Vertical Selection in the Information Domain of Children

Sergio Duarte Torres, Djoerd Hiemstra and Theo Huibers
University of Twente
The Netherlands
duartes,hiemstra,huibers@cs.utwente.nl

ABSTRACT

In this paper we explore the vertical selection methods in aggregated search in the specific domain of topics for children between 7 and 12 years old. A test collection consisting of 25 verticals, 3.8K queries and relevant assessments for a large sample of these queries mapping relevant verticals to queries was built. We gather relevant assessment by envisaging two aggregated search systems: one in which the Web vertical is always displayed and in which each vertical is assessed independently from the web vertical. We show that both approaches lead to a different set of relevant verticals and that the former is prone to bias of visually oriented verticals. In the second part of this paper we estimate the size of the verticals for the target domain. We show that employing the global size and domain specific size estimation of the verticals lead to significant improvements when using state-of-the art methods of vertical selection. We also introduce a novel vertical and query representation based on tags from social media and we show that its use lead to significant performance gains.

Categories and Subject Descriptors

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

Keywords

vertical selection, aggregated search, children, social media, evaluation

1. INTRODUCTION

In aggregated search, content is retrieved from different search services in the web and the content retrieved is integrated in a meaningful and consistent way. These search services are often referred as verticals, which are defined as domain specific collections, (*e.g.* entertainment, shopping, news) or collections from specialized types or genres (*e.g.* videos, images, songs). Generally, aggregated search systems are assumed to have complete access to the verticals, which implies having access to the query logs and to detailed statistical descriptors of the verticals.

In this work, we envisage a search system that integrates heterogeneous content from verticals, which are not fully accessible for the system (third party verticals). The system we envisage is intended to integrate content of different genres as it is done in aggregated search. In particular, we are interested in verticals that contain high quality information for children from 7 to 12 years old. In this system parents, teachers and other specialist on children care are allowed to add resources for children. For instance, they could add a vertical dedicated to coloring pages: `http://ivyjoy.com/colouring/search.html`, which only returns sheets of paper to be colored and that are suitable for children, or a vertical dedicated to search only videos: `http://www.youtube.com`, in this case the vertical provide content for all kind of public segments. We believe that an aggregated search system is a better solution for searching in the web content for children than simply crawling and indexing websites, or listing suitable content (*e.g.* `http://www.kids.yahoo.com`) because *(i)* it is more scalable, *(ii)* we can leverage and exploit the knowledge of parents and other experts by using hundreds of services suggested by them. Nonetheless, we believe a similar system would be also highly beneficial in other domains (*e.g.* business, health, sports) and other demographical segments (*e.g.* seniors, teenagers).

Under this scenario, once a query is submitted, the system has to decide first which are the most relevant verticals for the query. This problem has been characterized as a multi-class classification problem [3, 26] in which the objective is to predict the set of relevant verticals (or single vertical in [2]) from a set of predefined verticals that are accessible by the system. This problem is referred as vertical selection and it has been widely studied [2, 26, 15].

Our problem differs from those addressed in previous studies in that: *(i)* the vertical selection is carried out under the restriction of targeting a specific information domain (*e.g.* content for children); *(ii)* these users search for domain specific content in a set of verticals that may not be completely suitable for them, thus some verticals may provide only suitable content (*e.g.* coloring pages) but others may or may not contain suitable content for their information needs (*e.g.* youtube) since it contains content for all type of public and *(iii)* a test collection for this domain has not been built until this work and the process of gathering assessments is not straight forward given the nature of the targeted users.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.
Copyright 2013 ACM 978-1-4503-2077-1/13/07 ...$15.00.

The contributions of this paper are summarized as follow: *(1)* We provide a novel test collection for the problem of vertical selection in the domain of content for children. We describe its construction in sections 2 and 3. Given the novelty of the collection we describe two methodologies to gather relevant assessments using crowd sourcing: first assuming that the Web vertical is always displayed (which is the case in state-of-the-art aggregated search interfaces) [2] and the second methodology easing this restriction. We found that the two methodologies lead to a different set of relevant verticals and we show under which conditions this difference can be mitigated. We also observed that the first methodology is more prone to visual verticals preference bias, which has consequences in the design of the content aggregation algorithm and aggregated interfaces (Section 4).

(2) We estimate the children suitability of verticals by estimating the amount of child-friendly content and we compare them against the estimation of general content. We propose a simple approach to combine both estimations (children and non children vertical sizes) and we show that their used in *ReDDe* [19], a state-of-the-art vertical selection method, lead to significant improvements. We found that this method also outperform other state of the art methods such as *Clarity* (sections 5 and 6). *(3)* We present a novel method of vertical selection for domain specific scenarios based on a vertical representation using tags from social media and language models. Concretely, verticals are represented based on tags describing the urls from a sample of each vertical and a language model on these tags is employed to rank the relevancy of the verticals given a query. As far as we know social media has not been employed before as a source of evidence in the problem of vertical selection (sections 6.1 and 6.2).

2. COLLECTION CONSTRUCTION

The collection built consists of a carefully chosen set of queries and verticals. A set of vertical results was also retrieved for each (*query*,*vertical*) pair and relevant assessments mapping a set of relevant verticals to each query were gathered. Using this annotation schema, we can test and compare any pair of vertical selection methods. In the following sections we described in detail the methodology carried out to build the collection and we justify our decisions in the design of the collection.

2.1 Query set selection

The query set was extracted from the AOL query log [16]. We extracted queries landing on domains listed in the *Kids and Teens* directory. Given that only domains are displayed in the AOL log we carefully extracted those entries in which the exact domain is listed in the Dmoz directory (*i.e.* exact matching). An analogous procedure has been suggested for this information domain in [23, 7] and the Dmoz directory has been used in previous research in information retrieval for children [8, 10].

