Noname manuscript No.
(will be inserted by the editor)

The Impact of Result Diversification on Search Behaviour and Performance

Maxwell · Azzopardi · Moshfeghi

Received: date / Accepted: date

Abstract *Result diversification* aims to provide searchers with a broader view of a given topic while attempting to maximise the chances of retrieving relevant material. Diversifying results also aims to reduce search bias by increasing the coverage over different aspects of the topic. As such, searchers should learn more about the given topic in general. Despite diversification algorithms being introduced over two decades ago, little research has explicitly examined their impact on search behaviour and performance in the context of *Interactive Information Retrieval (IIR)*. In this paper, we explore the impact of diversification when searchers undertake complex search tasks that require learning about different aspects of a topic *(aspectual retrieval)*. We hypothesise that by diversifying search results, searchers will be exposed to a greater number of aspects. In turn, this will maximise their coverage of the topic (and thus reduce possible search bias). As a consequence, diversification *should* lead to

David Maxwell
School of Computing Science
University of Glasgow
Scotland
E-mail: d.maxwell.1@research.gla.ac.uk

Leif Azzopardi
Department of Computer and Information Sciences
University of Strathclyde
Scotland
E-mail: leif.azzopardi@strath.ac.uk

Yashar Moshfeghi
Department of Computer and Information Sciences
University of Strathclyde
Scotland
E-mail: yashar.moshfeghi@strath.ac.uk

performance benefits, regardless of the task, but how will diversification affect search behaviours and search satisfaction?

Based on *Information Foraging Theory (IFT)*, we infer two hypotheses regarding search behaviours due to diversification, namely that *(i)* it will lead to searchers examining fewer documents per query, and *(ii)* it will also mean searchers will issue more queries overall. To this end, we performed a within-subjects user study using the *TREC AQUAINT* collection with 51 participants, examining the differences in search performance and behaviour when using *(i)* a non-diversified system (*BM25*) versus *(ii)* a diversified system (BM25+*xQuAD*) when the search task is either *(a)* ad-hoc or *(b)* aspectual. Our results show a number of notable findings in terms of search behaviour: participants on the diversified system issued more queries and examined fewer documents per query when performing the aspectual search task. Furthermore, we showed that when using the diversified system, participants were: more successful in marking relevant documents, and obtained a greater awareness of the topics (i.e. identified relevant documents containing novel aspects). These findings show that search behaviour is influenced by diversification and task complexity. They also motivate further research into complex search tasks such as aspectual retrieval – and how diversity can play an important role in improving the search experience, by providing greater coverage of a topic and mitigating potential bias in search results.

Keywords Diversification · User Study · User Behaviour · User Behavior · Search Performance · Aspectual Retrieval

1 Introduction

Interactive Information Retrieval (IIR) is a complex (and often exploratory) process [19] in which a searcher issues a variety of queries as a means to explore the topic space [23]. Often, such tasks are *aspectual* in nature, where an underlying goal is to find out about the different facets, dimensions or aspects of the topic. This type of task is often referred to as *aspectual retrieval*. While aspectual retrieval has been heavily studied in the past (during the TREC Interactive Tracks [31]), there has been renewed interest in the search task as it represents a novel context to explore the idea of *"search as learning"* [9]. In this context, the goal of the system is to help the searcher learn about a topic [9] – and in doing so, the number of aspects that the searcher finds indicates how much they learned during the process [39]. If the goal of the system is to help people learn about a topic, then by returning results that are more diverse in nature and presenting a broader view on the topic, these changes *should* help searchers learn more about the said topic. This reasoning suggests that employing *diversification* will lead to an improved search and learning experience [39].

While there have been numerous diversification algorithms developed and proposed over the years [4,6,36,37,45], the focus here has been on addressing the problem of intents, rather than how diversification affects complex search

tasks, such as *ad-hoc* or aspectual retrieval. In this paper, we perform one of the first investigations into the influence and impact of result diversification on search behaviour and search performance when performing different search tasks (ad-hoc or aspectual). Our focus is on understanding how behaviours – in particular, how searching and stopping behaviours – change under the different conditions. We ground our study by drawing upon *Information Foraging Theory (IFT)* [32] (see Section 2) which derives the following hypotheses regarding diversification when performing aspectual search tasks: *(i)* diversification will lead to searchers examining fewer documents per query; and either *(ii)* issuing more queries, or *(iii)* completing the task in less time. However, these hypotheses seem to be counter to our intuition. If a system provides a more diversified set of results, then searchers *should* be able to exploit the diversification of results and find more varied aspects by examining more documents for a given query – and thus issue fewer queries. In order to explore the validity of the IFT hypotheses and test our intuitions, we designed a 2×2 within-subjects user study, where participants were tasked to learn about four different topics under the following conditions, using: *(i)* a non-diversified system (*BM25*); versus *(ii)* a diversified system (BM25+*xQuAD* [36]), and when the search task is either: *(a)* ad-hoc retrieval, where they need only to find relevant documents; or *(b)* aspectual retrieval, where they need to find documents that are both relevant and different – i.e. covering new, unseen aspects of the topic. We perform our experiments in the context of learning about a topic to write a report where participants use a standard search interface to search the *TREC AQUAINT* news collection.

2 Background and Motivation

When searching for information, searchers pose a varying number of queries, examine *Search Engine Result Pages (SERPs)*, and examine a number of documents (if any) before issuing a new query, or stopping their search altogether. This may be because they have found enough information to satisfy their underlying information need, have run out of time, were dissatisfied, or simply gave up their search [10, 11, 15, 26, 33, 44]. Prior work has shown that there are a variety of different factors that can influence an individual's search behaviours. Of particular relevance to this paper, it has been shown that different search tasks influence the search behaviour of searchers [23].

An interesting task that has not received much attention of late is aspectual retrieval. Aspectual retrieval is a type of search task that is concerned with the identification of different *aspects* of a given topic. This task type differs from traditional ad-hoc retrieval in the sense that ad-hoc retrieval is concerned only with what constitutes a *relevant* document to a given topic, rather than identifying relevant documents and whether they are *different* to what has been seen previously. A relevant and different document will contain unseen *aspects* associated with the topic in question. As an example, take the topic *Wildlife Extinction*, one of the topics in the *TREC 2005 Robust Track* [42]. If

the searcher under an ad-hoc search task finds several documents concerning `'Pandas in China'`, these would then all be considered relevant. However, for the aspectual retrieval task where *different* examples must be found, the first document concerning `'Pandas in China'` is considered relevant/useful. Other aspects (in this case, species of endangered animals) would then need to be found, such as `'Sumatran Rhinos in Malaysia'`, `'Crested Ibis in Japan'`, `'Black-Necked Crane in India'` etc.

Aspectual retrieval found significant traction in the *TREC Interactive Tracks* from 1997–2002. The overarching, high-level goal of the TREC Interactive Tracks was to investigate searching – as an interactive task – by examining the process of searching, as well as the outcome [31]. Historically, interaction was considered from the inaugural *TREC-1* in 1993 [13], where one group investigated interactive searching under *"interactive query mode"* within an ad-hoc search task. From TREC-6 (1997) to TREC 2002, a substantial volume of research was directed toward the development of systems and search interfaces that:

1. assisted searchers in exploring and retrieving various aspects of a topic, such as cluster-based and faceted interfaces that explicitly showed different aspects [29, 41];
2. tiles and stacks to organise documents [14, 17, 18, 20]; and
3. mechanisms to provide query suggestions that would ultimately lead to different search paths [40, 22].

However, a disappointing conclusion from this initiative was that little difference was observed between such experimental systems and the standard control systems (typically represented by the *ten blue links*), both in terms of behaviour and performance [43].

As work on aspectual retrieval subsided, work related to determining the intent of a searcher's query began to take hold. Here, the goal of this problem is to diversify the results retrieved with respect to the original query [35]. Thus, this addresses the problem of *ambiguity* for short, impoverished queries. This led to a series of diversification algorithms (and intent-aware evaluation measures) being proposed, changing focus from the interface to the underlying algorithms and their evaluation measures (e.g. [1, 4–6, 16, 34, 36, 37, 45, 46]). While there have been numerous studies investigating the effectiveness of diversification algorithms for the problem of *intents* (e.g. one query, several interpretations), little work has looked at studying how such algorithms apply in the context of aspectual retrieval (e.g. one topic, many aspects). This is mainly because a majority of these algorithms were developed after the TREC Interactive Track was concluded in 2002.

More recently, there has been a growing interest in new, more complex and exploratory search tasks – especially in the aforementioned context of *"search as learning"* [9]. Syed and Collins-Thompson [39] hypothesised that diversifying the results presented to searchers would improve their learning efficiency and that this would be observed by a change in vocabulary expressed in their queries. This study motivates our interest in examining the effects of

diversification (or not) when considering the task of aspectual retrieval (where a searcher needs to learn about different aspects). Thus, in this paper, our aim is to better understand how search performance and search behaviour changes when people undertake different types of search task – using search systems that diversify the ranked results, and those that don't. To ground this study, we first consider how search behaviour is likely to change by generating hypotheses given Information Foraging Theory.

