Composite Retrieval of Heterogeneous Web Search

Horatiu Bota
University of Glasgow
h.bota.1@research.gla.ac.uk

Ke Zhou
University of Edinburgh
ke.zhou@ed.ac.uk

Joemon M. Jose
University of Glasgow
joemon.jose@glasgow.ac.uk

Mounia Lalmas
Yahoo Labs
mounia@acm.org

ABSTRACT

Traditional search systems generally present a ranked list of documents as answers to user queries. In aggregated search systems, results from different and increasingly diverse verticals (image, video, news, etc.) are returned to users. For instance, many such search engines return to users both images and web documents as answers to the query "flower". Aggregated search has become a very popular paradigm. In this paper, we go one step further and study a different search paradigm: *composite retrieval*. Rather than returning and merging results from different verticals, as is the case with aggregated search, we propose to return to users a set of "bundles", where a bundle is composed of "cohesive" results from several verticals. For example, for the query "London Olympic", one bundle per sport could be returned, each containing results extracted from news, videos, images, or Wikipedia. Composite retrieval can promote exploratory search in a way that helps users understand the diversity of results available for a specific query and decide what to explore in more detail. In this paper, we propose and evaluate a variety of approaches to construct bundles that are *relevant*, *cohesive* and *diverse*. Compared with three baselines (traditional "general web only" ranking, federated search ranking and aggregated search), our evaluation results demonstrate significant performance improvement for a highly heterogeneous web collection.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

Keywords

Composite retrieval; bundle; vertical; diversity; relevance; cohesion

1. INTRODUCTION

Consider the following user information need "finding all information to plan a trip to Korea". Answering this information need typically involves submitting several queries to gather information about airports and visa policies, to read online reviews about hotels, and to check the geographic proximity of places to visit. Current search engines aggregate results from multiple verticals. However, the presentation of search results is limited to blocks where each block contains homogeneous information of one type (vertical). It is our conjecture that, though useful, this is not an effective solution. As the web has made available a large variety of diverse search engines – so-called *verticals* – e.g., news, image, video, blog, it is becoming important to return to users organised answers, made of information extracted from heterogeneous data sources. Doing so will not only support users in complex search tasks, but also allow them to understand the diversity of the information space and select what matters to them most.

For example, users typing the query "olympic" during the London 2012 Olympic Games may have been interested in different on-going game results with detailed statistics, video summaries and players' post-match commentary quotes. Returning such results per sport would have been of immense help to users, allowing them to explore all that is happening within each sport and focus on specific sports, events or athletes. We propose to return users such results into what we refer to as a *composite page*, where results from diverse verticals are *grouped* to form "bundles". Thus a composite page is a set of bundles, where a bundle is composed of "cohesive" information extracted from various sources (e.g. videos, statistics and quotes). What cohesive means and how it is operationalised is what we address in this paper.

Previous work [2] modelled the task of looking for restaurants in a city as *composite retrieval* and proposed to organise results into item *bundles* that together constitute an improved exploratory experience over the typical ranked list of results. In their work, composite retrieval was explored in the context of a homogeneous information space: selecting restaurants in a newly visited city. However, complex web search involves searching through diverse verticals and integrating their results: a challenging problem when it comes to forming bundles.

To construct a high-quality exploratory composite page, several criteria should be satisfied: **bundle relevance**, where the items within the bundles should be topically relevant to the query; **bundle cohesion**, where the items within the bundles should be similar with each other and therefore coherent to a topic; **bundle diversity**, where the items within the bundle should cover a diverse set of results from various verticals; and **page diversity**, where the bundles within the page should cover various aspects/topics of the query. As in traditional search, *relevance* is generally a priority as this is the key factor to a satisfied user experience; promoting cohesion and diversity is important but only when the results are relevant.

In this paper, we study *composite retrieval* in a *heterogeneous* web search environment. We propose several approaches to construct composite pages. Two challenges arise due to the heteroge-

WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2744-2/14/04.
http://dx.doi.org/10.1145/2566486.2567985.

neous nature of the data. First, relevance score distributions are not comparable across verticals, making the estimation of relevance of bundles more complex. Second, different factors can sometimes be contradictory with each other (relevance vs. diversity), thus requiring to determine the right trade-off between them.

We build on existing composite retrieval frameworks by adapting them for the heterogeneous web search context. In addition, we put forward a new approach for composite retrieval, which we refer to as the *central-satellite* composite retrieval paradigm, where a retrieved object is defined as a central package (a set of "general web" documents) and a set of satellite packages consisting of items retrieved from verticals that are coherent to the central package. This paper aims to answer the following research questions:

- **RQ1**: Can we construct composite pages that are more relevant than existing solutions, such as "general web search only" ranking, aggregated search and federated (merged) search ranking?
- **RQ2**: Which features are essential (or optional) to form bundles in composite retrieval? When constructing bundles, can we utilise query-related entities as an anchor to bridge the semantic gap between items retrieved from heterogeneous sources?
- **RQ3**: As bundles are created by selecting items from traditionally ranked lists of documents, how robust are different composite retrieval approaches to the quality of the initial ranking of documents?

The contributions of the paper are the following: we proposed a novel approach and demonstrated its effectiveness in constructing bundles to form composite pages. We conducted extensive experiments comparing our proposal with a number of baselines and approaches. To our knowledge, this is the first endeavour to use entities to bridge the semantic gap between items retrieved from heterogeneous web resources. Moreover, we investigated and provided insights on the usefulness of different sources of evidence for building bundles.

The paper is organised as follows. Section 2 discusses previous work. We describe our composite retrieval framework in Section 3 and propose a variety of approaches to form bundles in Section 4. Details of the test collections and experimental setup are provided in Section 5. Section 6 reports our experimental results. We discuss the implications of our results and conclude in Section 7.

2. RELATED WORK

Composite Retrieval Composite retrieval has been studied in recent years, however, the applications so far are in semi-structured scenarios (or in scenarios with implicit structure) such as shopping or finding a restaurant. In addition, they are experimented mostly on homogeneous environments and their applicability for general heterogeneous web search is not obvious. Within the shopping domain, [5] explored retrieving bundles of items that are subject to budget and compatibility (e.g. an iPhone and all its accessories). CARD [7] is a framework for finding the top-k recommendations of composite packages of products or services. [22] used a Fagin-style algorithm for variable size bundles to provide approximate solutions to composite retrieval and proved its optimality. [11] designed a graph traversal algorithm that retrieves itineraries in a city where different constraints are imposed (time, city chain).

Recently, [2] studied the complexity of the problem of composing bundles with constraints (like budgets), and proposed different algorithmic formulations to solve the problem. Their approach introduces a framework based on a produce and choose approach. In the production stage bundles are formed to promote cohesion and diversity. The application is in a homogeneous space and is amenable to simple clustering solutions. This framework, though general, suffers from a number of deficiencies. First, it is not trivial to employ this framework in heterogeneous scenarios. Second, the bundle formulation does not consider the concept of relevance in forming the bundles. This eventually leads to unsatisfactory solutions when forming bundles across heterogeneous scenarios in the context of web search.