Query	Dmoz Category
1950's television shows	News
science fair projects ideas	School - Time/Science
barometric pressure	School - Time/Science
bingo song copyright	Arts/Music
rabbit ears	Sports - Hobbies/Crafts
secret code game	Games/Word_Play

Table 1: Example of queries and their Dmoz category

We leave out navigational queries by filtering out query-click pairs in which the domain is mentioned in the query. For instance, the query *sesame street* is filter out if it lands on the domain *www.sesamestreet.org*. We also filter out query containing tokens such as *.com www.*, *http* and *.org*. The Levenshtein distance between the query and domain mentioned in the url was also employed to filter out queries that misspell a domain (*e.g.* pbkids). Concretely, the queries were selected from search sessions satisfying either of the following restrictions:

1. There is at least one click event after submitting the query which lands in a domain listed in Dmoz and the duration of the click event is of at least 60 seconds

2. There is a click on the Dmoz domain is the last event of the session

The first restriction is employed to capture only the cases for which the click event has a long duration, that is, if users spend more than 1 minute on a web result is a strong indication that the result clicked is relevant for the query [13]. The second condition is also employed because previous query log studies have shown that the last click of the session can be often associated to successful searches [6]. We extracted a set of 3.8K queries by using both restrictions. Table 3 provides examples of the queries extracted and the Dmoz *Kids and Teens* category in which they belong.

2.2 Selection of verticals

The list of verticals selected is shown in Table 2. The verticals were manually selected to cover the information needs found in the query set and the distribution of topics targeted by children between 7 to 12 years old in the Web [9]. We found that several genre verticals need to be split into fine-grained information services to fit the topic interests of children. For instance in our system the vertical *games* may refer to the *video games* or to the *online gaming* vertical. Similarly the genre *images* may refer to *coloring pages*, *printable worksheets* or the standard *pictures* vertical. In previous literature, these services are wrapped under a single vertical (either images or games respectively) [3]. This fine-representation of verticals is highly convenient for young users because it increases the accessibility to rich media and this niche of users have been found to struggle identifying and searching information in non-web verticals [11]. Dedicated verticals offering content for children may be found in the Internet (*e.g.* video-games, stories, health), however some of the verticals displayed in Table 2 were constructed artificially by modifying the search parameters of general web services. For instance, the vertical *coloring pages* is constructed by using the *line-drawing* parameter of the Google Image search service. For the case of the *printable worksheets*, we employed the *line-drawing* parameter of the Google Image service and we modified the user's query by expanding it with the terms *worksheets.*

3. DATA CHARACTERISTICS

We describe the collection based on the number of queries covered per vertical and the distribution of the numbers of

Vertical	Websites
web	google.com
games online	onlineflashgames.com
games	gamespot.com
images	google.com/imghp
coloring pages	google.com/imghp (line drawing option)
worksheets	google.com/imghp (query + worksheets)
books	books.google.com
question/answers	worldoftales.com
stories	worldoftales.com
shopping	amazon/toys
music	allmusic.com
videos	youtube.com
movies	rottentomatoes.com
enciclopedia	wikipedia.com
reference	dictionary.kids.net.au
how-to	www.instructables.org
school aid	livescience.com
	howstuffworks.com
	dsc.discovery.com
school activities	sciencekids.co.nz
	enchantedlearning.com
	howtosmile.org
lyrics	lyricsdrive.com
health	kidshealth.org

Table 2: List of verticals and their urls

verticals covering a query. Figure 1 depicts the vertical coverage, which refers to the proportion of queries covered by each vertical. A vertical is said to cover a query if it returns at least one result when the query is submitted. From this figure, we observe that large verticals such as *web* and *videos* tend to cover most of the queries in the dataset. Note that even relatively small verticals such as *how-to* or *stories* have high query coverage, which suggest that the verticals chosen are appropriate for the information needs targeted by the chosen queries. Nonetheless, we observed that the verticals *movies* and *reference*, which are widely used in previous vertical selection studies, cover less than 25% of the queries in our data set. This result may indicate that these verticals are less suitable for the audience we are interested in. Figure 2 shows the number of queries in the collection that are covered by a specific number of verticals. For instance from this figure we can observe that there are no queries in the collection that are covered by exactly two or three distinct verticals. Similarly around 75 queries are covered by exactly 10 verticals. Interestingly we observed almost a normal distribution in which most of the queries are covered between 10 and 21 verticals (84% of the queries). On average, queries from the entire collection are covered by 16.8 verticals. This result shows that the problem of vertical selection in the domain of children topics is not straight forward since each query has on overage a set of 16 verticals from which to choose relevant verticals.