2.1 Information Foraging Theory

To motivate our hypotheses, we draw upon *Information Foraging Theory (IFT)* [32], and, in particular, the *patch model* (a constituting model of IFT) to ground our research and provide insights into how search behaviours should change. The patch model predicts how long foragers will stay in a given *patch* before moving to a new patch. Under this model, the analogy with an information seeker is as follows. Moving between patches is like expressing a new query (and thus incurs a moving/querying cost) while staying within a patch is akin to assessing documents. Fig. 1 graphically shows the predictions given the theory for two systems (diversified and non-diversified), and the corresponding hypothetical gain curves. In the top left plot (Fig. 1 **(a)**) where a non-diversified system is being used, the gain curve for the ad-hoc retrieval task is higher, as any relevant document contributes to the overall level of gain. However, for the aspectual task, the gain curve is lower. This is because similar relevant documents (discussing an already observed aspect) do not contribute to the overall gain accrued by the searcher.

From IFT, the optimal stopping point would be different between the two tasks. Graphically, we can find this point by drawing a line from the origin to the tangent of the gain curve – the red and blue dots indicating the optimal stopping points for ad-hoc and aspectual retrieval, respectively. Thus, IFT suggests that when subjected to the non-diversified system, searchers will examine more documents per query for aspectual retrieval tasks than when compared to ad-hoc retrieval tasks.

In Fig. 1 **(b)** where a diversified system is being used, the gain curves for ad-hoc and aspectual retrieval will be similar. This is because relevant but different documents are discovered earlier. In the case of ad-hoc topic retrieval, these relevant (even if different) documents will still contribute to the overall gain. In the case of aspectual retrieval, the relevant and different documents will also contribute to the searcher's overall gain – but only up to the point where the documents are similar to the previously retrieved material (i.e. after this point, similar but relevant documents do not contribute to the gain). Therefore, IFT appears to suggest that similar stopping behaviours would be observed when searchers utilise a system that diversifies results.

Fig. 1 **(c)** shows the predicted stopping behaviour for the aspectual task, where we have plotted the aspectual gain curves from the system plots described above. Interestingly, IFT suggests that searchers will stop sooner when

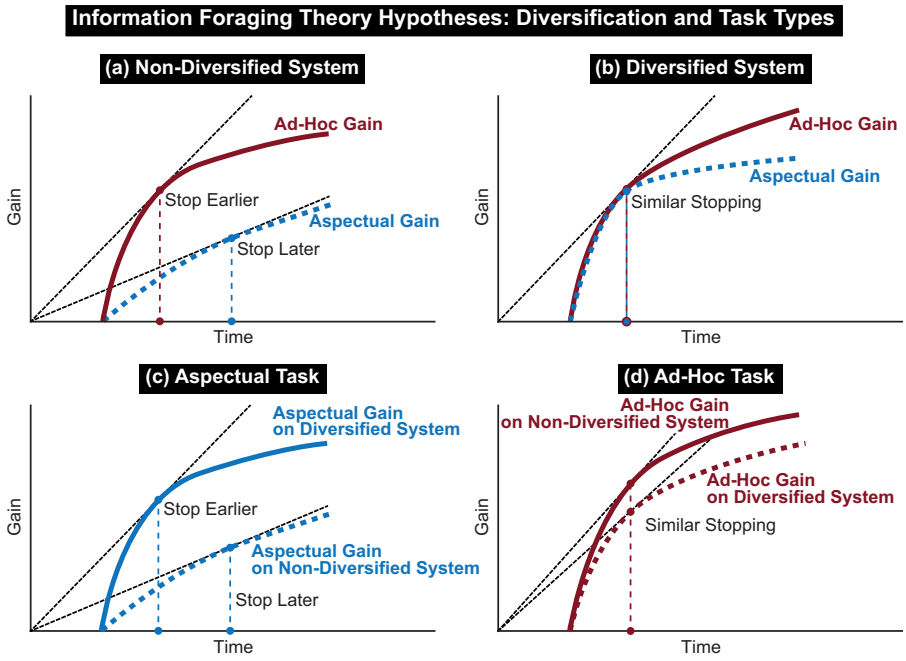

Fig. 1 A graphical depiction, using Information Foraging Theory, of how stopping behaviour is likely to be affected with: a system that diversifies results **(a)**; a system that does not diversify **(b)**; under an aspectual retrieval task **(c)**; and under an ad-hoc retrieval task **(d)**.

using the diversified system. Therefore, if searching for the same amount of time, searchers would thus issue more queries. Finally, Fig. 1 **(d)** shows the predicted stopping behaviour for ad-hoc retrieval tasks, where again we have plotted the respective gain curves on each system. Note that the gain curve for the diversifying system may be a little lower as some non-relevant but different material may ascend up the rankings – but as can be seen, we expect little difference between systems. Therefore, the expected gains and behaviours that we hypothesise will be approximately the same. Consequently, IFT suggests that there will be little difference regarding stopping behaviours between the two systems under ad-hoc retrieval tasks.

In contrast to the hypotheses from IFT, our intuitions suggested that searchers would behave differently such that: when using a standard, non-diversified system, searchers would be more likely to issue a greater number of queries because they would likely need to issue more queries to explore the topic. Indeed, Kelly et al. [23] showed that more complex search tasks require a greater number of queries to provide sufficient coverage of the topic. For example, if a searcher submits a query such as `'protecting Pandas in China'` that retrieves relevant material about pandas, we would expect them to only select one or two examples before issuing another query – rather than examining more results given the current query. In the case of ad-hoc topic retrieval, we would expect that they would issue fewer queries, and examine more doc-

uments per query. This is because they don't need to find multiple aspects of the said topic. However, when using a diversified system that attempts to promote different aspects of the topic, we would expect that a searcher's behaviour would change such that when undertaking aspectual retrieval, they would issue fewer queries and examine more documents per query. Below we test our intuitions against the theory.

2.2 Research Questions and Hypotheses

The primary **research question** of this study is: *how does diversification affect the search performance and search behaviour of people when performing ad-hoc topic and aspectual retrieval tasks?* Based on the theoretical analysis above using IFT, we can formulate the specific following hypotheses regarding performance and behaviour.

- Considering aspectual retrieval tasks, diversification will lead to:
  (H1) fewer documents examined per query; and
  (H2a) more queries issued; or
  (H2b) a decrease in task completion time.
- Considering ad-hoc retrieval tasks, diversification will lead to:
  (H3) no difference in the documents examined; and
  (H4) no difference in the number of queries issued.

However, the contradiction between IFT and our intuitions also provides an ulterior hypothesis. Furthermore, given the findings presented by Syed and Collins-Thompson [39], we also hypothesise that diversification will lead to a greater awareness of the topic, regardless of the task. Therefore, we expect searchers to encounter and find a greater variety of aspects when using the diversified system.

3 Experimental Method

To address our research questions and examine the hypotheses as outlined in Section 2.2, we conducted a within-subjects experiment with two factors: system and task. For the system factor, our baseline control system was based on BM25 (no diversification) and a diversified system based on BM25, re-ranked with xQuAD [36]. For the task factor, we used the standard ad-hoc retrieval task and compared against the aspectual retrieval task. This resulted in a 2×2 factorial design. Therefore, each participant completed four different search tasks, one in each of the four conditions (see below). Conditions were assigned using a Latin square rotation to minimise any ordering effects.

- **(D.As)** A **diversified system** with an **aspectual retrieval** task.
- **(ND.As)** A **non-diversified system** with an **aspectual retrieval** task.
- **(D.Ad)** A **diversified system** with an **ad-hoc retrieval** task.
- **(ND.As)** A **non-diversified system** with an **ad-hoc retrieval** task.

3.1 Corpus and Search Topics

For this experiment, we used the *TREC AQUAINT* test collection that contains over one million articles from three newswires, collected over the period 1996-2000. The three newswires were: the *Associated Press (AP)*; the *New York Times (NYT)*; and *Xinhua*. From the TREC 2005 Robust Track [42], we selected five contemporary topics that have been used in prior works [3, 25, 28]. These were: 341 *(Airport Security)*; 347 *(Wildlife Extinction)*; 367 *(Piracy)*; 408 *(Tropical Storms)*; and 435 *(Curbing Population Growth)*. These topics were chosen based on evidence from a previous user study with a similar setup, where it was shown that the topics were of similar difficulty and interest [25]. Topic 367 was used as a practice topic. The remaining four topics were used as part of the main experimental study.

3.2 Aspectual and Ad-Hoc Retrieval Tasks

Participants were asked to imagine that they needed to learn about a number of topics on which they were to write a report on. Given a topic, they were further instructed on whether to focus on finding *relevant* articles in the case of ad-hoc retrieval, or *relevant articles that discussed different aspects* of the topic in the case of aspectual retrieval. For example, for the *Airport Security* topic, participants were required to learn about the efforts taken by international airports to better screen passengers and their carry-on luggage under the ad-hoc retrieval task. For the aspectual retrieval task, they were also asked to find relevant documents that are different and mention *new, previously unseen* airports. Thus, participants were explicitly instructed to find a number of examples from different airports, as opposed to a similar or the same example based in the same airport multiple times.