We study composite retrieval in the context of heterogeneous web search and provide solutions to tackle the challenges arising from the heterogeneous nature of the data. In addition, as relevance is highly correlated to a satisfactory user experience in search [16], different from other works, we treat this as our main criteria of optimisation, whereas criteria such as cohesion and diversity are considered secondary. We explore the framework developed in [2] and adapt it to suit the needs of heterogeneous scenarios. Specifically, we enhance the definition of bundles by incorporating the concept of relevance explicitly. Secondly, we incorporate diversity and relevance at the time of ranking (choosing) the bundles, thus leading to effective solutions. Thirdly, we exploit entities to link between relevant items across noisy verticals and finally we incorporate query intent into the formation of bundles. Extensive experiments on the TREC federated web track data set demonstrate the power and effectiveness of heterogeneous composite retrieval.

Aggregated and Federated Search Aggregated search is the task of searching and assembling information from a variety of resources (or verticals) and placing it into a single interface [4, 24]. Aggregated search can be compared to federated search [18] (also known as distributed information retrieval), which deals with merging result rankings from different search engines into one single ranking list. The major difference between federated search and aggregated search is the heterogeneity of the data and the presentation of the results. The main challenges involved with aggregated search and federated search are resource selection and result composition. The former deals with deciding which resources contain the most relevant contents to a given query and the latter deals with selecting a subset of items from selected resources and presenting them as results. In aggregated search, the most common result presentation strategy is to present the results into a ranked list of so-called blocks where each block contains homogeneous information of one type (vertical). In federated search, the results are merged into one unified ranked list of items, disregarding the item type. Similar to aggregated search, selecting and organising results from heterogeneous data is the main focus of composite retrieval. However, rather than presenting the results of each selected vertical as a homogeneous block, composite retrieval aims to present results into cohesive bundles, where each bundle is composed of heterogeneous items (from several verticals). This potentially enhances the user experience since cohesive bundles are formed and presented to the user for exploration.

Cluster-based Retrieval The cluster hypothesis (closely associated documents tend to be relevant to the same requests) [6] gave rise to a large body of work [14, 21] on using query-specific document clusters for improving retrieval effectiveness. Clusters are generally created from documents that are highly ranked by an initial search performed in response to the query. It has been shown [14, 21] that for many queries there are query-specific clusters that contain a very high percentage of relevant documents. Furthermore, positioning the documents of these clusters at the top of the result list [14, 21] yields highly effective retrieval performance (in terms of relevance) than ranking documents directly. Our work is similar to cluster-based retrieval as we form bundles based on

a cluster-inspired optimisation approach (selecting items that are similar to each other to form a bundle). Similar to cluster-based retrieval, we rank the verticals (clusters) based on their estimated relevance and ultimately select the top ranked verticals to choose items from. The heterogeneous nature of the data and our approach to constructing semantic links between documents are what differentiate our work from traditional cluster-based retrieval.

Diversity in information retrieval (IR) Recently, IR research has investigated "diversity-based", or, "subtopic" retrieval approaches for modelling user search intents during search tasks [10] for ambiguous or multi-faceted information needs. An intent-diversified result ranking can be created by interleaving documents sampled from possible search intents (subtopics), with the importance of each intent indicated by several features such as prior search intent click-through rate or original document relevance. Our composite retrieval approach also takes "subtopic" diversity into account when forming composite pages with diverse bundles. However, rather than forming a homogeneous ranking list covering various subtopics, we construct heterogeneous composite pages that consist of "bundles" where each "bundle" corresponds to a "subtopic".

3. COMPOSITE RETRIEVAL FRAMEWORK

We formally define the heterogeneous composite retrieval problem below, followed by a review of the associated challenges.

3.1 Problem Formulation

We propose a new framework for composite retrieval in a heterogeneous web environment, which is motivated by previous work [2]. The latter is mostly applicable to a semi-structured environment, such as finding restaurants, and focuses solely on page *cohesion* and *diversity*, without considering *relevance*. Hence, the applicability of this framework for heterogeneous web search is not straightforward. In addition, we believe that *relevance* is the most crucial component of a positive user experience in search and therefore should be incorporated into the optimisation of the objective function. The heterogeneous nature of the multi-vertical environment also requires novel ways to model and estimate the various components of the proposed framework.

The goal of composite retrieval in a heterogeneous environment is to retrieve a set of bundles $P = \{S_1, ..., S_k\}$ to form a composite page P, where a bundle $S_i \in 2^I$ is a set of items that originate from a subset of verticals $V = \{V_1, ..., V_n\}$. The *objective* of the optimisation is to find a set of bundles to form a composite page $P = \{S_1, ..., S_k\}$ that maximises the utility $util(P)$. We assume that the utility of the page $util(P)$ solely depends on the following four criteria:

- **Relevance**: the expected probability of the items in the bundle that are relevant to the query:

$$rel(P) = \frac{\sum_{1 \leqslant i \leqslant k} \sum_{u \in S_i} r(u|q)}{\sum_{1 \leqslant i \leqslant k} \sum_{u \in S_i}} \quad (1)$$

  where $r(u|q)$ is the probability that a user finds an item u relevant as a function of the editorial grade g_u of that item u. $r(u|q)$ can be chosen in different ways. In accordance with the common gain function for DCG, we define it as:

$$r(u|q) = \frac{2^{g_u} - 1}{2^{g_{max}}}, \quad g \in \{0, ...g_{max}\} \quad (2)$$

  when the item is non-relevant ($g = 0$), the probability that the user finds it relevant is 0, whereas when the item is highly relevant ($g = 4$, if a 5 point scale is used), then the probability of relevance is near 1. When a binary relevance grade is used, $rel(P)$ corresponds to the precision at a cut-off metric $P@n$ of the page P.

- **Topical Cohesion**: the expected accumulated similarity of the items within the bundle:

$$tcoh(P) = \frac{\sum_{1 \leqslant i \leqslant k} \sum_{u,v \in S_i} s(u,v)}{\sum_{1 \leqslant i \leqslant k} \sum_{u,v \in S_i}} \quad (3)$$

  The similarity $s(u, v)$ between an item pair (u, v) can be computed implicitly by a given representation of the items. This $tcoh(P)$ corresponds to the normalised expected cohesion of bundles (clusters).

- **Topical Diversity**: the expected inter-bundle separation for bundle pairs:

$$tdiv(P) = \frac{\sum_{1 \leqslant i < j \leqslant k} (1 - \max_{u \in S_i, v \in S_j} s(u,v))}{\sum_{1 \leqslant i < j \leqslant k}} \quad (4)$$

  where the inter-bundle separation is defined as the minimum distance between two items from separate bundles.

- **Vertical Diversity**: the expected number of verticals the relevant items belong to in the bundle:

$$vdiv(P) = I\text{-}rec(P) \quad (5)$$

  where $I\text{-}rec(P)$ is the intent (vertical) recall metric [23] for the page P. Basically, it calculates the recall of those verticals that return relevant items on the composite page.

A page P with high $util(P)$ should consist of a set of *topically diverse* bundles that each contains *relevant* items that originated from a set of *diverse verticals* and are *cohesive* to one aspect of the query. Since this is a novel search task, the importance of each factor to the user experience has not been well understood and how to combine those factors into one utility function is outside the scope of this paper. Therefore, we evaluate the performance of each factor separately. However, as **relevance** is key to the user experience in search [16] and is a prior to construct any composite pages with high utility, we use it as our main criteria of interest for the purpose of evaluation.