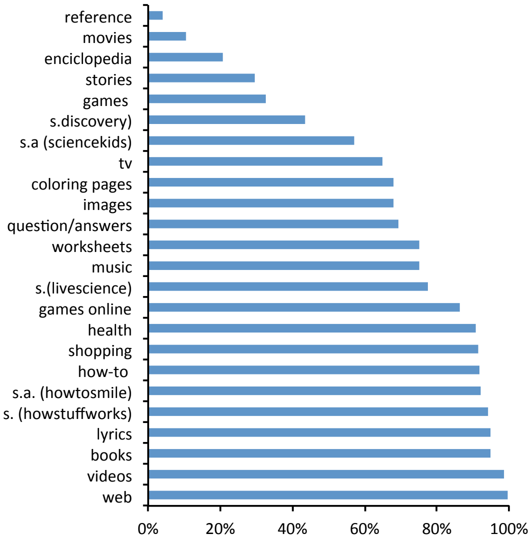

Figure 1: Queries covered by each vertical

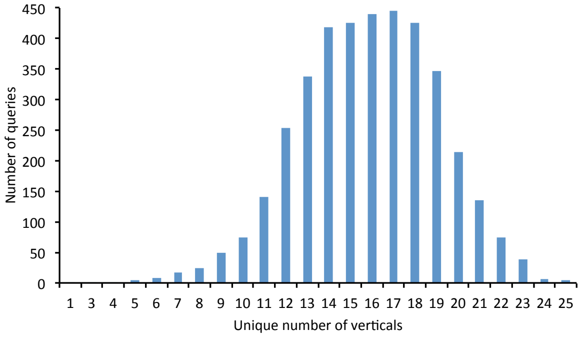

Figure 2: Distribution of unique verticals per query

4. GATHERING VERTICAL RELEVANCE ASSESSMENTS

We gather assessments by employing the crowd-sourcing engine Crowdflower [1] and a sample of 90 queries from the set of queries described previously. This sample was chosen to be representative of all the possible topics in the collection (based on the Dmoz categories of the queries). We carried out two experimental protocols for gathering the assessments. In the first protocol assessors were asked to chose between two set of vertical results: the results of the target vertical against the results returned by the web vertical (*i.e.* Google Web), each result set was displayed in a column next to each other in the survey page. Concretely, the top 4 results of each vertical were shown in each column and the order in which the columns appear was randomized. Nonetheless, the ranking of the results of each vertical was preserved. Adult assessors were able to choose *the most relevant set of results for the query (given that the content has to be suitable for users between 7 to 12)* between the two columns. The special option *none of the sets are relevant and suitable for children* was also given. This option was provided to avoid false positives since with this option users are not forced to chose a vertical when both sets are inadequate. The motivation of comparing each vertical against the web vertical is that in modern search engines the web results are always displayed and the results from other verticals are only displayed when the vertical results *add value* to the current web results, that is, when their results are preferred over the standard web results [1]. The main drawback

[1] http://crowdflower.com/

of this protocol is that we are unable to identify the cases in which the results from the web verticals are unsuitable for the query.

In the second protocol, we asked assessors to judge vertical results independently. This protocol is motivated by the fact that in an information system for children we may not always want to present the results from the *Web* vertical since the user may be requiring a different type of content (*e.g.* coloring pages, videos) or because the results from the web vertical are not suitable for children given query. For this reason we asked users to assess each set of vertical results independently using a scale-graded system of 5 points, from *bad* to *excellent* in terms of relevance to the query and appropriateness for children in the targeted age range. An advantage of this method arises in the possibility of evaluating directly the quality of the web vertical results when the information needs are targeted at young users. We can also rank the verticals based on the graded score assigned.

For both protocols, adult users were asked to make the judgments. We consider it reasonable to assume that adults can easily discern between content oriented for adults and children and thus that they are able to judge the results in the context of the domain. We also provide the queries to the assessors without a close description. The motivation for this was to let the assessors identify all the possible content that can be relevant for children given a query. To ensure the quality of the assessments a set of 120 golden judgments was created by the authors of this paper for each experimental protocol. These gold judgments were employed to avoid spammers by: *(i)* forcing the users of *CrowdFlower* to complete a training session in which they are shown only *units* (survey page) from the gold judgement set. They are allowed to start the task if they answer correctly at least 6 of these units; *(ii)* during the task the gold units are mixed with the units under evaluation, users answering incorrectly more than 8% of the gold units shown to them during the survey are ignored. Only users from the United States were allowed to carry out the survey to ensure language proficiency and domain knowledge (the queries were extracted from the US market). Each unit work was paid with 0.01 dollar cents and each unit was evaluated by at least 3 assessors. In the following paragraphs, we describe the assessments gathered and we compare the results obtained by both experimental protocols.

4.1 Distribution of relevant verticals

A gold test set was created by mapping each query in the sample to a set of relevant verticals by using the assessments collected in *Crowdflower*. For the first protocol (*i.e.* paired assessments), we map a query to a vertical if at least a certain percentage of annotators select the vertical as relevant for the query. The threshold 60% and 80% were used in the results reported. As a point of reference 60% means that more of half of the assessors agreed on classifying the target vertical as relevant while 80% means that most of the assessors agreed.

For all the thresholds we observed in Figure 3 a long tail distribution in which visual-oriented verticals are preferred, that is *YouTube*, *Google Images*, *Coloring pages* and *Worksheets* are the most frequent verticals assessed as relevant. This result may be due to the bias generated by visual content in the paired assessments. We believe that the exposition of visual content is more appealing to the user when they are asked to make an assessment against text based results. This hypothesis is also supported by the fact that the best performing verticals at higher threshold values are the image oriented verticals (*Google Images* and *Worksheets*). The bias towards visual oriented verticals has also been reported before in the literature [4, 20]. Anecdotally we also observed that children oriented educational websites (*e.g.* science for kids, instructables) were more frequently chosen as relevant than *Wikipedia*, which is a relatively trusted source. All the other verticals (*e.g.* music, lyrics, games) were less frequent and were located in the bottom of the long tail distribution.

For the second experiment protocol a vertical is said to be relevant if the averaged score assigned by all the assessors to a (*query*, *vertical*) pair is greater that a given threshold. Figure 4 shows the distribution obtained when using the values 3.0 and 4.0 as threshold (recall we used a graded system from 1 to 5 to judge each result set).

Additionally, we also employed as thresholds the averaged score obtained by the *Google Web* vertical (on a query basis) and the maximum score between the *Google Web* vertical averaged core and 3.0 respectively. The last two thresholds were employed in order to make a fair comparison between the relevant verticals obtained with the two experimental protocols. Recall that in the first experimental protocol we compare each set of vertical results against the *Google Web* vertical, for this reason we employed the Web vertical score as threshold for the experiments.