Participants were instructed to find and save at least four *useful* documents. Depending upon the task being undertaken, *useful* related to a document being either relevant or relevant and different.

3.3 Relevance Judgments and Aspects

For each topic, we used the corresponding TREC QRELs from the TREC 2005 Robust Track to provide the relevance judgements for the study. However, to assess how many aspects were retrieved, we needed to commission additional labels, as existing labels were not available for all the selected topics. For each topic, we first examined the topic descriptions to identify what dimensions could be considered aspects of the topic. We noted that for each topic there were at least two ways this could be achieved: entity- or narrative-based. For example, in the topic *Population Growth*, a document could be relevant if it stated the country (entity-based) or measure that was taken to reduce population growth (narrative-based).

For this study, it was decided that we should focus on entity-based aspects. This was because *'different narratives'* were subject to greater interpretation than *'different entities'*. For each relevant document, two assessors extracted different aspects, and we found that there were substantially higher agreements (95% vs 67%) between assessors across the entity based aspects: (341) airports; (347) species; (367) vessels; (408) storms; and (435) countries; as opposed to the more narrative-based aspects: (341) the security measures taken; (347) the protection and conservation efforts; (367) the acts of piracy; (408) the death and destruction; and (435) the population control methods. Entity-based aspects that we considered for each topic are listed below.

- **Topic 341** ***(Airport Security)*** Different *airports* in which additional security measures were taken, e.g. *John F Kennedy International Airport*, *Boston Logan International Airport*, or *Leonardo Da Vinci International Airport*.
- **Topic 347** ***(Wildlife Extinction)*** Different *species of endangered animals* under protection by states, e.g. *golden monkey*, *Javan rhino*, or *Manchurian tiger*.
- **Topic 367** ***(Piracy)*** Different *seaworthy vessels* that were boarded or hijacked, e.g. *Petro Ranger*, *Achille Lauro*, or *Global Mars*.
- **Topic 408** ***(Tropical Storms)*** Different *tropical storms* where people were killed or there was major damage, e.g. *Hurricane Mitch*, *Typhoon Linda* or *Tropical Storm Frances*.
- **Topic 435** ***(Curbing Population Growth)*** Different *countries* where population control methods were employed, e.g. *China*, *India* or *Zimbabwe*.

The total number of aspects identified for each topic were: 14 for 341; 168 for 347; 18 for 367; 43 for 408; and 26 for 435. Created judgments were saved in the TREC Diversity QREL format, as used by already established evaluation tools that consider aspectual retrieval, such as `ndeval`[1].[2]

3.4 Baseline System and User Interfaces

Two experimental search systems were developed. These were identical except regarding branding/logo and the retrieval algorithm used. First, in terms of branding, we created two fictional retrieval system names, ***Hula Search*** and ***YoYo Search***, for which different colour schemes were devised. The names were chosen as they were not associated with any major retrieval system (to the best of our knowledge), nor did they imply that one of the systems performed better than the other. The colour schemes were chosen to provide the greatest difference in visual appearance to those with colourblindness (two variants of colourblindness, *protanopia* and *deuteranopia*, were both considered). This was to ensure that participants could later indicate which system that they

[1] `https://trec.nist.gov/data/web/10/ndeval.c`

[2] In the interests of promoting reproducibility and repeatability, the aspectual judgements created as part of this study are available for download at `https://git.io/fpAX8`.

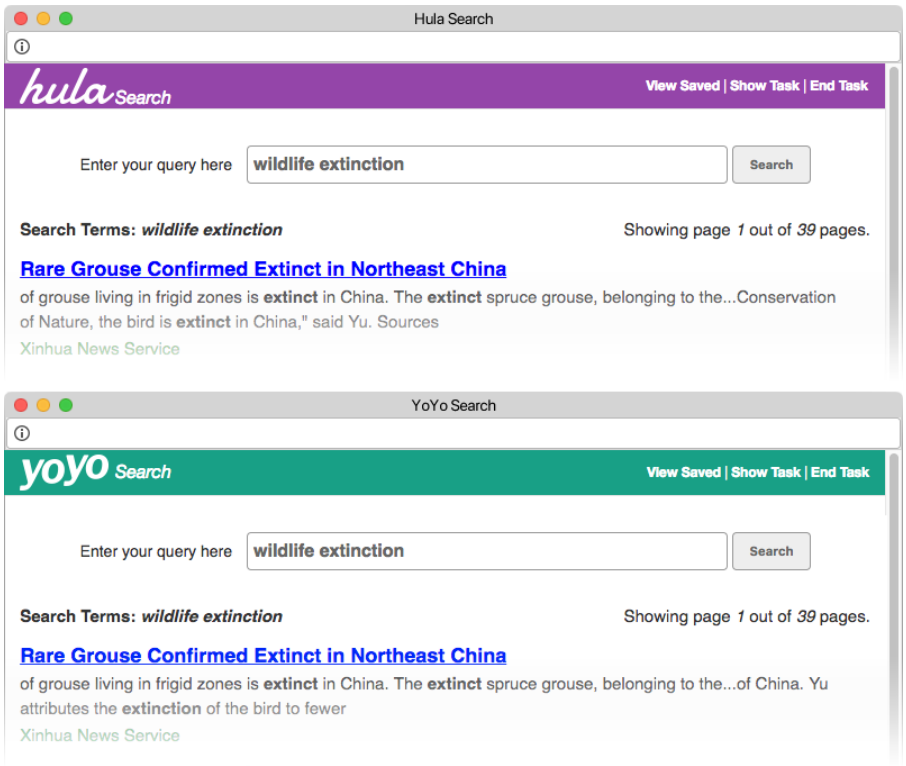

Fig. 2 The two different retrieval systems and their interfaces, as used in this study. The top screenshot shows *Hula Search* (non-diversified, baseline), with *YoYo Search* (diversified) underneath. Note the different colour schemes that were designed to emphasise the fact that different search systems were being used, without creating too much of a visual distraction.

preferred, etc. Screenshots of the two systems in action are provided in Fig. 2. Note that a generic ***NewsSearch*** system – complete with a blue header – was used for the practice task. This was to allow participants to familiarise themselves with how to mark and save documents, and how the search functionality worked – all without becoming favourably or unfavourably biased to one particular system.

For the underlying retrieval system, we used the *Whoosh Information Retrieval (IR)* toolkit.[3] We used BM25 as the retrieval algorithm ($b = 0.75$), but with an implicit `AND`ing of query terms to restrict the set of retrieved documents to only those that contained all query terms provided. This was chosen as most retrieval systems implicitly `AND` terms together. BM25 served as the baseline control for the non-diversified system condition.

[3] Whoosh can be accessed at `https://pypi.python.org/pypi/Whoosh/`.

Input: Original ranking, *existingResults*
Diversification depth ($k = 30$)
$\lambda = 0.7$, weighting for diversification scoring component

Output: Manipulated array of results, diversified to depth k

Helpers: *getEntities(x,y,z)* Returns an array results x, consisting of entities present in documents in array x from range y to z
getLength(x) Returns the length of array x
getUnseenEntities(x,y) Returns entities in document x that have not yet been observed in ranked document array y
sortByScore(x) Sorts documents array x by *.score* in descending order
array.pop() Removes the top entry from an array, returning its value

```
entities = [];
newRankings = [];
i = 1;
newRankings[0] = existingResults.pop();
while i ≤ k do
    entities = getEntities(existingResults, 0, i-1);
    j = 0;
    while j ≤ getLength(existingResults) do
        newEntityCount = getUnseenEntities(document, existingResults);
        existingResults[j].score = existingResults[j].score + (λ·newEntityCount);
        j = j + 1;
    end
    sortByScore(existingResults);
    newRankings[i] = existingResults.pop();
    i = i + 1;
end
```

Alg. 1: The algorithm employed to diversify results. Input for this algorithm assumes a ranked list of results, as ranked by the BM25 baseline discussed in the narrative. The diversification algorithm presented is based upon the xQuAD framework [36]. Included in the pseudo-code above are a list of input parameters (including the original BM25 ranking) and simple helper functions used within the algorithm.

3.5 Diversifying Results

For the diversified system, we used BM25 as outlined above to provide the initial ranking and then used xQuAD [36] to diversify the ranking. xQuAD has been shown to provide excellent performance for web intent-based diversification. The algorithm used is presented in Alg. 1, complete with a description of the various inputs, the output and helper functions used.

To select a reasonable approximation for the algorithm's two tuneable input parameters, i.e. k (how many documents to re-rank) and λ (how much focus on diversification), we performed a parameter sweep using a set of 715 training queries from a prior user study [28]. Results from this pilot study are presented in Table 1. As can be seen from the table, we explored a range of k and λ values, with 10–50 trialled for k, and 0.1–1.0 for λ. We selected $k = 30, \lambda = 0.7$ as this configuration provided the best results ($P@10 = 0.36$, $\alpha DCG@10 = 0.075$, $AR@10 = 6.61$, see below for AR) in terms of performance and efficiency – i.e. a higher k only slightly increased performance but took longer to compute.