If the query subtopics and each corresponding relevance assessments of items are available, as in the collections used to evaluate diversity based IR [10], *topical relevance* and *topical cohesion* can be evaluated based on existing diversity metrics (e.g., intent-aware metrics [1]). However, we rely on a collection which has multiple verticals (federated search collection) for which subtopic relevance assessments are not available. Considering the high cost involved in collecting subtopic relevance assessments, we evaluate those two criteria (*topical relevance* and *topical cohesion*) using evaluation metrics of clustering quality (cluster cohesion and separation) as described above.

3.2 Challenges

Composite retrieval is challenging in particular in the context of a heterogeneous web environment for several reasons. The challenges are (1) computational **complexity**, (2) term **mismatch** with heterogeneous information and (3) appropriate estimation and optimisation of the **multiple criteria** function (in our case, cohesion, diversity and relevance).

For effective user experience, we aim to create optimal bundles, ones that meet all our criteria. However, this in itself is an *NP-hard* optimisation problem [2]. More precisely, it was proved that optimising for *cohesion* and *diversity* in the objective function reduced

to the well known NP-complete problem *Maximum Edge Subgraph* in composite retrieval. Therefore, we require efficient greedy algorithms to optimise the utility of a composite page.

The term distributions in the different verticals vary widely [17] and are therefore *not comparable*. This makes it difficult to calculate the similarity $s(u, v)$ between two items u and v that originated from two different verticals. Therefore, it is important to bridge the different term distributions of diverse verticals in a way that allow to compare them and combine them in a meaningful manner.

Different sources of evidence (for instance, the query-item similarity, the vertical where the item originated from, etc.) have to be considered when estimating the relevance of the item and also to decide whether to include it in a bundle and in which bundle. A comprehensive study is required to understand the optimal way to estimate the relevance of an item in heterogeneous composite retrieval. In addition, how to appropriately combine these sources of evidence to account for cohesion and diversity is not obvious.

Our contributions lie in addressing these three challenges: (1) We propose a new approach for bundle formation and experiment with a number of variations of produce and choose approach [2]; (2) we investigate an entity based representation to bridge the heterogeneous nature of the information contained in various verticals; and (3) we propose different approaches to incorporate relevance into the optimisation and analyse the usefulness of various features in estimating relevance.

4. BUNDLE SELECTION & RANKING

We introduce our adapted greedy approach for optimisation and describe our methodology for estimating the different sources of evidence in this section.

As we discussed above, this problem is NP-hard. Previous work [2, 3] showed that *Maximum Edge Subgraph* and *composite retrieval* are two counterparts. If we generate candidate bundles and we consider each candidate bundle as a node of a bundle-graph, where inter-bundle distances are the edge weights, then the composite retrieval problem can be reduced to the *Maximum Edge Subgraph* problem. This suggests that composite retrieval can be solved by generating a set of candidate bundles and then selecting the best possible subset. They call this approach Produce-and-Choose (PAC) and we choose this as a fundamental paradigm for solving our problem. We chose the PAC method because it fits our problem better and it has been proved to produce better (more cohesive) bundles than other approaches (Cluster-And-Pick) [3].

The PAC approach consists of two stages: (1) Produce bundles and (2) Choose bundles. There are ways to generate bundles: Bundles One-by-One (BOBO) and Constrained Clustering. We employ just one of these approaches, BOBO (Section 4.1.1), as it is representative of a wide class of clustering algorithms and is a better fit to the structure of our problem. However, the application of this framework in a heterogeneous environment is not trivial. We propose to address this problem by using an entity based representation. This approach also allowed us to diversify the results (Section 4.2.1) based on captured query intents that are represented by the entities. In summary, unlike [2], at the production stage we compute cohesion and vertical diversity for producing good bundles and at the choose stage (ranking), we integrate topical diversity and relevance. In addition, based on the unique characteristics of the multi-vertical environment, we proposed a novel approach (Central-Plus-Satellite, CPS) that better suits a heterogeneous environment, such as the one we are exploring in this paper. For choosing bundles (Section 4.2), previous research [2, 3] showed that the known results for *Maximum Edge Subgraph* can be exploited to preserve the approximation guarantees. We employ their approach and incorporate relevance into it.

Algorithm 1: Produce Bundles: BOBO

Input: set of items I, a cost function f that checks vertical diversity and bundle size constraints, a threshold β on the number of items in a bundle, minimum bundle score μ, number of bundles c
Output: a set of c valid bundles
$Cand \leftarrow \varnothing$
$Pivots \leftarrow I$
while $Pivots \neq \varnothing$ *and* $|Cand| < c$ **do**
  $\omega \leftarrow Pivots[0]$
  $I \leftarrow I \setminus \omega$
  $S \leftarrow pick\_bundle(\omega, I, f, \beta)$
  if $score(S) \geqslant \mu$ **then**
    $I \leftarrow I \setminus S$
    $Pivots \leftarrow Pivots \setminus S$
    $Cand \leftarrow Cand \cup S$
  else
    $Pivots \leftarrow Pivots \setminus \omega$
return $Cand$

Algorithm 2: pick_bundle

Input: pivot ω, set of items I, a cost function f that checks vertical diversity and bundle size constraints, a threshold β on the number of items in a bundle
Output: bundle s
s = $\{\omega\}$; active $\leftarrow I \setminus \omega$; $finish$ = false
while *not* $finish$ **do**
  $i \leftarrow argmax_{\{i \in active\}} s(i, \omega)$
  if $f(s \cup \{i\})$ **then**
    $s \leftarrow s + i$;
  else
    $finish = true$
  $active \leftarrow active \setminus \{i\}$
return s

In summary, unlike [2, 3], we apply the composite framework in a heterogeneous environment, and tackle the problems that arise in multi-vertical composite retrieval by proposing a novel, entity based approach to assess item similarity and relevance.

4.1 Produce Bundles

We introduce two approaches for producing bundles: BOBO (adapted from previous approach) and CPS (newly proposed).

4.1.1 Bundles One-by-One (BOBO)

This method of producing a set of candidate bundles is inspired by *k-nn* clustering: a pivot is chosen at each step and a valid bundle is built around that pivot. If the bundle internal cohesion score is above a certain threshold μ, it is kept, otherwise it is discarded. The pseudo-code for this algorithm is shown in Algorithm 1.

Basically, BOBO approach starts with an empty set of candidate bundles, and considers each element in the item set as a possible pivot. The item set originates from the initial federated search rankings that merge results from all verticals. At each iteration an item is picked from the set of $Pivots$, and in our case, we choose the item (document) in $Pivots$ with the highest relevance estimation. Once a pivot is selected, we build a bundle S around it. This is done by the routine $pick\_bundle$ described in Algorithm 2. The routine greedily keeps picking the closest element to the pivot ω, as long as the bundle s does not exceed the pre-defined maximum number

of items for a bundle. The function f in Algorithm 2 also ensures vertical diversity, by enforcing the constraints that the bundle s is required to contain items from at least two different verticals.

Once a candidate bundle is created (by pick_bundle), the algorithm checks whether its internal cohesion ($score$ function in Algorithm 1) is larger than a pre-defined threshold μ. More precisely, to reflect cohesion, the $score$ function is defined as the expected similarity of item pairs $score(S) = \sum_{u,v \in S} s(u,v) / \sum_{u,v \in S}$. If this check has passed, then the bundle enters the bundle candidate set $Cand$ and the items within this bundle are removed from I and $Pivots$ so that they can no longer be selected again. Otherwise if the bundle S has a score lower than μ then it is discarded. In both cases the pivot ω is removed from $Pivots$ so that it is no longer considered.

