at $|\mathcal{D}| = 100{,}000$, with $n = 100$ documents being reranked.

ments, and grows fast, reaching a maximum around 60-80 aspects for xQuAD, and around 100 aspects for IA-Select. The difference between both diversifiers is due to the fact that when neutralizing the baseline retrieval function in the novelty component, IA-Select treats redundancy in a binary way: once an aspect c is covered by one document in the reranked subset S at some point in the diversification, any further document covering the aspect is considered as totally redundant (since $p(c|d)$ is either 0 or 1), regardless of the number of documents covering c in S. xQuAD on the contrary captures degrees of redundancy for the covered aspects, as reflected in $p(d|c)$ which is non-binary. As a result, IA-Select needs more aspects to keep diversifying, whereas xQuAD can take advantage further times of the same few aspects.

If we continue taking further aspects beyond the optimum, the amount of diversification slowly degrades until eventually becoming zero, when using as many aspects as documents in the collection.

We thus see that for these settings, 60-80 aspects would be optimal for xQuAD, and 80-100 for IA-Select. We now turn to examine whether these optima depend on the number of documents being reranked, or the collection size. Figure 2 shows the variation of the optimum number of aspects (the number that results in the higher reranking distance) for a range of collection sizes, from 1,000 to 100,000. Given the regular behavior of the distance vs. the number of aspects, and the range for maximization being reasonably small and discrete, we find these optima by simple brute force, scanning a large enough range of aspect space sizes. It can be seen that the optimum fluctuates, but does not seem to depend on the collection size. The average optimum is around 100 aspects for IA-Select, and 60 for xQuAD, not far from the previous results.

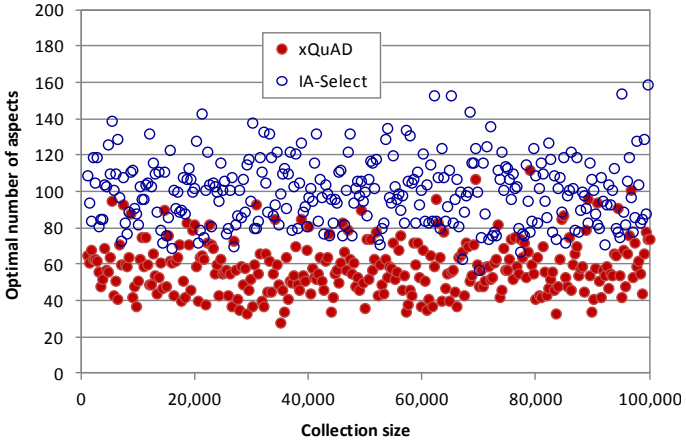

Figure 2. Optimum number of aspects maximizing the ranking distance of diversification for different collection sizes. The size of the document set to be diversified is fixed at $n = 100$.

Finally, we study the dependence of the optimal number of classes for different top n cuts for diversification. Figure 3 shows the optimal aspect space size for n ranging from 50 to 150. A linear growth trend can be observed in the optimal number of aspects with respect to n. The linear fit of the plots gives a slope of ~0.9 for IA-Select, and ~0.6 for xQuAD. This is not far from the ratio optimum aspect space size / diversification depth in the previous observations, and a more precise convergence of the estimates may be expected by increasing the range variation scale.

In the experiments so far, the aspects are sampled from their background distribution, which results in a particular number of aspects being covered in the set of documents to be reranked, which is smaller, in general, than the total number of aspects in the collection. In order to observe the effect that a specific number of different aspects in the result set has in the diversification, we repeat the previous simulation but this time we force a fixed number of aspects in the top n documents, evenly distributed (i.e. same number of documents covering each aspect). Figure 4 displays the result. The number of aspects has here a more direct effect on the result, where the linear relation to the size of the result set is clearer. In this case the slope is ~0.7 for both diversifiers. Note that the number of aspects in the result set and in the whole collection are different variables and the sampling in this experiment and the previous one is not equivalent, hence the slight difference in the observed slope.

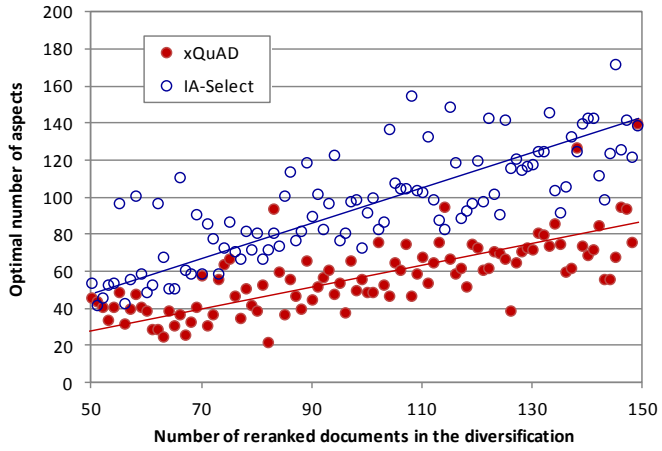

Figure 3. Optimum number of aspects maximizing the ranking distance of diversification for different sizes of the document set being diversified. The collection size is fixed at $|\mathcal{D}| = 100,000$.