Table 1 Table illustrating the effects of varying λ and diversifying rank cutoff k using the diversification algorithm as outlined in Alg. 1. Values in the table represent the number of new aspects found (aspectual recall, *AR*) in the top 10 documents after re-ranking on average, over 715 queries issued from a prior user study [28]. When $\lambda = 0.0$, diversification ***(D)*** is not applied – this configuration therefore enjoys the same performance as our non-diversified ***(ND)***, baseline system that utilises BM25.

	Diversification Cutoff (k)				
Weighting (λ)	**10**	**20**	**30** ***(D)***	**40**	**50**
0.0 ***(ND)***	3.64				
0.1	3.64	4.94	5.51	5.95	6.37
0.3	6.58	6.58	6.64	6.59	6.59
0.5	6.58	6.58	6.58	5.58	6.58
0.7 ***(D)***	6.56	6.56	**6.61**	6.51	6.60
0.9	6.52	6.52	6.61	6.57	6.63
1.0	6.63	6.63	6.59	6.61	6.56

3.5.1 Aspectual Retrieval Measures

To measure the performance of the retrieval systems and participants with respect to aspectual retrieval, we utilise two additional measures, reported in tandem with traditional IR measures such as $P@k$. These measures are *Aspectual Recall (AR)* and αDCG.

Aspectual recall is defined by Over [30] as *"...the fraction of the submitted documents which contain one or more aspects."* Given a ranking, AR can be computed by summing the number of *unseen, novel aspects* for a topic up to some rank k, and dividing by k. Therefore, this provides us with a useful means of determining how successful systems and searchers were at identifying documents containing novel aspects.

αDCG also provides us with this ability. An extension of *Discounted Cumulative Gain (DCG)* [21], αDCG employs a position-based searcher model [8] – similar in nature to its functionally related counterparts, such as $\alpha NDCG$. αDCG takes into account the position at which a document is ranked, along with the aspects contained within the documents. It ranks by rewarding newly-found aspects, and penalising redundant (seen) aspects geometrically, discounting all rewards with a discounting rank function. As the name may imply, α is a tuneable parameter that controls the severity of redundancy penalisation. As used in prior TREC experimentation [31], all values of αDCG reported in this paper are computed with $\alpha = 0.5$.

3.6 Experimental Procedure

Participants were provided with a link to an online experimental system that first presented the information sheet regarding the experiment. This was then

followed by the consent form which participants needed to agree to in order to proceed.[4] Participants were then asked to fill in a brief demographics survey before undertaking the practice task to familiarise themselves with the interface. Once comfortable with the system, participants could then proceed to undertake the four search tasks. Depending upon the Latin square rotation, participants would then be provided with one of the four conditions on one of the four topics. For each task, participants first completed a pre-task questionnaire. They then moved onto the search task itself. After completion of the task, they were asked to fill in a post-task questionnaire. After completing all four search tasks, participants were then asked to fill in an exit questionnaire regarding which system they preferred. The experiment would then conclude.

3.7 Recruitment and Controls

Participants for the experiment were recruited via the crowdsourcing platform *Amazon Mechanical Turk (MTurk)*. Previous work has shown that crowdsourced studies provide similar results as traditional lab-based user studies [24, 47]. However, the caveat here is that this is true only if sufficient controls are in place. If not, workers may attempt to game the system and could, in theory, complete the task poorly [12]. Therefore, it is important to ensure that quality control mechanisms are in place to mitigate this risk.

First, we ensured that the device being used was desktop-based, and the device's screen resolution was at least 1024×768 or greater in size. As the experiment was conducted via a web browser (i.e. *Chrome*, *Firefox*, *Safari*, *Edge*, etc.), we wanted to ensure that only the controls provided by the experimental apparatus were used. The experimental system was therefore launched within a popup window of size 1024×768. Within the popup, all other browser navigation controls (i.e. back buttons, etc.) were disabled (to the best of our abilities). The experimental system was tested on several major (aforementioned) browsers, across a range of different operating systems. This gave us confidence that similar experiences would be had across different systems.

Based upon the suggestions from prior work [12, 47], workers were only permitted to begin the experiment on the MTurk platform that:

- were from the United States;
- were native English speakers;
- had a *Human Intelligence Task (HIT)* acceptance rate of at least 95%; and
- had at least 100 HITs approved.

Requiring the latter two criteria increased the likelihood of recruiting individuals who wanted to maintain their reputation, and would be more likely to complete the study in a satisfactory manner.

Participants were informed that from our pilot study, it would take approximately 7-10 minutes to find at least four relevant documents per task – and

[4] Ethics approval was sought before the experiment from the Department of Computer and Information Sciences at the University of Strathclyde (ethics approval number 622).

the duration of the entire experiment would be approximately 40-50 minutes. Since we did not impose any time constraints on how long they searched for, we imposed an accuracy-based control. We informed participants that their accuracy in identifying relevant material would be examined and that they should aim to find four useful documents with at least 50% accuracy (using the TREC relevance judgments as the gold standard). Note that from a previous lab-based study [28] for this set of topics, the accuracy of participants was between 25% and 40% on average, depending on the topic. While we stipulated a higher accuracy, this was to motivate participants to work diligently. Since we anticipated the experiment to take just under an hour, participants were compensated with nine dollars (USD).

In all, 64 participants performed the experiment. However, 13 participants were omitted either because they failed to complete all search tasks (five participants were removed), failed to mark at least four documents (two participants were removed), or spent less than two minutes per task and failed to retrieve any relevant documents (six participants were removed).

Of the 51 participants who successfully completed the experiment, 26 females and 25 males participated. The average age of the participants was 38.66 years ($min = 20$; $max = 71$; $stdev = 11.43$). 22 of the participants reported having a bachelor's degree or higher, with the remaining 29 possessing an associate degree or lower. All participants but one expressed *Google* as their everyday retrieval system of choice. All participants indicated that they conducted many searches for information via a retrieval system per week. Nearly three-quarters of the participants (i.e. 38 participants) reported using a mouse for the experiment, with the remaining 13 using some form of a trackpad.

3.8 Logging and Measures

Below we note the measurements taken while participants used the experimental system. Our system logged a variety of different events associated with querying and assessing. The generated logs permitted us to measure three different aspects of the search user experience, being: *(i)* interaction; *(ii)* performance; and *(iii)* time.

Interaction Measures included the number of queries issued by participants, the number of documents that were examined, the number of different SERPs viewed, and the depths to which participants clicked on (and hovered over) result summaries. It should be noted that components recorded such as hover depths over result summaries were inferred from the movement of the mouse cursor – eye-tracking equipment was not used in this study. In prior studies, the position of the mouse cursor on the screen has correlated strongly with the user's gaze on the screen [7,38].

Performance Measures included the number of documents that were saved by participants, denoting that they were either relevant (for ad-hoc retrieval), or

relevant and contained new information (for aspectual retrieval). From this, we could also break this number down into the number of documents that were saved and TREC relevant – as well as TREC non-relevant – and $P@k$ measures at varying depths for the performance of the queries issued by the participants. Using the diversity QRELs (generated as per the description in Section 3.3), we were able to determine how well the query performed with respect to how many new entities were in the top k results, and the αDCG scores for each query. In addition, using the list of saved documents, we could also then identify how many entities that participants had found, and how many documents contained one or more unseen entities – both in terms of the context of results of the current query, and the overall search session (over all the queries issued).

From the log data, we could also compute additional performance measures such as the accuracy that searchers reached during each session, as well as the *probabilities of interaction.* In the context of this study, accuracy referred to the ratio of documents that were TREC relevant, versus the total numbers of documents saved. For example, if a searcher saved four documents during a search session, with three of them being TREC relevant, the searcher's accuracy was $3/4 = 0.75$. The interaction probabilities that we considered included: the probabilities of clicking on a result summary link ($P(C)$) – given that it was either TREC relevant ($P(C|R)$) or TREC non-relevant ($P(C|N)$), and the probabilities of marking a document that was clicked ($P(M)$) – again, given that it was either TREC relevant ($P(M|R)$) or TREC non-relevant ($P(M|N)$).

Time-Based measures included the time spent issuing queries (from query focus to issuance), the time spent on a SERP – as well as examining result summaries[5] – and the time spent examining documents. These times could then allow us to compute the total amount of time spent during the session.

3.9 User Experience

To capture their perceived experiences, we asked participants to complete both pre- and post-task surveys for each of the four experimental conditions that they were presented with during the experiment.

Pre-task surveys consisted of five questions, each of which was on a seven-point Likert scale (1 – strongly disagree to 7 – strongly agree). Participants were sought for their opinions on their: *(i)* prior knowledge of the topic; *(ii)* the relevancy of the topic to their lives; *(iii)* their desire to learn about the topic; *(iv)* whether they had searched on this topic before; and *(v)* the perceived difficulty to search for information on the topic.