4.1.2 Central-Plus-Satellite (CPS)

We introduce a different method, suggested by the observation that established vertical selection methods are prone to noise interference, and inspired by the approach to composite retrieval described in [5]. The basic idea is that we first produce bundles in a *central vertical* using BOBO and then we attach items from other *satellite verticals* onto the produced bundles. Our approach combines established vertical selection method (ReDDE [19]) with our entity based representation.

The pseudo-code for the CPS algorithm is shown in Algorithm 3. In the initial phase, vertical selection is performed using the ReDDE methodology to select the *central vertical* and *satellite verticals*. The top-1 vertical in the ReDDE vertical ranking is treated as the *central vertical*. A set of items from this central vertical forms the central item set I. Then, similarly, a set of *satellite verticals* is created (top-n verticals except the *central vertical*) and a set of satellite items coming from those verticals form the satellite item set S. In our experiments, we fixed the central vertical to "general web search" as we found this to minimise noise introduced by vertical selection, and set a threshold of $n = 2$ on the number of satellite verticals selected.

In the second phase, BOBO is used to generate and select a set $Cand$ of candidate bundles, using *only* items that originate from the central vertical. This provides us with a set of cohesive bundles to which we can attach items from the satellite verticals.

In the third phase, satellite items are attached to the bundles in $Cand$ only if a set of constraints (e.g. cohesion) are satisfied. The bundle-item similarity $s(b,i)$ is based on item-item similarity $s(u,i)$. We simply assume that a bundle b is represented by the elements common to all items within that bundle. In our experiments, after the bundle-item similarity estimation, we only add items to a bundle if the items contain at least a certain threshold (30%) of fraction of entities from that bundle.

4.2 Choose Bundles

Our approach of choosing bundles is different from the PAC approach [2] to the extent that we aim to select the bundles that have the highest degree of cohesion and relevance. Given the number of required bundles, k, a set $Cand$ of candidate bundles, and a similarity function between items, the algorithm (shown in Algorithm 4) selects the top k most cohesive bundles in the candidate set (determined by the function w). Relevance is considered using different relevance estimation approaches (Section 4.3.2).

Since we aim to provide bundles that are not only cohesive and relevant, but also topically diverse, we apply a post-diversification (Section 4.2.1) on the bundles we currently choose. We do not directly consider topical diversity when choosing bundles since doing so degraded the relevance of the bundles significantly.

Algorithm 3: Produce Bundles: CPS

Input: A central item set I, a set S of satellite items, a cost function f that checks diversity, cohesion and bundle size constraints, a threshold β on the number of items in a bundle, minimum bundle score μ, number of bundles c, the number of bundles to select k
Output: A set s of valid bundles

```
Cand ← BOBO(I, α, f, β, μ, c)
Cand ← ChooseBundles(k, Cand)
for b in Cand do
    i ← argmax_{i∈S} s(b, i)
    while f(b ∪ {i}) do
        b ← b ∪ {i}
        S ← S ∖ i
        i ← argmax_{i∈S} s(b, i)
return Cand
```

4.2.1 Post-Diversification

We propose two different diversification strategies to the set of bundles we select, based on how we estimate topical distance between bundles.

DT Diversification The first diversification strategy we consider is similar to Maximal Marginal Relevance (MMR) ranking strategy [9]. We denote this approach as *DT Diversity* and a suffix of DT is attached to the approach employed by this strategy (e.g. BOBO-DT). This is approach is based on applying MMR diversification to bundles (rather than documents) and the methodology to determine a distance d between two bundles S_i, S_j is defined as follows:

$$d(S_i, S_j) = 1 - argmax_{u \in S_i, v \in S_j} s(u,v) \tag{6}$$

The distance is computed as the maximum similarity between any two items in the two bundles. At each step, we select the bundle that is most cohesive and relevant, but at the same time the most dissimilar from the previously selected bundle.

DE Diversification We propose another diversification strategy that is based on explicitly diversifying query intents using entities (denoted as *DE Diversity*). The basic idea is that we estimate different subtopics (intents) of the query q by a set of query-specific entities q_e. We consider each entity $e \in q_e$ as being a subtopic of q, and we compute the probability of *aboutness* of an entity e to a document by simply using the frequency of the entities $freq(e)$ appearing in document d:

$$P(e|d_e) = \frac{freq(e)}{\sum_{e \in d_e} freq(e)} \tag{7}$$

After the estimation of entity-document "aboutness" as above, for every bundle $S \in Cand$ and every entity $e \in q_e$, we define a bundle-entity "aboutness" score that is calculated as an average of the entity-document "aboutness" score of all the documents in the bundle:

$$aboutness(S, e) = \frac{\sum_{d_e \in S, e \in q_e} P(e|d_e)}{\sum_{d \in S}} \tag{8}$$

Therefore, to diversify, for each entity e (treated as a subtopic or intent), we select the bundle that has the highest "aboutness" score to e and we assume that the bundle corresponds to the subtopic e of the query q.

4.3 Evidence and Estimation

We describe our approach for estimating similarity between items (Section 4.3.1) and estimating item relevance (Section 4.3.2).

Algorithm 4: Choose Bundles

Input: bundle number k, a set of candidate bundles $Cand$, $\omega(S \in Cand) = \sum_{u,v \in S} s(u,v)$
Output: a set Ω of valid bundles
$\Omega \leftarrow \varnothing$
while $|\Omega| < k$ **do**
  $u \leftarrow argmax_{\{v \in Cand\}} \omega(v)$
  $Cand \leftarrow Cand \setminus u$
  $\Omega \leftarrow \Omega \cup u$
return Ω

4.3.1 Estimating Similarity

Before attempting to produce cohesive bundles, we must define a measure of similarity between items (documents). There are two challenges when estimating the similarity of documents within a multi-vertical environment. Firstly, the documents are heterogeneous (general web documents vs. multimedia documents) and the different term distributions across verticals make the estimation difficult. Secondly, the retrieval ranking functions of documents from multiple verticals can vary. Therefore, we propose to use named entities as a bridge between verticals in assessing document similarity. We first assume that Wikipedia entries are representative of all the entities present in the documents. Then we used a state-of-the-art annotation tool that maps textual spots (contents) to Wikipedia entries [13] on our documents, such that every document d in our collection has a corresponding entity representation d_e. To further select the most representative entities from a document, we sort the entities in d_e using a traditional tf×idf measurement (entity frequency in the document multiplied by the inverse of its frequency across the collection), and select the top 100 entities for every document. In the end, we represent each document d by a 100-dimensional entity vector $d_e = \{e_1, e_2, ..., e_{100}\}$. Finally, we use the Jaccard coefficient of their entity sets to compute the similarity $s(u,v)$ of two documents u, v.

As mentioned in Section 4.1.2, we estimate the bundle-item similarity estimation $s(b,i)$ by simply assuming that a bundle b is represented by the entities that appear in all the items inside that bundle. Then we also use the Jaccard coefficient between entity sets of the bundle b and the item i to compute the similarity $s(b,i)$.

4.3.2 Estimating Relevance

To estimate the relevance of a document to a given query based on its entity representations, we annotate queries with entities by using a pseudo-relevance feedback technique. For a given query q, from the highest top 10 ranked documents in a BM25 retrieval setting, from each document we extract and sort entities as previously described. From the set of entities extracted from these documents, we select the top 10 most frequently occurring entities as the entity representation of the query $q_e = \{e_1, e_2, ..., e_{10}\}$.