In Figure 4 we observed similar distributions for the thresholds 3.0 and 4.0, although all the frequencies in the latter are lower as a consequence of the higher threshold value. Some of the verticals in the long tail also rank differently (e.g. *Amazon*, *KidsHealth*, *Tv*). Nonetheless, the most frequent verticals are the same when using both thresholds: *Google Web*, *Google Images*, *YouTube* and *Yahoo! Answers*. For the thresholds *google-score* and *max(google-score,3.0)* we observed large differences in the distribution in respect to the first two thresholds employed. Even though the top 2 most frequent relevant verticals are the same (*Google Web* and *Google Images*. Verticals such as *Youtube*, *Yahoo! Answers* and *Google Books* are not prominent as it was the case before. In general terms all the other verticals were assigned as relevant with less frequency. This result suggests that the score obtained by the *web vertical* is often higher that the score assigned to the other verticals given that the frequency of the relevant verticals are significantly trimmed when using *Google Web*'s score as threshold. It is important to mention that even though this is the case, in about 10% of the queries, *Google Web* is not chosen as relevant (when using as reference 3.0 and 4.0 as thresholds).

An important observation of the previous results is that in the second experimental protocol we did not observed the visual content bias observed with the first experimental protocol, as it was shown with all the thresholds. These results suggest that the second methodology can also simulate the first when using the *web vertical* score as threshold while avoiding the bias generated when comparing text results against visual content.

4.2 Inter-assessor agreement

We analyzed the inter-assessor agreement for both experimental protocols to quantify the quality of the assessments under different relevant thresholds and to identify the rel-

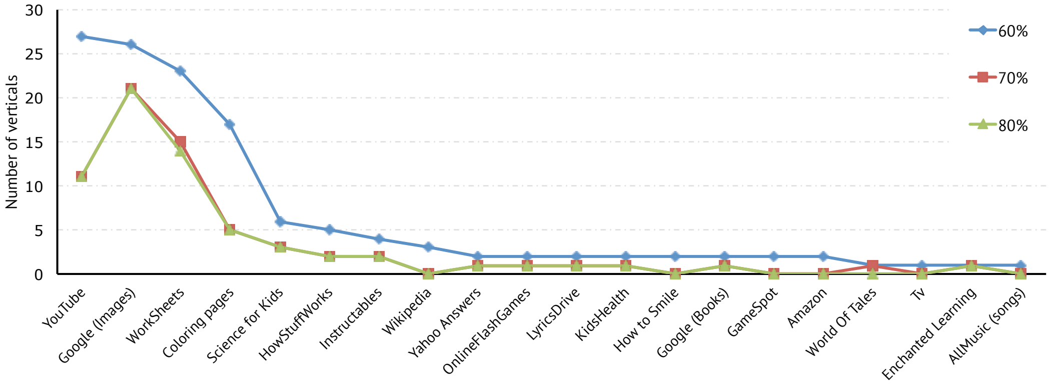

Figure 3: Frequency distribution of verticals for the first experimental protocol

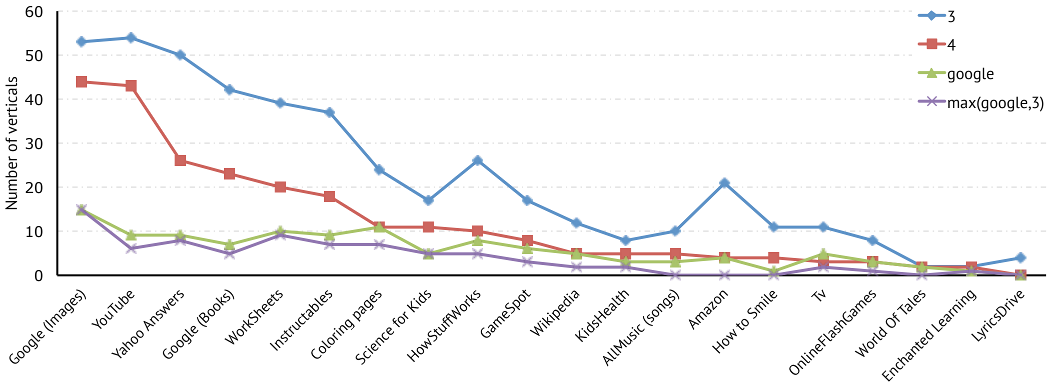

Figure 4: Frequency distribution of verticals for the second experimental protocol

Inter-agreement metric	Score
average pairwise percent agreement	82.86%
fleiss' kappa	0.683
FK agreement	0.828
average pairwise cohen's kappa	0.682
krippendorff's alpha	0.683

Table 3: Inter-agreement scores found for the task

evant threshold values that maximize assessor agreement. The motivation is to use these thresholds in our experiments of vertical selection. We are also interested in investigating the threshold for which both protocols lead to a similar set of relevant verticals. For the first experimental protocol, the survey consisted of a sample of 90 queries which lead to 3360 decisions (comparisons between the web vertical and each one of the other verticals) and each pair was evaluated on average by three assessors. Table 3 shows the inter-assessor scores obtained for the experiment. We list the most common metrics employed in IR and natural language processing. All the metrics show a substantial agreement between assessors. This result indicates that the task was consistently interpreted by the assessors and that assessors agreed in discerning content that is suitable for children in the age range specified.

For the second experiment protocol we measured the inter-assessor agreement by establishing that each pair *query, vertical* is assessed by n assessors (*e.g.* 3 coders) and the assessment is binary (relevant or non-relevant) based on the threshold defined in the previous section: 3, 4, web vertical score and *max(web vertical, 3.0)*. Note that this encoding is slightly different to the one employed in the first experimental protocol, in which we had three possible values: (relevant, non relevant (*web vertical* is preferred) and none of the two sets are relevant. Figure 5 shows the inter-assessor agreement using the Krippendorff's alpha score. We only report these results since the averaged Fleiss' kappa agreement obtained was almost identical to the Krippendorff's alpha scores. We found that at lower threshold values the inter-assessor agreement in higher, which is expected since lower threshold values represent a larger score interval to discern between relevant and non-relevant. It was observed that the lowest agreement is obtained when using the *web vertical* score as threshold. In general terms, the inter-assessor agreement was slightly lower that in the first experimental protocol.