[5] Result summary times were approximated by dividing the total recorded SERP time by the number of snippets hovered over with the mouse cursor. We believe this is a reasonable assumption to make – network latency issues beyond our control meant that mouse hover events occasionally were delivered at the wrong times, and as such were logged in the incorrect order.

Following the completion of each search task, participants were provided with a post-task survey, again using a seven-point Likert scale for responses. The survey considered aspects of *(i)* their behaviour, and *(ii)* how they felt the system performed. Considering their behaviours, participants were asked for their opinions on:

- how successful they thought they were at completing the task ***(success)***;
- how quickly they felt they completed the task ***(participant speed)***;
- whether they issued different queries to explore the topic ***(queries)***;
- if they only examined a few documents per query ***(documents)***;
- whether they checked each document carefully before saving ***(checks)***; and
- whether they saved more documents than was required, with a minimum of four being required ***(more)***.

Participants were also asked for their opinions on:

- whether they thought the system helped them complete the task quickly ***(system speed)***;
- whether they felt the system made it difficult to find useful information ***(difficulty)***;
- if the system made it easy to complete the task ***(ease)***;
- if they were happy with how the system performed ***(happiness)***;
- whether the system was cumbersome or not ***(cumbersome)***; and
- whether they were confident in the decisions they made ***(confident)***.

Upon completion of the experiment, participants were provided with an exit survey consisting of several questions. Here, we wanted to ascertain which of the two search system offered the best performance, and which one they preferred. Answers were provided on a scale this time from $1-6$, with: 1 denoting *definitely Hula Search* (non-diversified); 3 denoting *slightly Hula Search;* 4 denoting *slightly YoYo Search*; and 6 denoting *definitely YoYo Search* (diversified). We opted not to include a neutral position to force participants into deciding between one of the two systems. We asked participants:

- which system was most informative ***(informative)***;
- which system was more unhelpful ***(unhelpful)***;
- what one was easier to use ***(easiest)***;
- what system was less useful ***(least useful)***;
- what system returned more relevant information ***(most relevant)***;
- what system offered a more diverse set of results ***(most diverse)***; and
- what system they preferred overall ***(most preferable)***.

4 Results

We now address our research questions and hypotheses as outlined in Section 2.2. Both the behaviour and performance of each participant were analysed across each of the four experimental conditions, ***D.As***, ***ND.As***, ***D.Ad***

Table 2 Query statistics and performance measures across both of the experimental systems trialled, ***ND*** (Non-Diversified) and ***D*** (Diversified). Note the significant differences between the diversity-centric measures, $\boldsymbol{\alpha DCG}$ (where $\boldsymbol{\alpha = 0.5}$) and aspectual recall (***AR***), highlighting that the diversification algorithm did indeed provide a more diverse set of results to the participants.

		ND	***D***
	Queries Issued	718	555
	Terms per Query	3.59	3.80
	Unique Terms	345	292
Prec.	***P@5***	0.25 ± 0.01*	0.29 ± 0.01*
	P@10	0.22 ± 0.01	0.24 ± 0.01
αDCG	$\boldsymbol{\alpha DCG}$**@5**	0.02 ± 0.00*	0.04 ± 0.00*
	$\boldsymbol{\alpha DCG}$**@10**	0.03 ± 0.00*	0.04 ± 0.00*
AR	***AR@5***	1.40 ± 0.11*	3.39 ± 0.21*
	AR@10	2.11 ± 0.14*	4.07 ± 0.24*

and ***ND.Ad***. Task (***As.*** vs ***Ad.***) and system (***ND.*** vs ***D.***) effects were also examined. To evaluate these data, ANOVAs were conducted using the conditions, systems and tasks each as factors; the main effects were examined with $\alpha = 0.05$. Bonferroni tests were then used for post-hoc analysis. As discussed in Section 3.5.1, αDCG values are computed using $\alpha = 0.5$. All variance values reported denote the *standard deviation* from the mean.

To begin our analysis, we first examined whether the performance experienced by participants on the two systems was actually different (as indicated by our pilot study). We took the queries participants issued to each system and measured the performance according to αDCG, *aspectual recall (AR)* and precision (see Table 2). Statistical testing confirms that the two systems were significantly different in terms of diversity (i.e. αDCG@10: $F(1, 1272) = 28.74, p < 0.001$, and AR@10: $F(1, 1272) = 55.43, p < 0.001$). P@10 was however not significantly different. This suggests that the reranking promoted relevant and diverse documents, but only in the top 10 results on average. These results gave us confidence that the results presented by the diversified system ***D*** did indeed offer participants with a broader perspective of the topics being examined.

Aside from showing query performance, Table 2 also reports the number of terms issued per query over systems ***ND*** and ***D***. Of the 1273 queries issued, those issued to ***ND*** were shorter on average, with 3.59 terms compared to 3.80 terms for ***D***. However, the vocabulary used by participants issuing queries to ***ND*** was more diverse than ***D*** – queries issued to ***ND*** contained 345 unique terms, compared to 292 for ***D***. This provides our first finding of note. When using ***ND***, participants issued more queries to accomplish their tasks – but these queries were slightly shorter and more varied in terms of vocabulary.

Table 3 Behavioural and performance measures reported across the four experimental conditions (top table), systems and tasks (bottom table).

Actions	Experimental Conditions			
	D.As	***ND.As***	***D.Ad***	***ND.Ad***
#Queries	5.92± 0.88	5.25± 0.80	4.96± 0.74	5.20± 0.69
#SERPs/Query	1.78± 0.14	2.42± 0.24	2.28± 0.31	2.28± 0.20
Documents/Query	3.02± 0.39	3.65± 0.46	3.48± 0.51	3.23± 0.37
Depth/Query	12.85± 1.49	15.73± 1.45	16.19± 2.14	13.94± 1.93
#Saved	5.80± 0.26	5.96± 0.25	5.92± 0.25	5.78± 0.20
#TREC Saved	2.63± 0.22	2.18± 0.23	2.51± 0.23	2.22± 0.22
#TREC Non-Relevant	1.75± 0.22	1.96± 0.23	1.37± 0.22	1.82± 0.23
#Entities Found	7.22± 0.94*	4.31± 0.60*	5.82± 0.77	4.37± 0.59*
#Docs w/ New Entities	3.20± 0.21*	2.35± 0.20*	2.63± 0.23	2.02± 0.18*

Actions	Systems		Tasks	
	ND	***D***	***Ad.***	***As.***
#Queries	5.23± 0.53	5.44± 0.58	5.08± 0.51	5.59± 0.59
#SERPs/Query	2.35± 0.16	2.03± 0.17	2.28± 0.18	2.10± 0.14
Documents/Query	3.44± 0.29	3.25± 0.32	3.36± 0.31	3.34± 0.30
Depth/Query	14.84± 1.58	14.52± 1.31	15.07± 1.44	14.29± 1.47
#Saved	5.87± 0.16	5.86± 0.18	5.85± 0.16	5.88± 0.18
#TREC Saved	2.20± 0.16	2.57± 0.16	2.36± 0.16	2.40± 0.16
#TREC Non-Relevant	1.89± 0.16	1.56± 0.16	1.60± 0.16	1.85± 0.16
#Entities Found	4.34± 0.42*	6.52± 0.61*	5.10± 0.49	5.76± 0.57
#Docs w/ New Entities	2.19± 0.13*	2.91± 0.16*	2.32± 0.15*	2.77± 0.15*

4.1 Observed Behaviours

Interactions. Table 3 presents the mean of: *(i)* the number of queries issued; *(ii)* the number of SERPs that were examined by participants per query; *(iii)* the number of documents examined (clicked) per query; and *(iv)* the click depth (or search stopping depth) per query. Statistical tests reveal no effects across conditions, systems or tasks. However, there are several trends that are worth mentioning. Firstly, we notice that when participants used the diversified system to complete the aspectual retrieval task, they examined fewer documents per query than when completing the same task on the non-diversified system (12.85 vs. 15.73) – which is in line with ***H1***. We also observed that participants issued slightly more queries on the diversified system compared to the non-diversified system with the aspectual retrieval task (5.92 vs. 5.25) – which is in line with ***H2a***. To reiterate, these results were not significant.

Table 4 Interaction probabilities, as observed over the four experimental conditions trialled in this study. Refer to Section 3.8 for an explanation of each probability's meaning. Here, asterisks *(*)* denote that probabilities of interaction were significantly different when compared to other experimental conditions.