We describe the methods used to estimate a document relevance. We have three sources of evidence: **V**: the estimated query-vertical similarity based on the ReDDE resource selection approach; **T**: the estimated query-document similarity based on the term-based BM25 ranking; and **E**: the estimated query-document similarity based on the entity representation of the query and document.

For **V**, we compute a probability $P(v|q)$ of a query q orientation to vertical v using the ReDDE [19] approach. *ReDDE* scores a target vertical based on its expected number of documents relevant to the query. It derives this expectation from a retrieval of a central sampled index that combines documents sampled from every target vertical. Given this retrieval, *ReDDE* accumulates a vertical score $ReDDE_q(v_i)$ from its document scores $P(q|\theta_d)$, taking into account the difference between the size of the original vertical N^{v_i} and a sampled set size N^{samp}.

$$ReDDE_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d \in topm} I(d \in v_i) P(q|\theta_d) \quad (9)$$

where $I(.)$ is a indicator function. To be consistent with [19], we choose $m = 1000$ in our experiments.

For **T**, we compute $P(d|q)$ as follow:

$$P(d|q) = \frac{bm25(d,q)}{\sum_d bm25(d,q)} \quad (10)$$

where the $bm25(d,q)$ is the BM25 scoring function [6].

For **E**, we compute the similarity of a document to a given query based on the entity representation $P(d_e|q_e)$ as follow:

$$P(d_e|q_e) = \frac{\sum_{e \in q_e, e \in d_e} P(e|q_e) \times P(d_e|e)}{\sum_{e \in d_e, e \in q_e}} \quad (11)$$

where $P(e|q_e)$ is estimated as the probability of generating entity e from the entity representations of the top 10 BM25 retrieved documents to the initial query q.

We can incorporate any of these three estimations of relevance into our objective function (i.e. $w(v)$ in Algorithm 4) when choosing bundles by simply including them as a factor in the initial objective function ($w(v)$) and we study their effectiveness on choosing bundles. We add a prefix of a given relevance estimation source to a composite retrieval approach if we incorporate it into the choose bundle stage. For example, BOBO-VT means BOBO approach that incorporates both vertical orientation relevance estimation V and term-based relevance estimation T, whereas BOBO means the original approach without incorporating relevance estimations at all.

5. EXPERIMENTAL SETUP

We list the test collection, the evaluation metrics and the baseline systems used in our work.

5.1 Data

We used a federated search test collection [15] as our test collection. This is a new public dataset used in the TREC FedWeb track 2013.[1] The collection contains search result pages from 108 web search engines covering a variety of information sources, ranging from "general web search engine" (e.g. Google, Yahoo!), to vertical search engines that focus on specific media or genres (e.g. YouTube and Wikipedia). Examples of verticals are listed in Table 1.

To provide a representation of each vertical search engine, several query-based samplings were provided for the vertical selection. For items (textual or multimedia documents) returned by each search engine, the authors collected relevance judgements by judging both the snippet created by the engine, and the actual document content. The TREC Web Track 2010 queries were reused to collect documents. This is an ideal test collection to use in investigating heterogeneous web search problems.

5.2 Evaluation

As mentioned before, there are several factors that can affect user perceived usefulness of a composite page. First, as assumed in the Cranfield paradigm setting, the **topical relevance** of the items significantly contributes to the utility of the page. In addition, as we aim to form bundles, each of them forming a coherent "story", the **cohesion** of the items within a bundle (whether the items focus on the same aspect) also impact the utility of the page. Indeed,

[1] https://sites.google.com/site/trecfedweb/

Table 1: Verticals Used in this Paper.

Vertical	Document	Examples
Image	online images	Photobucket
Video	online videos	Hulu, YouTube
Jobs	job description pages	LinkedIn Jobs, Simply Hired
News	news articles	Google News, ESPN
Blog	blog articles	Google Blogs, WordPress
Q&A	answers to questions	Yahoo Answers, Answers.com
Shopping	product shopping page	Amazon, eBay, Discovery Channel Store
Academic	research paper or technical report	Nature, CiteSeerX, SpringerLink
Encyclopedia/Dict	encyclopedic entries	Wikipedia, Encyclopedia Britannica
Books	book search pages	Google Books, Columbus Library
Social	Social interaction or sharing services	Facebook, MySpace, Tumblr, Twitter
General web	standard web pages	Google, Yahoo, AOL, Bing, Baidu

as shown in [12], presenting results incoherently in terms of topicality resulted in user dissatisfaction. We hypothesise that with a more complex heterogeneous environment, this could be more severe. The **diversity** of the result set (the set of bundles) may also have an effect on individual user preference of the search results. It was shown in [16] that there was a preference amongst users for systems that were measured to have more **topical diversity** (e.g. α-$nDCG$ [10]) for faceted or ambiguous informational needs. Finally, [24] found that **vertical diversity** plays an important role in user preference of aggregated search pages in a heterogeneous web environment. All the above mentioned factors are important in evaluating the performance of composite retrieval. In this paper, we measure performance with **topical relevance**, **cohesion**, **topical diversity** and **vertical diversity**. We report performance results for each metric separately, allowing us to obtain a complete understanding on the effectiveness of our proposed approaches.

Since we are mostly concerned with results returned at the high rank of the page, we report our **relevance** performance based on precision metrics (P@5, P@10, P@30) and a set of top-heavy rank-biased metric ($nDCG$@5, $nDCG$@10, $nDCG$@30). For **cohesion** and **topical diversity**, we report normalised cohesion metric $tcoh(P)$ and normalised diversity metric $tdiv(P)$, respectively (described in Section 3.1). For **vertical diversity**, we use the expected intent-recall $vdiv(P)$ which is the fraction of verticals (intents) with relevant items retrieved, that are present on the composite page P.

5.3 Baseline Systems

We compare our composite retrieval system with three baselines: (1) "general web" search engine only (**GW**); (2) traditional federated search systems (**FS**); and (3) aggregated search systems (**AS**). Basically, "general web" search engine forms the basics of web search today and we use it as our basic baseline. Traditional federated search systems, rather than building bundles, aims to simply merge rankings of different search engines into one single ranking. Aggregated search is the most similar approach to ours. However, different from our composite retrieval work, in AS systems, result presentation is based on a block paradigm (a block only presents results from one single vertical) and does not have the concept of "bundle". Note that comparing our systems with "general web" is not "fair" as we have more information (results from other verticals). However, we include this to demonstrate the effectiveness of our approach in combining information from heterogeneous (sometimes more noisy) vertical sources on the web.

For the **GW** baseline, we index the general web collection only and use BM25 as our ranking function. For the **FS** baseline, we use the state-of-the-art ReDDE [19] resource selection approach to rank the relevant verticals and CORI [8] result merging approach to merge results from different resources. We have another **FS** baseline FSC (Federated Search Central). FSC is obtained by mixing items from all different verticals into a central index and all items are ranked by a traditional ranking function (BM25). For our **AS** baseline, we rank the verticals based on ReDDE (same as **FS**) while we use a simple fixed-threshold approach to select relevant verticals (always selecting top-3 verticals as relevant for the query). For embedding the selected vertical results into the "general web" results, we use a simple approach. Following previous work [25], there are three possible embedding positions: ToP (top of the page), MoP (middle of the page) and BoP (bottom of the page). We simply embed our top-1 vertical on ToP, second vertical on MoP and third vertical on BoP. Although we have not developed state-of-the-art AS systems based on machine learning techniques [4] (as this is not our main focus), we think this is sufficient to illustrate and compare our approaches.