Additionally we compared the agreement between the lists of relevant vertical collected using both experimental protocols. The agreement was also measured using the Krippendorff's alpha score. In this case, we have two coders (the results of each experimental protocol) and the coding is binary: relevant or not relevant. We compute the score for all the possible threshold combinations of the two methods. Recall that in the first protocol the threshold refers to the percentage of users assessing a vertical as relevant while in the second protocol the threshold refers to the graded score (between 1 and 5) and the cut-off is performed by considering non-relevant the pairs for which the averaged score of

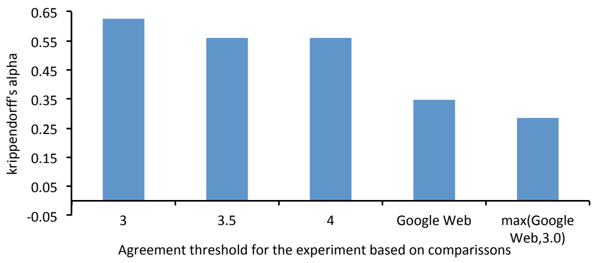

Figure 5: Inter-assessor agreement for the second experiment protocol for several thresholds

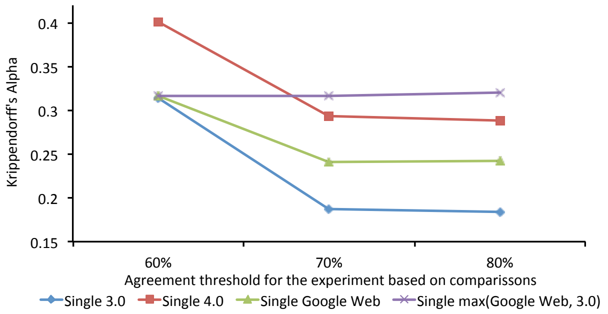

Figure 6: Agreement between the two experiment protocols

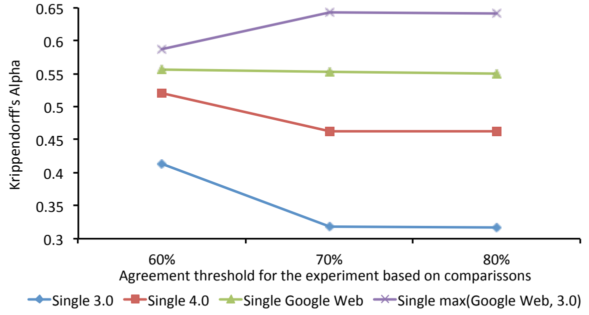

Figure 7: Agreement between the two experiment protocols (Including Google Web)

all the assessors fall below the threshold defined. It is important to mention that we estimated the agreement scores under two scenarios: with and without considering the *web vertical*. We made the distinction because this vertical is not assessed directly in the first protocol since it assumes that the results from verticals are relevant only if they add value to the results provided by the *web vertical*. Nonetheless, we artificially created an assessment for this vertical by setting it as relevant if for a given query it is preferred by all the assessors at least in one of the paired comparisons. Figure 6 and Figure 7 shows the results obtained using these two modalities (*i.e.* with and without the *web vertical*) respectively. For the former the maximum agreement score obtained was 0.41 using as threshold 4.0 for the second protocol and 60% for the first protocol. Scores between 0.4 and 0.6 are considered moderate agreement [5, 17]. Nonetheless we observed that the score values were more stable when we set the second protocol threshold as *max(web vertical, 3.0)*. This result is interesting because it suggest that by using the *web vertical* score we can simulate, at some extend, the vertical set obtained by the first protocol, having the advantage of avoiding the visual bias identified. In Figure 7, we observed larger agreement scores (maximum of 0.632). Similarly, higher values of agreement are obtained consistently when using the score of the Google Web vertical.

On overall, the previous results suggest that both approaches lead to relatively high inter-assessor agreement. However, the first protocol provides more consistent results since the inter-assessor agreement is higher. It is important to mention that the second protocol is not prone to visual bias and provide a wider set of relevant verticals per query, which is useful in the construction of exploratory information systems. In addition, we found that for the second protocol that the threshold *60%* lead to the highest assessor-agreement, thus we will employed this threshold for our experiments (Section 6.2).

5. VERTICAL SIZE ESTIMATION

The corpus size estimation is highly important to understand verticals' characteristics and quality. The size estimation of a corpus is also a key feature in the selection of search engines in federated search and distributed search [22, 25]. In our scenario, the estimation of the vertical size is crucial since this statistic is needed in the best performing resource selection methods such as *ReDDe*. Recall that the system for children we envisage has only access to the verticals through limited query interfaces in which is only possible to submit a query to receive a limited number of results. Under this non-cooperative environment, the search engines do not provide collection summaries from which global statistics about the vocabularies of the collection can be inferred, for this reason they need to be estimated from vertical samples. We will show how these estimations are used in vertical selection in Section 6.

Si and Callan [19] proposed the use of the *capture-recapture* method through query-based sampling to estimate the size of the collections for the problem of resource selection in non-cooperative environments. The capture-recapture method has been used traditionally in the Ecology field for the estimation of the population size of species. It works as follow: a predefined number of animals are captured, marked and then released. After a certain amount of time, a second sample of animals is captured in the same area and the new sample is inspected to estimate the intersection between the two samples. The population is estimated using the sizes of the two samples and their intersection using the following expression:

$$s' = \frac{|sample_A| * |sample_B|}{|sample_A \cap sample_B|} \quad (1)$$

In search engines this process is carried out using query-based sampling: A set of queries is sent to the target search engine and the documents returned by the search engines are collected. This process is repeated and the estimation is based on the size of the number of documents collected in the two samples.