Probability	***D.As***	***ND.As***	***D.Ad***	***ND.Ad***
$P(C)$	0.16± 0.01*	0.21± 0.02*	0.16± 0.01*	0.20± 0.01*
$P(C\|R)$	0.27± 0.03	0.30± 0.04	0.25± 0.03	0.31± 0.04
$P(C\|N)$	0.13± 0.02*	0.18± 0.02*	0.13± 0.01*	0.17± 0.02*
$P(M)$	0.67± 0.03	0.66± 0.03	0.70± 0.03	0.71± 0.04
$P(M\|R)$	0.78± 0.04	0.63± 0.05	0.74± 0.04	0.67± 0.05
$P(M\|N)$	0.59± 0.04	0.61± 0.04	0.65± 0.04	0.65± 0.04

Turning our attention to the ad-hoc retrieval tasks, while our hypotheses suggested that there would be no differences in terms of the number of documents examined ***(H3)***, or in the number of queries issued ***(H4)***. This was indeed found to be the case. However, we note that participants when using the diversifying system ***D*** inspected more results than when using the non-diversified system (16.19 vs. 13.94), all while issuing slightly fewer queries (4.96 vs. 5.20). We can see the trade-offs between queries and the number of results inspected per query, where more queries tend to lead to fewer results being examined, and vice versa. This trend suggests that participants when searching on the diversified system for relevance ***(D.Ad)***, may have had to examine documents to greater depths in order to find more relevant material (due to system performance). Alternatively, the system simply encouraged participants to examine to greater depths (which is what we intuitively anticipated when they were searching under the diversified system, ***D***). Either way, we find no conclusive evidence to support the studies main hypotheses – only trends.

Table 4 reports interaction probabilities associated with searcher interactions. These concern the probability of clicking a result summary (or snippet, with the probability denoted by $P(C)$) on a SERP, or the probability of marking a document as relevant (or relevant and new, with the probability denoted by $P(M)$). Also included are the conditional probabilities for the two, based upon whether the document saved or clicked was TREC (R)elevant or (N)on-Relevant. From the table, we can see that there was a significant difference between conditions (and systems, not shown) for the probability of a click, and the probability of clicking on non-relevant items. Comparing systems indicated that participants clicked more when using the non-diversified system, and clicked on more non-relevant documents. However, we did not observe any task effects. This suggests that the non-diversified system led to participants examining more documents, but often more non-relevant documents. This is reflected by the fact that across all the performance measures (see below), participants performed worse when using the non-diversified system.

Table 5 Interaction times across each experimental condition (top table), system and task (bottom table). Included is: the mean total session time *(Total Session)*; the per query time *(Per Query)*; the per document time *(Per Document)*; and the per result summary (snippet) time *(Per Snippet)*. Also included are mean total times. Results are presented in seconds.

	Experimental Conditions			
Time	***D.As***	***ND.As***	***D.Ad***	***ND.Ad***
Total Session	443.65± 45.05	430.50± 38.39	432.18± 49.87	447.55± 47.82
Total Query	45.26± 6.48	47.76± 8.41	46.40± 8.01	43.22± 6.55
Per Query	8.80± 0.89	9.99± 1.21	9.69± 0.79	8.69± 0.57
Total Doc.	162.93± 20.47	144.85± 16.73	139.58± 16.70	152.83± 27.69
Per Document	15.97± 1.96	13.03± 1.01	13.66± 1.02	15.09± 2.20
Per Snippet	1.59± 0.09	1.75± 0.15	1.71± 0.11	1.71± 0.13

	Systems		**Tasks**	
Time	***ND***	***D***	***Ad.***	***As.***
Total Session	439.02± 30.52	437.91± 33.44	439.86± 34.38	437.08± 29.45
Total Query	45.49± 5.31	45.83± 5.13	44.81± 5.15	46.51± 5.28
Per Query	9.34± 0.67	9.25± 0.59	9.19± 0.49	9.39± 0.75
Total Doc.	148.84± 16.10	151.26± 13.19	146.21± 16.10	153.89± 13.18
Per Document	14.06± 1.21	14.81± 1.10	14.37± 1.21	14.50± 1.11
Per Snippet	1.73± 0.10	1.65± 0.07	1.71± 0.08	1.67± 0.09

Time-Based Measures. Table 5 reports the time taken for various interactions across each condition, system and task. We report: the mean total session time (from the first query focus to ending the task); the mean time spent entering queries; the mean per document examination time; and the mean time spent examining a result summary (or snippet). Also included are mean total values, with the mean calculated by averaging over total times for each of the 51 participants. All values are reported in seconds. Surprisingly, no significant differences were found between any of the comparisons over the total session times, the per query times, the per document times, and the per snippet times. Results, however, do show a relatively constant mean session time over each of the four experimental conditions, at ≈ 438.5 seconds, which is about 7 minutes, on average. This was in line with the time taken to find four documents in our previous user studies with similar crowdsourced workers [28] and lab-based participants [27].

Considering hypothesis ***H2b***, no evidence was found to support that under the diversity system ***D*** with an aspectual task that completion times would be lower. Here, we can see that they were in fact slightly higher (443 seconds

D vs. 430 seconds on ***ND***, i.e. the difference of about examining one more document on average).

Performance. In Table 3, we also report a number of performance measures: the number of saved documents – also broken down into the number of TREC saved and TREC non-relevant and saved, along with the number of new entities found (within saved documents, with new being in the context of a search session) – and the number of documents containing at least one new entity. In terms of the documents saved, there were no significant differences between conditions, systems or tasks. On average, participants saved around six documents on average, which was two more than the goal set, 4. This finding suggests that participants wanted to make sure that they found a few extra documents (at the expense of potentially sacrificing some accuracy), just in case some of the documents they had marked were not relevant/useful.

However, when we look at the entity-related measures, we note that participants found more documents that contained new entities, and found more entities overall when using the diversifying system, ***D***. This was significantly different (6.52 ± 0.61 compared to 4.34 ± 0.42 respectively, where $F(1, 203) = 8.70, p < 0.05$). When examining each condition, the Bonferroni follow-up test highlighted significant differences between condition ***D.As*** and conditions ***D.Ad*** and ***ND.Ad***, where $F(3, 203) = 3.49, p < 0.05$. We also notice that participants found more documents with entities and more entities overall when undertaking the ad-hoc retrieval task using the diversifying system, in comparison to when they used the non-diversifying system, ***ND*** (documents with entities: 2.63 vs. 2.02, new entities: 5.82 vs. 4.37). Although this was again not significantly different, it does suggest that when participants utilised the diversifying system, they did learn more about the different aspects of the topic (or at least a greater number of aspects associated with the topics) than when using the non-diversifying system.

Post Task and Post System Questionnaires. There were no significant differences between conditions, tasks, or system for any of the post-task questions. For the post system questionnaires, participants were roughly evenly split between their preference for the diversified or non-diversified system – again with no significant differences. This finding suggests that despite the substantial (and significant) difference in aspectual recall and other system performance measures shown between the systems, participants of this study seemed largely ambivalent to the different influence of the two systems. Their observed behaviours do however suggest that the system (and task) did ultimately affect their performance.

4.2 Gain over Time

We motivated this study using IFT where we constructed a number of gain curves that reflected our beliefs about how the search performance experienced by searchers would look like over each system and task. This was done

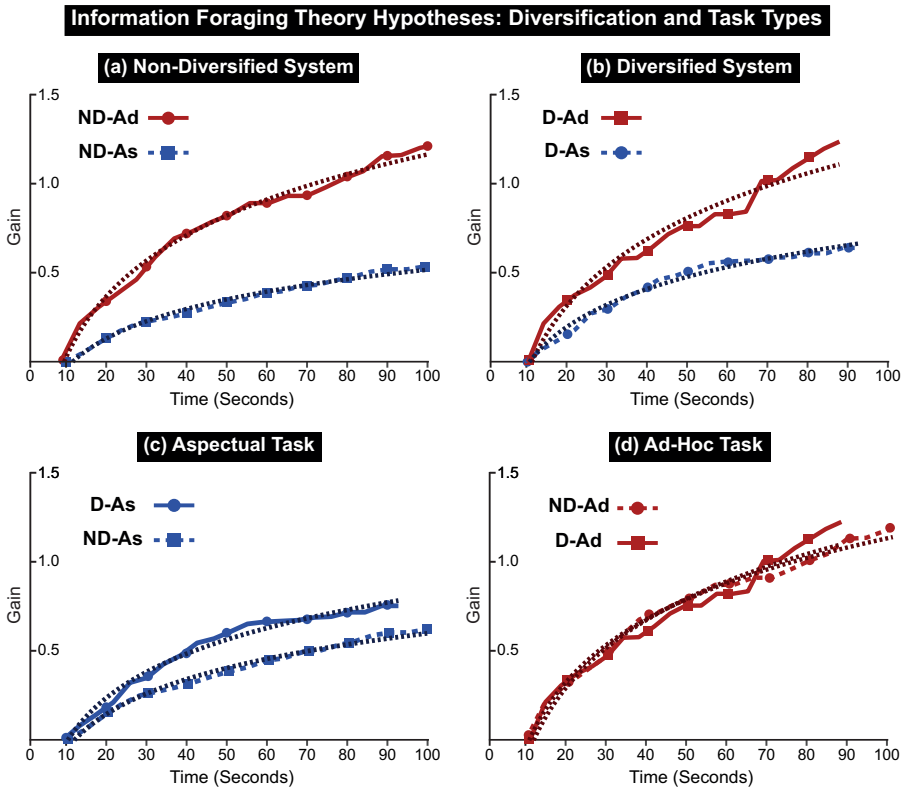

Fig. 3 Plots illustrating the *Cumulative Gain (CG)* attained by participants of the study, on average, over the first 100 seconds of a search session. Plots are analogous to the four presented in Fig. 1. Also included in are fitted curves (dashed lines).

to generate the hypotheses mentioned in Section 2.2. Here, we examine how participants performed over time for each of the systems and conditions to infer the gain curves. We then compare that to our expectations (which were shown previously in Fig. 1).