6. EXPERIMENTAL RESULTS

We report our experimental results in a homogeneous environment in Section 6.1, followed by results in a heterogeneous environment in Section 6.2. In the latter section, we study the importance of the different sources of evidence in estimating relevance for our two approaches as well as the effectiveness of using entity representations of heterogeneous items in our framework. Finally, we study the robustness of composite retrieval approaches. We aim to answer the following research questions (elaboration of our three main research questions specified in Section 1):

- **(RQ1).** In the homogeneous space, can we use composite retrieval to improve relevance compared to traditional one?
- **(RQ2).** In the heterogeneous space, can we build composite pages that are more relevant than existing solutions, such as aggregated search and federated (merged) ranking?
- **(RQ3).** Which sources of evidence are mostly essential (or optional) to form bundles for composite retrieval?
- **(RQ4).** Can we extract query-related entities as an anchor to bridge the semantic gap between the items retrieved from heterogeneous sources to form bundles?
- **(RQ5).** How robust are the different composite retrieval approaches to the quality of the initial rankings?

We use the following settings for our composite retrieval algorithms. We set the constraint of the maximum number of items allowed within the bundle to be 3 to mimic current web search settings (in aggregated search). Obtaining the optimal setting of this

Table 2: Performance of various composite retrieval approaches in the homogeneous web environment "General Web" only (GW). Pairwise t-test is applied for significance test, and values marked with △(p-value=0.05), ▲(p-value=0.01) and ▽(p-value=0.05), ▼(p-value=0.01) indicate respectively significant improvements or deterioration over GW baseline.

	GW	BOBO	BOBO-DT	BOBO-DE
P@5	0.540	0.624	0.440	0.636
P@10	0.562	0.564	0.416 ▼	0.586
P@30	0.577	0.343 ▼	0.343 ▼	0.343 ▼
nDCG@5	0.333	0.479	0.350 △	0.461
nDCG@10	0.373	0.453 △	0.342 ▼	0.452 △
nDCG@30	0.428	0.352 ▼	0.314 ▼	0.349 ▼

constraint is left for future work. For a composite page, we assume 10 bundles are presented.

6.1 Homogeneous Composite Retrieval

To answer **RQ1**, we conducted our experiments on the "general web" only search engine. The results are given in Table 2. The baseline is GW (traditional BM25 ranking function). Table 2 reports the retrieval performance of various composite retrieval approaches (BOBO, BOBO-DT, BOBO-DE) in the homogeneous web environment ("General Web" only). We do not include the CPS approach since it does not apply to a homogeneous environment (no satellite items can be attached). Note that as in the web setting, generally, top results (e.g. top-10 or top-5) are the major evaluation concern. Pairwise t-test significance test is conducted to identify significant improvement or deterioration over the GW baseline. We identify several trends from Table 2:

- Generally speaking, composite retrieval approaches perform comparatively as well as the baseline in the top rankings. The BOBO-DE approach performs the best in all the composite retrieval approaches in this setting and can significantly improve the relevance performance over the traditional retrieval baseline (GW), especially in the top ranking (indicated by both precision and nDCG metrics).
- All BOBO approaches perform significantly better in the early ranking (e.g. top 5 results) but significantly worse in the latter ranking (e.g. top 30 results). This might be due to the more irrelevant items introduced from the latter ranking (i.e. higher probability to pick irrelevant items as a pivot).
- Comparing BOBO-DT and BOBO-DE, we can observe that promoting inter-bundle topical diversity after ranking bundles can either deteriorate (BOBO-DT) or keep (BOBO-DE) unchanged. This suggests that our approach to promote diversity is not effective to boost relevance in the homogeneous environment. This is not surprising since relevance and diversity have been empirically demonstrated to act against each other in homogeneous web settings [10].

Note that composite retrieval in the homogeneous setting is similar to cluster-based retrieval in IR. It is not surprising that our approach performs well in this setting since cluster-hypothesis is well understood and it has been shown that positioning the documents of query-specific clusters at the top of the result list [14] yields highly effective retrieval performance (in terms of relevance) rather than ranking documents directly.

Returning to **RQ1**, we can conclude that composite retrieval can significantly improve retrieval performance in the top ranking in a homogeneous web search environment.

6.2 Heterogeneous Composite Retrieval

In this section, we aim to test the effectiveness and robustness of our proposed composite retrieval approaches.

6.2.1 Effectiveness

To answer **RQ2** and **RQ3**, we conducted our experiments on the heterogeneous test collection. The results are presented in Tables 3 and 4. Part (a) of Table 3 reports the performance of various composite retrieval approaches (BOBO, BOBO-DT, BOBO-DE, CPS, CPS-DT, CPS-DE) in the heterogeneous web environment with federated search baseline FSC (Federated Search Central). FSC is obtained by mixing items from all different verticals into a central index and all items are ranked by a traditional ranking function (BM25). This system is generally assumed to be the oracle system (the upper-bound system performance) in the federated search area. Part (b) of Table 3 presents the results related to the different relevance estimation approaches over BOBO and aims to track the performance change. In addition to relevance, results are compared based on a set of other criteria defined for evaluating composite retrieval (Section 3.1): topical cohesion (*tcoh*), topical diversity (*tdiv*) and vertical diversity (*vdiv*). These evaluation criteria can only be applied to compare the composite retrieval approaches and therefore the baseline system FSC is excluded in these comparisons. Pairwise t-test significance test is conducted to indicate significant improvement and deterioration of composite retrieval approaches over the baseline (FSC). The findings can be summarised as follows:

- Federating heterogeneous information is a challenging problem. Even the central federated approach FSC performs significantly worse than the homogeneous "general-web only" ranking GW. For example, when we look at $nDCG@10$ in Tables 2 and 3, we observe that FSC(0.287) < GW(0.373) and this is statistically significant at p-value=0.01.
- Most of the composite retrieval approaches perform significantly better than the baseline (FSC) in the early rankings (top 10). BOBO is the worst performing approach whereas CPS-DT approach performs the best. Compared with BOBO, we can observe that promoting inter-bundle topical diversity after ranking the bundles (BOBO-DT, BOBO-DE) can significantly improve the performance of composite retrieval.
- An interesting fact is that CPS-DT performs significantly better than both FSC and GW [2] in the early rankings. This is partly because of the conservative nature of the approach. It uses GW as the anchor and therefore it is more careful when selecting items from verticals.
- When comparing cohesion and topical diversity, we observe that the CPS based approaches generally produce bundles that are more cohesive and more topically diverse. However, compared with the BOBO based approaches, they are less vertically diverse. This can be explained by the fact that CPS is more conservative in the sense that it favours cohesion and topical diversity when forming bundles and only adds vertical items when it is sufficiently confident. On the other hand, the BOBO based approaches favour vertical diversity.
- When adding relevance estimation of items in the bundle, BOBO-E approach incorporating the entity-based relevance $P(d_e|q_e)$ performs best and it significantly improves over the baseline. Indicated by $nDCG@5$ and $nDCG@10$ from

[2]We calculate the significance of CPS-DT against GW and found that it significantly outperforms over GW at $nDCG@5$ (p-value=0.05) and $nDCG@10$ (p-value=0.01).