Shokouhi et al. [18] explores the problem of resource selection in non-cooperative environments and propose a query-based sampling *QBS* method based on the capture-recapture method [19] referred as *multiple capture-recapture*. They showed with a large set of heterogeneous collections that their method outperforms previous approaches using search engines in non-cooperative environments. We employed their method to estimate the sizes of the verticals chosen for our collection. In their method, the *capture-recapture* process is repeated T times using samples of size m and the estimation is carried out by counting the size of the intersection of each pair of samples. Concretely the estimation is performed with the following expression:

$$s' = \frac{T(T-1)k^2}{2D} \qquad (2)$$

where T is the size chosen for each sample, k is the number of samples and D is the accumulated number of duplicates found in the intersection within each pair. In our approach, we are not only interested in estimating the size of the collection but also in the size of the content available for the target domain in each vertical. We employed the *multiple capture-recapture method* and a random samples of queries from the query set of our collection to estimate the size of the content of interest for children in each vertical . Recall that these queries were chosen to be representative of the topics of interest for children, as it was described in 2.1.

We carried out an analogous process to estimate the size of the verticals of content that is not oriented to children. For this purpose we employed a set queries known to be submitted to extract information in other domains (*i.e.* non for children). For this set we employed the same methodology described in section 2.1 but instead of using the *Kids and Teens* seed urls we employed the global categories of Dmoz which are not present in the categories for children. In this fashion, we obtained size estimations for each vertical of content that is oriented for children and non-children. The results of the estimates are shown in Table 4. These values were obtained by using a set of 2K queries. We set the parameters T and k to 50 and 25 respectively. The sample was constructed by choosing randomly 5 queries from the set of 2K queries (with replacement) and collecting the top 10 results for each query. Similar parameter values have been used in previous studies [18]. It is important to mention that the set of 90 queries employed to gather user assessments were not employed in the samples generated for the size estimation process to avoid bias in the evaluation of vertical selection methods.

Consistently, we observed large ratios between the estimations using the *grown up* and *kids* queries with verticals known to be large and targeting all kind of public. For instance, the ratios for the verticals *web*, *question/answers*, and *images* were 34.0, 1.6 and 13.0 respectively. Inverse ratio trends were observed when considering verticals focused on children topics, such as the gaming and educational verticals.

The ratio found for the verticals *games online* and *education (livescience)* were 0.6 and 0.4 respectively. These results are interesting for resource selection because the ratio gives us an estimation of the likelihood to find content for children in the vertical. In the following section, we will explore the use of this ratio along with the vertical size estimation in the problem of vertical selection.

Vertical	**Kids**	**Grown ups**
web	803,297	27,377,757
games online	51,235	31,282
games	79,910	36,965
images	4,749,306	63,878,640
coloring pages	776,072	1,258,471
worksheets	931,287	1,006,777
books	963,956	30,115,533
question/answers	8,613,190	14,553,214
stories	3,928	3,478
shopping	562,308	22,485,899
music	121,920	540,501
videos	728,842	177,227
movies	61,443	131,131
encyclopedia	61,186	409,160
how-to	172,481	114,445
school(livescience)	11,666	5,308
school(howstuffworks)	3,267	3,674
school(discovery)	18,241	23,117
s. activities (sciencekids)	4,851	3,597
s. activities (howtosmile)	3,267	11,027
lyrics	455,305	161,271
health	7,185	2,939
tv	13,390	310,538

Table 4: Vertical size estimations using the set of *kids* and *grown up* queries

6. RESOURCE SELECTION METHODS IN IR FOR CHILDREN

We employed two well-known vertical selection methods: *ReDDe* and *Clarity*. The former has been shown as one of the most effective methods for resource selection in federated search in cooperative and non-cooperative environments [19]. More recently, this method was adapted for state-of-the-art aggregated search systems [2] and it was proven to be one the most discriminative sources of evidence in vertical selection. It is important to mention that models built on query logs have been shown to provide higher performance than *Redde* [2, 3]. However query logs are inaccessible in the settings of the aggregated search system we are interested in since the verticals belong to third parties, contrary to the case of the aggregated search considered in [2]. For this reason, we employed in this work *ReDDe*, which is defined in equation 3.

$$ReDDe_q(V_i) = |V_i| \sum_{d \in R} I(d \in S_i^*) p(q|M_d) p(d|S_i^*) \qquad (3)$$

where $p(d|S_i^*) = \frac{1}{|S_i^*|}$, $|V_i|$ is the size of the vertical, S_i^* is the set of sampled vertical documents and $p(q|M_d)$ is the query likelihood score of the document d. The score is estimated from an index combining the document samples of all the verticals. In this work we index the documents using the Terrier search engine[2] and score them using the Hiemstra's Language Model [14].

The second baseline employed is *Clarity*, which was originally used as a measure of retrieval effectiveness. *Clarity* is

[2] http://terrier.org

defined in the following way:

$$Clarity_q(C) = \sum_{w \in V}^{|V|} p(w|M_q) log \left(\frac{p(w|M_q)}{p(w|M_c)} \right) \quad (4)$$

where $P(w|M_q)$ is defined as:

$$p(w|M_q) = \frac{1}{z} \sum_{d \in R} p(w|M_d) p(q|M_d) \quad (5)$$

where $p(w|M_c)$ is the probability of w which is estimated using the collection language model and $p(w|M_c)$ is the query likelihood score of the document and it is estimated using a index of the documents of the target collection only: $p(q|M_d)$. The variable z is defined as $\sum_{d \in R} p(q|M_d)$. The documents are also scored using language model with linear interpolation smoothing [14]. For *ReDDe* and *Clarity* the top 100 documents where employed to calculate the document scores.

We also redefine *ReDDe* to exploit the size estimations obtained with the children and non-children queries. The intuition is to use the ratio between these two size estimations to boost those verticals that contain more content available for children since these verticals have a higher likelihood of providing content that is suitable for them. Equation 6 shows the new definition:

$$ReDDe'_q(V_i) = \frac{|V_i^{kids}|}{|V_i^{adults}|} \sum_{d \in R} I(d \in S_i^*) p(q|M_d) \quad (6)$$

where $|V_i^{kids}|$ and $|V_i^{adults}|$ are the size estimations obtained using the two set of queries. We will show in the next section that this definition lead to better performance when using our test collection.