To create empirical gain curves, we plotted *Cumulative Gain (CG)* over time. Here, we defined gain to be the number of saved relevant documents (in the case of ad-hoc retrieval), and gain as the number of saved relevant but different documents (in the case of aspectual retrieval). These definitions are what we said would constitute a useful document in the two separate tasks. As the gain is measured in the same units, we can plot the gain for both tasks on the same axes to aid comparability.

Fig. 3 shows the corresponding empirical gain curves for: ***(a)*** the non-diversified system on both tasks, ***(b)*** the diversified System on both tasks, ***(c)*** the aspectual task for both systems, and ***(d)*** the ad-hoc task for both systems. Compared to our expectations in Fig. 1, we see on visual inspection that our predictions were roughly in line with the gain experienced. For

example, in Fig. 1 ***(a)***, we hypothesised that on the non-diversified system, participants would experience greater levels of gain; the empirical gain curves reported in Fig. 3 ***(a)*** show this. A critical difference though is for Fig. 1 ***(b)***, where we hypothesised that the gain curves would be similar on the diversified system (up until a point) before the aspectual gain would drop. Examination of Fig. 3 ***(b)*** shows that participants had a very different experience – and experienced lower gains earlier in their search sessions – motivating a revision of our expectations.

To do so, we first fit a logarithmic function to each of the gain curves given time (as done by Athukorala et al. [2]), such that:

$$gain = b \cdot log(time) - a. \tag{1}$$

Table 6 shows the parameters and correlation coefficients for fit (r^2) for each condition. We then could calculate how many documents a participant would examine by drawing the tangent line to the estimated gain functions from the origin. This resulted in the predicted number of documents examined – which we see are in line with the actual documents examined. With respect to Fig. 3 ***(b)***, we see that for the diversified system, the theory, given their performance, suggests that participants should examine more documents per query on the aspectual task than when undertaking the ad-hoc task (i.e. 4.98 to 3.36, respectively). We observed that they examined 3.48 and 3.02 documents per query – which follows the same trend – but not to the same magnitude. Thus, we revised our expectations regarding how people would search differently between these tasks. With respect to ***H1***, we see that the theory, given their performance, suggests that participants, when undertaking the aspectual task, would examine fewer documents per query when using the diversified system. This is compared to the non-diversified system (4.36 vs 4.92). Again, we see that they examined 3.02 and 3.65 documents per query respectively, again following the same trend – but not to the same magnitude. This post-hoc analysis has justified some of our initial hypotheses regarding how search behaviour would change under the different conditions – but it has also led to us revising our expectations based on the observed, empirical data.

5 Summary and Conclusions

In this paper, we investigated the effects of diversifying search results when searchers undertook complex search tasks, where one was required to learn about different aspects of a topic. We inferred a number of hypotheses based upon IFT in which diversification would lead to searchers examining fewer documents per query and subsequently issuing more queries. We tested our hypotheses by conducting a within-subjects user study, using *(i)* a non-diversified system; versus *(ii)* a diversified system, when the retrieval task was either: *(a)* ad-hoc; or *(b)* aspectual in nature.

Our findings lend evidence to support the IFT hypotheses broadly. However, we only observed statistically significant differences across a subset of

Table 6 Fitting parameters for the gain curves illustrated in Fig. 3 over each experimental condition. Also included are the estimations from the model for the time to examine a document, and the depth to which participants should go *(Pred. Docs.)* – as well as the observed number of documents examined *(Actual Docs.)*, and stopping depth (on average).

	Model Fitting			Pred.	Actual
Experimental Condition	a	b	r^2	**Docs.**	**Docs.**
ND.Ad	−1.08	0.48	0.989	3.68	3.23
ND.As	−0.57	0.23	0.987	4.92	3.65
D.Ad	−1.22	0.52	0.959	4.98	3.48
D.As	−0.68	0.29	0.985	4.36	3.02

behavioural and temporal measures. This was despite the fact that there were significant differences in the performance of the two systems – the diversified system was able to, on average, return a ranked list of results with a greater number of documents containing new, unseen entities. This finding is in line with past work which found that interface-based interventions seemingly had little influence on search performance and search behaviours. Clearly, bigger differences need to be present – or larger sample sizes are required – to determine if the difference between systems over all examined indicators is significant. Despite these results, there were a number of clear trends.

When performing the aspectual task on the diversified system ***D*** (in contrast to the Non-Diversified System): participants examined fewer documents per query (3 vs. 3.7 documents/query), issued slightly fewer queries (5.9 vs. 5.2 queries), and didn't go to as great a depth when examining SERPs (depths of 12.8 vs 15.7). Taken together this resulted in a lower probability of clicking ($P(C) = 0.16$ vs 0.21, which was significantly different) and interestingly a lower probability of clicking on non-relevant ($P(C|N) = 0.13$ vs. 0.18, which was also significantly different). While participants spent a similar amount of time searching on both systems, participants on the diversified system spent slightly more time examining each document (16 seconds vs. 13 seconds), and more time in total examining documents (163 seconds vs. 145 seconds) - suggesting that more effort was directed to assessing rather than searching. However, participants found significantly more entities (7.2 vs. 4.3 entities) and found more documents that contained new/different entities (3.2 vs 2.4). Both of these findings were statistically significant. This shows that the diversification algorithm led to a greater awareness of the topics and provided participants with greater coverage of the topic - which suggests that participants were able to learn more about the topic, and were exposed to less bias.

When performing the ad-hoc task over the diversified system ***D*** (in contrast to the non-diversified system ***ND***): participants examined more documents per query (3.48 vs. 3.23 documents/query), issued slightly more queries (4.96 vs. 5.20 queries), and examined content to greater depths presented on SERPs (depths of 16.2 vs. 13.9). Again, this meant that the probability of clicking

was lower on the diversified system (0.16 vs. 0.20); this was significantly so. Participants spent similar amounts of time searching on both systems. However, unlike on the aspectual tasks, participants spent less time examining potentially relevant documents on system ***ND*** (13.7 vs. 15.1 seconds), and they spent less time in total assessing documents (139.6 vs, 152.8 seconds). This suggests that less effort was directed at assessing, rather than searching. This could be possibly due to the performance of the diversified system being higher than the non-diversified system ($P@5 = 0.29$ vs. 0.25, which was significantly different). Alternatively, it could be because the results returned were easier to identify as relevant as the probability of marking a document given it was relevant was higher (0.74 vs. 0.67). This suggests that participants may be more confident when using the diversified system. Although not explicitly requested in the task description, participants encountered more novel entities when using the diversified system (5.8 vs. 4.4). Participants also found more documents with new entities using the diversified system (2.6 vs. 2.0). Taken together, this suggests that participants again implicitly learn more about the topic because the diversified system surfaced content that presented a more varied view on the topic.

With regards to the application of IFT, we showed that generated hypotheses were largely sound, but the empirical data prompted us to revise the hypotheses. Initially, we hypothesised that the performance and behaviour on both tasks would be similar when using the diversified system (see Fig. 1(b)). However, post-hoc analysis revealed that the performance (and subsequent behaviour) was different (see Fig. 3(b)). Here, participants obtained higher levels gain for the ad-hoc task. Thus, under such conditions, IFT would stipulate that they would examine more documents per query (3.48 vs 3.02 documents/query) and issued fewer queries (4.9 vs. 5.9 queries) when undertaking the ad-hoc retrieval task vs. the aspectual retrieval task (as opposed to there being no difference). Encouragingly, our application of IFT (before and after the experiment) led to new insights into how behaviours are affected under different conditions. This shows that IFT is a useful tool in developing, motivating and analysing search performance and behaviours. Furthermore, counter to our intuition about how we *believed* people would behave in these conditions, the theory provided *more informed and accurate hypotheses* which tended to hold in practice.

This work motivates further research into complex search tasks and the impact of diversifying search results. Diversification can play an important role in improving the search experience by providing greater coverage of a topic, and mitigating potential biases that may exist in search results. One such avenue for further exploration is a per-topic analysis. As our results were presented as a mean over each experimental condition (or system, or task, as reported in the bottom table of Table 3), we may have missed important per-topic differences. This argument can be motivated from the large variance in the number of different entities identified for each topic (with *Airport Security* having only 14 different airports, and *Wildlife Extinction* possessing 168 different species of endangered animal). However, a reported low variance

in the number of entities found (see the *#Entities Found* rows in Table 3) suggests that this may not be the case. Further work includes an investigation into how such search behaviours and performance would vary with larger sample sizes, yielding increased experimental power. An examination of how these behaviours and performance under different retrieval tasks and search contexts would also be an interesting area for future exploration. For example, *would different retrieval tasks affect the perceptions of individuals and their decisions?* Finally, an examination of different diversification algorithms would provide us with a better understanding of how diversification influences search behaviours and performance.