Table 3: Performance of various composite retrieval approaches in the heterogeneous web environment based on FSC (Federated Search Central): central ranking of mixture of all verticals. Pairwise t-test significance test is applied, and values marked with △(p-value=0.05), ▲(p-value=0.01) and ▽(p-value=0.05), ▼(p-value=0.01) indicate respectively significant improvements or deterioration over baseline (FSC).

	(a) Composite Retrieval Approaches							(b) Adding Relevance Estimation			
	FSC	BOBO	BOBO-DT	BOBO-DE	CPS	CPS-DT	CPS-DE	BOBO-VT	BOBO-VE	BOBO-E	BOBO-T
P@5	0.448	0.500	0.568 △	0.536	0.536 ▲	0.604 ▲	0.560 △	0.568	0.508	0.572 △	0.600
P@10	0.460	0.510	0.540	0.534 △	0.568 ▲	0.588 ▲	0.568 ▲	0.514	0.520	0.538 △	0.568
P@30	0.505	0.472	0.472	0.472	0.296 ▼	0.296 ▼	0.296 ▼	0.436 ▼	0.456 ▽	0.466	0.464
nDCG@5	0.260	0.327	0.401 ▲	0.364 ▲	0.331 ▲	0.395 ▲	0.380 ▲	0.427 ▲	0.355 △	0.393 ▲	0.413
nDCG@10	0.287	0.345	0.400 ▲	0.367 ▲	0.373 ▲	0.412 ▲	0.398 ▲	0.404 △	0.367	0.388 △	0.415
nDCG@30	0.351	0.362	0.383	0.371	0.291 ▽	0.308	0.306	0.368	0.349	0.369	0.382
tcoh	-	0.301	0.301	0.301	0.676	0.676	0.676	0.289	0.262	0.271	0.297
tdiv	-	0.180	0.180	0.180	0.268	0.268	0.268	0.174	0.159	0.161	0.167
vdiv	-	0.260	0.260	0.260	0.115	0.115	0.115	0.241	0.255	0.266	0.273

Table 4: Performance of best-performing composite retrieval approaches in the heterogeneous web environment against all baselines. Values marked with △, ▲indicate, respectively, significant improvements over BOBO-DT and CPS-DT (in this order). Similar convention with ▽, ▼indicates values below BOBO-DT and CPS-DT. Statistical significance is established by paired t-test of p-value<0.05 in all cases.

	FSC	FS	AS	OVS-FS	OVS-AS	GW	BOBO-DT	CPS-DT
P@5	0.448 ▽▼	0.352 ▽▼	0.388 ▽▼	0.532	0.444 ▽▼	0.540	0.568	0.604
P@10	0.460 ▼	0.368 ▽▼	0.346 ▽▼	0.568	0.510 ▼	0.562	0.540	0.588
P@30	0.505 ▲	0.363 ▽	0.323 ▽	0.576 △▲	0.563 △▲	0.577 △▲	0.472	0.296
nDCG@5	0.260 ▽▼	0.192 ▽▼	0.229 ▽▼	0.303 ▽▼	0.241 ▽▼	0.333 ▽▼	0.401	0.395
nDCG@10	0.287 ▽▼	0.206 ▽▼	0.219 ▽▼	0.350	0.303 ▽▼	0.373 ▽▼	0.400	0.412
nDCG@30	0.351	0.229 ▽	0.219 ▽▼	0.409 ▲	0.388 ▲	0.428 △▲	0.383	0.308

BOBO-VT and BOBO-VE, we observe that adding vertical orientation estimation $P(v|q)$ can also boost retrieval performance. BOBO-T, which incorporates only term-based relevance estimation $P(d|q)$, performs the worst and does not significantly improve over the baseline. This demonstrates the effectiveness of using entity representation for relevance estimation across heterogeneous verticals.

In Table 4, we compare the best performing approaches (BOBO-DT and CPS-DT) to a set of baselines and demonstrate their effectiveness. We aim to find out whether the best performing composite retrieval approaches in the heterogeneous web environment can outperform other baselines (different search paradigms): "General Web" Search only (GW), Federated Search (FS) and Aggregated Search (AS). Table 4 compares each baseline against BOBO-DT and CPS-DT respectively and shows whether each baseline performs significantly worse. Since we found in our experiments that vertical selection greatly affects retrieval performance, we also add two artificial systems (OVS-FS, OVS-AS) that use the oracle vertical selection (using OVS as prefix, indicating the upper bound of VS performance). To obtain OVS, assuming that we have relevance assessments for the items, we rank all verticals based on the recall of relevant items and set a simple cut-off threshold (verticals with fraction of relevant items smaller than 10% are not selected). Those two systems aim to reflect the oracle performance of FS and AS. Several trends can be observed in Table 4:

- When comparing different baseline search paradigms, we observe that FS and AS are similar, and AS outperforms slightly at top ranks, indicated by $nDCG@5$ and $P@5$. When the oracle vertical selection is added, the performance of each search paradigm increases significantly and OVS-FS outperforms OVS-AS. As demonstrated before, GW performs well and the performance is similar to OVS-FS.
- Composite retrieval approaches generally outperform all other search paradigms (GW, FS and AS) significantly in the early rankings (top 5 or top 10). One interesting fact is that they significantly outperform GW in the early ranking, which suggests that incorporating results from other vertical can improve retrieval performance. This is different from the conclusions we obtain when comparing FS and AS against GW where we found that heterogeneous federation degrades retrieval performance whereas FS and AS performed significantly worse than GW.
- Another interesting fact is that composite retrieval, especially CPS-DT, can even beat other search paradigms where the oracle vertical selection is applied. This might be due to the fact that the proposed composite retrieval is more conservative when adding vertical results as only results that are cohesive and potentially relevant are added.

Going back to **RQ2**, the composite retrieval paradigm can outperform both aggregated search and federated (merged) ranking on relevance performance in a heterogeneous environment. Returning to **RQ3**, relevance estimation is useful for improving composite retrieval approaches whereas both entity-based item relevance estimation and vertical orientation can effectively improve performance. Since we observe that by using entity, all our composite retrieval approaches (BOBO and CPS based) perform comparatively well in the heterogeneous environment, therefore, we answer **RQ4** that entities can be used as a bridge to semantically link heterogeneous items.

6.2.2 Robustness

To answer **RQ5**, we varied the initial rankings that composite retrieval approaches are based on and investigated the robustness of different composite retrieval approaches. The results are reported in Table 5. The robustness in this context refers to the extent to which the composite retrieval approach can still perform well when the initial ranking is degraded in terms of relevance performance. Therefore, it is compared against the original ranking where the

Table 5: Robustness of various composite retrieval approaches to initial ranking quality. All systems are measured by nDCG@10. Pairwise t-test significance test is applied, and values marked with △(p-value=0.05), ▲(p-value=0.01) and ▽(p-value=0.05), ▼(p-value=0.01) indicate respectively significant improvements or deterioration over the original ranking.

nDCG@10	Original	BOBO	BOBO-DT	BOBO-DE
FS	0.206	0.221	0.243	0.234
AS	0.219	0.211	0.220	0.211
FSC	0.287	0.345	0.400 ▲	0.364 ▲
OVS-AS	0.303	0.308	0.377 ▲	0.308
OVS-FS	0.350	0.307	0.371	0.370
GW	0.373	0.453 △	0.342 ▼	0.451 △

first column in Table 5 specifies the original ranking (ranked by the relevance performance $nDCG@10$ in a ascending order). The column headed by "Original" specifies $nDCG@10$ of the original ranking where the intersection of a given column and row specifies the $nDCG@10$ score of a given composite retrieval approach that is based on the corresponding original ranking. Note that, because CPS constructs central bundles using "general web search" documents to which it attaches documents from satellite verticals, it is not influenced by the quality of various initial rankings and it is not included in our robustness analysis.