6.1 Representing verticals through social media

Nowadays, social media is widely used to describe and share web resources in the Internet. We believe that the descriptions provided by these thousands of users can be beneficial for the problem of vertical selection, particularly on specialized domain environments.

In this work, we utilize bookmarks from the social website *Delicious*[3] in which users can share bookmarks of their favorites websites by providing a list of describing tags. For instance, tags describing the domain *www.howtosmile.org* (such as *science*, *math*, *lessons*) can be used to emphasize this vertical if we are able to infer that the tags are related to the intent of the query (*e.g.* which is the case for the query *school science fair*).

For this purpose we create a tag representation of the query and the vertical. A language model is employed to assign a retrieval score to the verticals. For both representations we used the Delicious crawl collection built in [24], which contains around 130 million bookmarks.

The query is represented as a bag of tags using the top *10* results of an index containing the documents in the Dmoz *kids and teens* section. The intuition is that the tags describing the top results from this index are a fair representation of the intent that the query has in the domain of information for children. Similarly, a vertical is represented as a bag of tags associated to the urls from the sample of documents extracted from the target vertical. It is important to mention that we associate each url to a set of tags by *(i)* finding the url in the collection provided in [24], and by *(ii)* extracting tags from the title and snippet of the results by using the vocabulary of tags. The latter strategy was used given the low coverage of the collection for the small verticals.

Based on this vertical representation we rank the verticals for a query using a language model: $p(V|Q)$, which is defined in the following fashion:

$$p(V_i|Q) = \frac{p(Q|V_i)p(V_i)}{p(Q)} \propto p(t) \prod_{j=1}^{|Q|} p(q_j|V_i) \quad (7)$$

$$p(q_j|V_i) = \frac{cf(q_j, V_i) + \mu \, p(q_j)}{|V_i| + \mu} \quad (8)$$

where $p(q_j)$ is the prior probability of q_j and μ is the Dirichtlet smoothing parameter. These probabilities are estimated using MLE on the artificially documents of tags created for each vertical and query.

We combine this probability score with the *ReDDe* score defined in Equation 3. For this purpose, we normalize *ReDDe* scores across verticals for each query and we weighting in the following manner:

$$LMReDDe_q(V_i) = p(V_j|Q) * ReDDe_q(V_i)^* \quad (9)$$

where $ReDDe_q(V_i)^* = ReDDe_q(V_i) / \sum_{k=1}^{|V|} ReDDe_q(V_k)$. An analogous definition was applied to *ReDDe-R* expressed in Equation 6.

6.2 Experimental Results and Discussion

We compared the performance of *ReDDe-R* (Equation 6), the social media language model (Equation 7), *LMReDDe* and *LMReDDe-R* (Equation 9) against the state-of-the-art methods*ReDDe* and *Clarity*. For our experiments, we employed the 90 queries annotated with human assessments using both protocols: paired comparisons and single vertical assessments, which will be referred in our results as protocol *A* and *B* respectively. We employed as threshold values 60% and 3.0 to define a relevant vertical in the experimental protocols *A* and *B* respectively. For the language model we set experimentally the parameter μ to 2500. It is important to mention that other threshold values lead to consistent results, nonetheless due to space restrictions we only reported results with the values mentioned.

Figures 8 and 9 shows the precision and recall curves obtained for all the methods using the gold set obtained through the experimental protocol *A*. We found that *Clarity* is consistently outperformed by all the other methods and that the best performing methods are the ones based on *ReDDe*. We also observed that our three *ReDDe* variations outperformed both baselines being *LMReDDe* the best performing method for protocol *A* and *LMReDDe-R* for protocol *B*.

We noted that the *web* verticals is generally the first ranked vertical by most of the methods and that several queries in out test collection were only associated to this *web vertical*. Concretely, we found that from the 90 queries of the test set, 27 queries were associated only to the *web vertical*. For this reason, we decided to repeat our experiments ignoring this vertical from the collection. We believe that by ignoring this vertical we will have a clearer picture of the performance of the methods, especially on smaller verticals.

[3] `http://delicious.com`

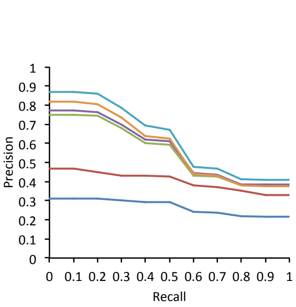

Figure 8: Protocol A (with *web vertical*)

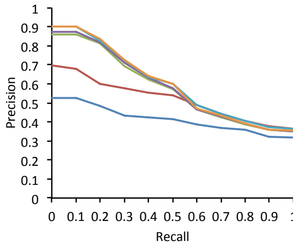

Figure 9: Protocol B (with *web vertical*)

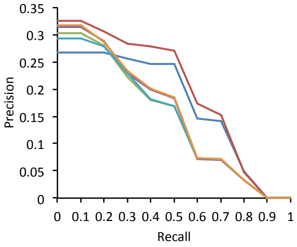

Figure 10: Protocol A (without *web vertical*)

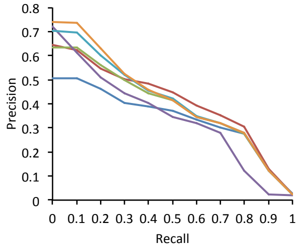

Figure 11: Protocol B (without *web vertical*)

Protocol	Web vertical	Clarity	LM	ReDDe	ReDDe-R	LMRedde	LMReDDe-R
A	Yes	0.242	0.375	0.527	0.548	**0.606**	0.564
	No	0.150	**0.179**	0.137	0.146	0.137	0.149
B	Yes	0.305	**0.588**	0.552	0.556	0.573	0.564
	No	0.287	0.377	0.360	0.382	0.382	**0.393**

Table 5: MAP results

Figure 10 and 11 shows the precision-recall curves obtained by ignoring the effect of the *web vertical*. As it was the case before, *clarity* was outperformed by all the other methods. However we observed that *LM* outperform all the other methods using protocol *A*, similarly with *LMReDDe-R* for protocol *B*. We believe this result is interesting because it shows the potential of using social media for vertical representation. Recall this metric makes only used of social media tags to rank the verticals while all the other methods make uses of the entire content of the sampled documents.