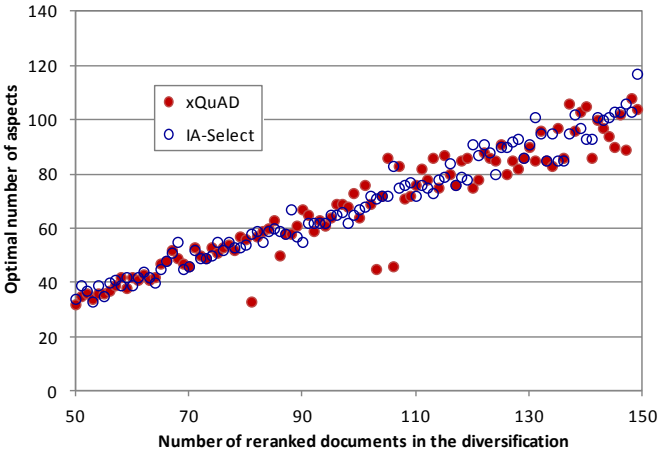

Figure 4. Optimum number of aspects maximizing the ranking distance for different result set sizes. This simulation is similar to the one displayed in Figure 3, but the number of aspects is forced to occur exactly as such within the result set, instead of randomly sampling aspects from the whole collection.

5. DISCUSSION AND CONCLUSION

We have addressed in this paper a relevant question when one designs a diversification framework, namely the choice and configuration of the aspect space. We show that this is a fundamental decision that strongly impacts the results and the potential range for the action of a diversifier.

We have characterized extreme cases that are easier to run into than one might think, and we have described their effect. We further

analyze the intermediate spectrum between both ends aiming to provide some observations of trends by means of empirical simulations. The underlying goal of the experiments is to seek some invariant that helps recognize an optimum aspect space in terms of such an invariant, independently from the variability of other factors such as the collection size or the number of documents to be reranked.

In our study the observed invariant for the ideal prior aspect distribution seems to be –under the simplifying assumptions of the experiments– $p(c) \sim \rho/n$ (since the optimal number of classes shows a linear growth trend $|\mathcal{A}| \sim n/\rho$), where n is the size of the set to be diversified, and ρ seems to range around a constant value for each diversification algorithm. As future work, we see interest in finding a more general invariant in terms e.g. of the mutual information itself, which would require much lighter –if any– assumptions on the specific aspect distribution.

Our study and experiments are intentionally neutral with respect to factors in the system or collection properties, other than the aspect coverage by documents and the collection size. These properties would add up their share in the effect of diversification. Pursuing this direction would be a system specific investigation in principle, although generic simplified steps could be introduced such as, for instance, taking $f_q(d) = 1/|\mathcal{D}|$ rather than $f_q(d) = 1$. Another extension worth being addressed is the analysis of other dimensions such as the degree of aspect coverage overlap (i.e. documents covering more than one aspect), the introduction of which we envision to be reasonably feasible in our framework of study.

Beyond its theoretical interest, the question researched here has a direct practical motivation and potential uses, in common decisions such as the choice of one among several available document features for diversification, the granularity of subqueries from a query log, or the appropriate level (i.e. the number of classes) one should use from a classification taxonomy such as ODP. This paper presents some steps in addressing a question which we see as important enough to warrant further research, beyond the study presented here.

6. ACKNOWLEDGMENTS

This work is supported by the Spanish Government (TIN2011-28538-C02-01), and the Government of Madrid (S2009TIC-1542).

7. REFERENCES

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

[2] Capannini, G., Nardini, F. M., Perego, R., Silvestri, F. Efficient diversification of search results using query logs. 20th Int. Conf. on World Wide Web (WWW 2011). Hyderabad, India, March 2011, 17-18.

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

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

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

[6] Clarke, C. L. A., Craswell, N., and Soboroff, I. Overview of the TREC 2009 Web Track. TREC 2009, Gaithersburg, MD, USA.

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

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

[9] Clements, M., de Vries, A. P., and Reinders, M. J. T. The task-dependent effect of tags and ratings on social media access. ACM Trans. on Information Systems 28(4), November 2010.

[10] He, J., Meij, E., and de Rijke, M. Result diversification based on query-specific cluster ranking. Journal of the American Society for Information Science and Technology 62(3), March 2011, 550-571.

[11] Rafiei, D., Bharat, K., and Shukla, A. Diversifying web search results. 19th Int. Conf. on World Wide Web (WWW 2010). Raleigh, NC, USA, April 2010, 781-790.

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

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

[14] Spärck-Jones, K., Robertson, S. E., and Sanderson, M. Ambiguous Requests: implications for retrieval tests, systems and theories. ACM SIGIR Forum 41(2), December 2007, 8-17.

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

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

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

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

[19] Zhang, M. and Hurley, N. Avoiding Monotony: Improving the Diversity of Recommendation Lists. 2nd ACM Int. Conf. on Recommender Systems (RecSys 2008). Lausanne, Switzerland, October 2008, 123-130.

[20] Ziegler, C-N., McNee, S. M., Konstan, J. A., and Lausen, G. Improving recommendation lists through topic diversification. 14th Int. Conf. on World Wide Web (WWW 2005). Chiba, Japan, 2005, 22-32.