We can observe the following trends: when the relevance performance of the original ranking is low (e.g. $nDCG@10$ lower than 0.25), the composite retrieval BOBO approach suffers from the large number of irrelevant items within the ranking and could not produce bundles that contain many relevant items. In addition, there is a general trend that when the relevance of the original ranking improves, the performance of the BOBO approach increases. Specifically, BOBO-DT performs the best on robustness.

Returning to **RQ5**, we conclude that, in general, composite retrieval is robust with regard to the initial ranking quality.

7. CONCLUSIONS

Our objective was to investigate whether composite retrieval can promote *relevance*, *cohesion* and *diversity* in a heterogeneous multi-vertical web search environment. Our results indicate that composite retrieval can significantly improve the performance over various current search paradigms, such as traditional "general web only" ranking, federated search ranking and aggregated search.

We showed that composite retrieval can significantly improve retrieval performance in both homogeneous and heterogeneous web search environments (**RQ1**). In particular, in the heterogeneous environment, our proposed composite retrieval approach CPS-DT outperformed all current state-of-the-art search paradigms ("general web" search only, federated search and aggregated search). We also demonstrated (**RQ2**) that incorporating our proposed entity-based relevance estimation of items and vertical orientation estimation (based on the state-of-the-art resource selection approach ReDDE) improves composite retrieval approaches compared to those that disregard them (e.g., BOBO). Finally, we found that our proposed BOBO-DT is the most robust approach for composite retrieval with respect to the initial ranking quality in the heterogeneous web environments (**RQ3**).

Our results have implications for work in heterogeneous information access and diversity in IR. The composite retrieval search paradigm aims to promote a diverse information space for users to explore. Diversity is promoted in two dimensions, topical and vertical. For an exploratory task, rather than issuing multiple queries with respect to different aspects of an information need to several vertical search engines, composite retrieval provides a unified page that consists of *relevant*, *cohesive* and *diverse* stories to explore. Promoting both topical diversity and multi-vertical information aggregation is however challenging [10, 4] and our work is the first substantial endeavour to push along both lines. We have demonstrated that without losing relevance (actually our approach can perform better in terms of relevance performance), we can promote diversity for both dimensions. The IR and web search research communities have already gained numerous insights into single dimension diversification. Our work opens a fruitful research avenue as heterogeneous information access is becoming ubiquitous.

Our work has several limitations:

- *Relevance Evaluation*: We used $P@n$ and $nDCG@n$ to evaluate the relevance of a composite page. However, the user model underlying these metrics (user reads from the ranked list from top to bottom) will probably not hold for composite retrieval. A more elaborate user model and associated metrics are required.
- *Diversity Evaluation*: We used intent-recall to evaluate vertical diversity and both cohesion and inter-bundle separation to evaluate topical diversity. For vertical diversity, although we understand that composite retrieval promotes more verticals containing relevant items, how to capture vertical diversity among bundles requires further investigation. For topical diversity, we need to further understand whether our cohesive bundles reflect actual subtopics of an information need and whether intra-separation correlates with the coverage of subtopics. Nonetheless, our current, albeit simple, evaluation approach already provides insights on performance with respect to diversity.
- *Optimisation*: Our optimisation is based on two greedy approaches that were adapted to our context. There are parameters (e.g. maximum number of items per bundle) that have not been explored. We only set these parameters to mimic current web search scenarios. Thus the optimisation of such parameters (e.g. by machine learning) is left for future work. In addition, other optimisation approaches could be employed (e.g. cluster-and-pick [2]). Due to space limits, these were not explored.

Future works include conducting user studies similar to those performed in the aggregated search domain [20] to better understand user browsing behaviour on the composite pages and the impact of different factors (e.g. task type, etc.). In addition, more rigorous evaluation metrics for this search paradigm must be developed for further in-depth and reliable investigation of composite retrieval performance.

Acknowledgments

This work was partially funded by the Linguistically Motivated Semantic Aggregation Engines (LiMoSINe[3]) EU project.

[3]www.limosine-project.eu

8. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM, 2009.

[2] S. Amer-Yahia, F. Bonchi, C. Castillo, E. Feuerstein, I. Méndez-Díaz, and P. Zabala. Complexity and algorithms for composite retrieval. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 79–80. International World Wide Web Conferences Steering Committee, 2013.

[3] S. Amer-Yahia, F. Bonchi, C. Castillo, E. Feuerstein, I. Méndez-Díaz, and P. Zabala. Composite retrieval of diverse and complementary bundles. 2013. http://www.optimization-online.org/DB_FILE/2013/02/3785.pdf.

[4] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 315–322. ACM, 2009.

[5] S. Basu Roy, S. Amer-Yahia, A. Chawla, G. Das, and C. Yu. Constructing and exploring composite items. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD '10, pages 843–854, New York, NY, USA, 2010. ACM.

[6] D. C. Blair. Information retrieval, 2nd ed. c.j. van rijsbergen. london. *Journal of the American Society for Information Science*, 30(6):374–375, 1979.

[7] A. Brodsky, S. Morgan Henshaw, and J. Whittle. Card: a decision-guidance framework and application for recommending composite alternatives. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, pages 171–178, New York, NY, USA, 2008. ACM.

[8] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–28. ACM, 1995.

[9] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.

[10] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM.

[11] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. Automatic construction of travel itineraries using social breadcrumbs. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 35–44, New York, NY, USA, 2010. ACM.

[12] S. Dumais, E. Cutrell, and H. Chen. Optimizing search by showing results in context. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 277–284. ACM, 2001.

[13] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1625–1628, New York, NY, USA, 2010. ACM.

[14] O. Kurland and C. Domshlak. A rank-aggregation approach to searching for optimal query-specific clusters. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 547–554. ACM, 2008.

[15] D. Nguyen, T. Demeester, D. Trieschnigg, and D. Hiemstra. Federated search in the wild: the combined power of over a hundred search engines. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1874–1878. ACM, 2012.

[16] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 555–562. ACM, 2010.

[17] R. L. Santos, C. Macdonald, and I. Ounis. Aggregated search result diversification. In *Advances in Information Retrieval Theory*, pages 250–261. Springer, 2011.

[18] M. Shokouhi and L. Si. Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102, 2011.

[19] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 298–305. ACM, 2003.

[20] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. Factors affecting click-through behavior in aggregated search interfaces. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 519–528. ACM, 2010.

[21] A. Tombros, R. Villa, and C. J. Van Rijsbergen. The effectiveness of query-specific hierarchic clustering in information retrieval. *Information processing & management*, 38(4):559–582, 2002.

[22] M. Xie, L. V. Lakshmanan, and P. T. Wood. Breaking out of the box of recommendations: from items to packages. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 151–158, New York, NY, USA, 2010. ACM.

[23] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 10–17. ACM, 2003.

[24] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating aggregated search pages. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 115–124. ACM, 2012.

[25] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Which vertical search engines are relevant? In *Proceedings of the 22nd international conference on World Wide Web*, pages 1557–1568. International World Wide Web Conferences Steering Committee, 2013.