In conclusion, we found that in terms of search behaviour: participants on the diversified system issued more queries and examined fewer documents per query when performing aspectual search tasks. Furthermore, we showed that when using the diversified system, participants were more successful in marking relevant documents, and obtained a greater awareness of the topics (i.e. identified relevant documents containing novel aspects). This was also the case even when they were not specifically instructed to do so (i.e. when performing the ad-hoc search task). These findings suggest that diversification should be employed more widely, particularly where bias is a potential issue (such as in news search). Here, diversification algorithms would the be able to present a broader overview of the aspects within a topic.

Acknowledgements The authors would like to thank the 51 Amazon Mechanical Turk workers for their participation in the study. We would also like to thank the anonymous reviewers for their insightful feedback related to the initial submission of this article. The lead author acknowledges support from the *EPSRC*, under grant number `1367507`.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proc. 2nd ACM WSDM, pp. 5–14 (2009)
2. Athukorala, K., Oulasvirta, A., Glowacka, D., Vreeken, J., Jacucci, G.: Narrow or broad?: Estimating subjective specificity in exploratory search. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14, pp. 819–828 (2014)
3. Azzopardi, L., Kelly, D., Brennan, K.: How query cost affects search behavior. In: Proceedings of 36th ACM SIGIR, pp. 23–32 (2013)
4. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proc. 21st ACM SIGIR, pp. 335–336 (1998)
5. Carterette, B., Chandar, P.: Probabilistic models of ranking novel documents for faceted topic retrieval. In: Proc. 18th ACM CIKM, pp. 1287–1296 (2009)
6. Chen, H., Karger, D.R.: Less is more: probabilistic models for retrieving fewer relevant documents. In: Proc. 29th ACM SIGIR, pp. 429–436 (2006)
7. Chen, M.C., Anderson, J.R., Sohn, M.H.: What can a mouse cursor tell us more?: Correlation of eye/mouse movements on web browsing. In: Proc. 19th ACM CHI Extended Abstracts, pp. 281–282 (2001)
8. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proc. 31st ACM SIGIR, pp. 659–666 (2008)

9. Collins-Thompson, K., Hansen, P., Hauff, C.: Search as learning (dagstuhl seminar 17092). In: Dagstuhl Reports, vol. 7 (2017)
10. Diriye, A., White, R., Buscher, G., Dumais, S.: Leaving so soon?: Understanding and predicting web search abandonment rationales. In: Proc. 21st ACM CIKM, pp. 1025–1034 (2012)
11. Dostert, M., Kelly, D.: Users' stopping behaviors and estimates of recall. In: Proc. 32nd ACM SIGIR, pp. 820–821 (2009)
12. Feild, H., Jones, R., Miller, R., Nayak, R., Churchill, E., Velipasaoglu, E.: Logging the search self-efficacy of amazon mechanical turkers. In: Proc. CSE SIGIR Workshop, pp. 27–30 (2010)
13. Harman, D.: Overview of the first text retrieval conference (trec-1)
14. Harper, D.J., Kelly, D.: Contextual relevance feedback. In: Proc 1st ACM IIiX, pp. 129–137 (2006)
15. Hassan, A., Shi, X., Craswell, N., Ramsey, B.: Beyond clicks: Query reformulation as a predictor of search satisfaction. In: Proc. 22nd CIKM, pp. 2019–2028 (2013)
16. He, J., Meij, E., de Rijke, M.: Result diversification based on query-specific cluster ranking. J. Am. Soc. Inf. Sci. Technol. **62**(3), 550–571 (2011)
17. Hearst, M.A.: Tilebars: Visualization of term distribution information in full text information access. In: Proc. 13th ACM SIGCHI, pp. 59–66 (1995)
18. Hearst, M.A.: Texttiling: Segmenting text into multi-paragraph subtopic passages. Comput. Linguist. **23**(1), 33–64 (1997)
19. Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context (2005)
20. Iwata, M., Sakai, T., Yamamoto, T., Chen, Y., Liu, Y., Wen, J.R., Nishio, S.: Aspectiles: Tile-based visualization of diversified web search results. In: Proc. 35th ACM SIGIR, pp. 85–94 (2012)
21. Järvelin, K., Kekäläinen, J.: Cumulative gain-based evaluation of IR techniques. TOIS **20**(4), 422–446 (2002)
22. Kato, M.P., Sakai, T., Tanaka, K.: Structured query suggestion for specialization and parallel movement: Effect on search behaviors. In: Proc. 21st WWW, pp. 389–398 (2012)
23. Kelly, D., Arguello, J., Edwards, A., Wu, W.c.: Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In: Proc. 1st ACM ICTIR, pp. 101–110 (2015)
24. Kelly, D., Gyllstrom, K.: An examination of two delivery modes for interactive search system experiments: Remote and laboratory. In: Proc. 29th ACM SIGCHI, pp. 1531–1540 (2011)
25. Kelly, D., Gyllstrom, K., Bailey, E.: A comparison of query and term suggestion features for interactive searching. In: Proc. 32nd ACM SIGIR, pp. 371–378 (2009)
26. Kiseleva, J., Kamps, J., Nikulin, V., Makarov, N.: Behavioral dynamics from the serp's perspective: What are failed serps and how to fix them? In: Proc. 24th ACM CIKM, pp. 1561–1570 (2015)
27. Maxwell, D., Azzopardi, L.: Stuck in traffic: How temporal delays affect search behaviour. In: Proc. 5th IIiX, pp. 155–164 (2014)
28. Maxwell, D., Azzopardi, L., Moshfeghi, Y.: A study of snippet length and informativeness: Behaviour, performance and ux. In: Proc. 40th ACM SIGIR, pp. 135–144 (2017)
29. McDonald, J., Ogden, W., Foltz, P.: Interactive information retrieval using term relationship networks. NIST Special Publication pp. 379–384 (1998)
30. Over, P.: Trec-6 interactive track report pp. 73–82 (1998)
31. Over, P.: The trec interactive track: an annotated bibliography. Inf. Proc. & Mgt. **37**(3), 369–381 (2001)
32. Pirolli, P., Card, S.: Information foraging. Psychological Review **106**, 643–675 (1999)
33. Prabha, C., Connaway, L., Olszewski, L., Jenkins, L.: What is enough? Satisficing information needs. J. of Documentation **63**(1), 74–89 (2007)
34. Radlinski, F., Dumais, S.: Improving personalized web search using result diversification. In: Proc. 29th ACM SIGIR, pp. 691–692 (2006)
35. Rose, D.E., Levinson, D.: Understanding user goals in web search. In: Proc. 13th WWW, pp. 13–19 (2004)
36. Santos, R.L., Macdonald, C., Ounis, I.: Exploiting query reformulations for web search result diversification. In: Proc. 19th WWW, pp. 881–890 (2010)

37. Santos, R.L., Macdonald, C., Ounis, I.: Intent-aware search result diversification. In: Proc. 34th ACM SIGIR, pp. 595–604. ACM (2011)
38. Smucker, M., Guo, X., Toulis, A.: Mouse movement during relevance judging: Implications for determining user attention. In: Proc. 37th ACM SIGIR, pp. 979–982 (2014)
39. Syed, R., Collins-Thompson, K.: Retrieval algorithms optimized for human learning. In: Proc. 40th ACM SIGIR, pp. 555–564 (2017)
40. Umemoto, K., Yamamoto, T., Tanaka, K.: Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search. In: Proc. 39th ACM SIGIR, pp. 405–414 (2016)
41. Villa, R., Cantador, I., Joho, H., Jose, J.M.: An aspectual interface for supporting complex search tasks. In: Proc. 32nd ACM SIGIR, pp. 379–386 (2009)
42. Voorhees, E.: Overview of the trec 2005 robust retrieval track. In: Proc. TREC-14 (2006)
43. Voorhees, E., Harman, D.: TREC: Experiment and Evaluation in Information Retrieval. The MIT press (2005)
44. Zach, L.: When is "enough" enough? modeling the information-seeking and stopping behavior of senior arts administrators: Research articles. J. American Soc. for Info. Sci. and Tech. **56**(1), 23–35 (2005)
45. Zhai, C., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: ACM SIGIR Forum, vol. 49, pp. 2–9. ACM (2015)
46. Zuccon, G., Azzopardi, L.A., van Rijsbergen, K.: The quantum probability ranking principle for information retrieval. LNCS: Advances in IR Theory **5766**, 232–240 (2009)
47. Zuccon, G., Leelanupab, T., Whiting, S., Yilmaz, E., Jose, J., Azzopardi, L.: Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. Information Retrieval **16**(2), 267–305 (2013)