Table 5 shows the MAP values obtained by all the methods with the two test sets and with and without the *web vertical*. The results are in line with the performance observed in the precision-recall curves. We verify the statistical significance of our results by comparing each pair of methods using the paired t-test for the equality of means with unequal variance. A statistical significance was acknowledged if the probability of the two means being equal (*e.g.* null hypothesis) is smaller than 5%. We found that all the differences were statistical significant except for the pairs: *ReDDe - LMReDDe* (using protocol A without the web vertical), *ReDDe-R - LMReDDe* (with *B* without the web vertical) and *ReDDe-R - LMReDDe-R* (with *A* and without the web vertical).

As a final remark, we observed that the models can behave differently according the test set. For example, the *LM* method seems to provide more gain for the test set created with the first protocol. We believe that further research is required to provide more robust mechanism to combine the scores of the different methods and to understand the scenarios in which each methods is more beneficial.

7. RELATED WORK

7.1 Aggregated Search

In [2, 3] is proposed a machine learning approach to combine the scores of several resource selection methods. *ReDDe*, which was initially proposed for federated search and databases[19], is adapted for large-scale aggregated search systems. Our work differs from theirs in that: *(i)* we are interested on domain specific areas (*e.g.* content for children); *(ii)* our methods are applied on public available resources instead of proprietary environments (absent of query logs); *(ii)* we relax the restriction of having only one relevant vertical per query since we observed that even though the number of relevant verticals is small, the average is far greater than 1 (around 3.5); *(iv)* a collection is provided to the research community. We hope this will motivate the further study of vertical selection in the domain of children.

In [2] vertical relevance assessments are gathered by asking trained annotators to find the most suitable vertical for the query. In [1] aggregated search interfaces are assessed using paired comparisons. This methodology was used in IR initially in [21]. In this work, we employed the paired comparison methodology to gather assessments and we show that these may lead to preferences towards visual oriented verticals. We proposed an alternative methodology that does not suffer of this bias.

7.2 IR for children

Duarte et al. [23, 7] presented a methodology to extract queries focused on retrieving information for children using public data. Their motivation is to carried out a query log analysis to compare the behavior of children and grown ups. In [9], the authors contrast their first results against the results obtained with a set of queries actually submitted by children. They found consistent results in terms of topic trends and topic distribution in both data sets. For this reasons we used their work as a starting point to build our collection. Our work differs in that we are interested in the evaluation of vertical selection methods. Our collection builds on top of this query set by adding a set of verticals, their documents and vertical relevance assessments. In [12] is proposed a variation of page rank to promote websites appropriate for children. Eickhoff et al. [10] proposed a learner to classify web pages in appropriate and non-appropriate for children. In [8] is proposed a biased random walk to recommend queries for children. All these studies employed the Dmoz *kids and teens* directory, which is also employed by us. However, these studies focus only on the web vertical while our work involves a large set of verticals of heterogeneous genres, and our interest is on choosing the best collection for children instead of filtering content or re-ranking solely on the web vertical.

8. CONCLUSIONS AND FUTURE WORK

We presented a detailed description of a simple methodology to build a collection for the problem of vertical selection in the domain of content for children. We contrast two methodologies to gather relevant assessments using *CrowdFlower*. We proved that the two methods lead to a different set of relevant verticals and that the former is prone to visual bias. We show that the different sets obtained by these methods can also lead to differences in the performance of vertical selection methods. We believe that the choice of either methodology is highly dependent of the targeted aggregated search system. For instance if the web vertical is always displayed it may be more beneficial to employ the paired comparison method since it has higher inter-assessor agreement. Nonetheless, further refinements are needed given the visual bias obtained by this method. We found that tags from social media are an effective resource for the problem of vertical selection given that for several experiment settings was the best performing feature. Similarly, the ratio between the sizes estimations (*i.e.* children and grown ups) lead to a significant performance gain.

There are several directions for future work. We would like to verify that grown ups judgments correlate with the judgments of children in the targeted age group. We showed a simple language model to rank verticals using tags from social media. However, more sophisticated methods to exploit these resources are worth to explored (*e.g.* random walk, semantic latent models). We also consider worth verifying the applicability of the methods proposed in this paper in other information domains. (*e.g.* teenagers, elderly)

9. ACKNOWLEDGEMENTS

This research was partially funded by the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 231507.

10. REFERENCES

[1] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A methodology for evaluating aggregated search results. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 141–152, Berlin, Heidelberg, 2011. Springer-Verlag.

[2] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *SIGIR '09*, pages 315–322, New York, NY, USA, 2009. ACM.

[3] J. Arguello, F. Diaz, and J.-F. Paiement. Vertical selection in the presence of unlabeled verticals. In *SIGIR*, pages 691–698, 2010.

[4] J. Arguello, W.-C. Wu, D. Kelly, and A. Edwards. Task complexity, vertical display and user interaction in aggregated search. SIGIR '12, pages 435–444, New York, NY, USA, 2012. ACM.

[5] A. Bermingham and A. F. Smeaton. A study of inter-annotator agreement for opinion retrieval. SIGIR '09, pages 784–785, New York, NY, USA, 2009. ACM.

[6] D. Downey, S. Dumais, D. Liebling, and E. Horvitz. Understanding the relationship between searchers' queries and information goals. CIKM '08, pages 449–458, New York, NY, USA, 2008. ACM.

[7] S. Duarte, D. Hiemstra, and P. Serdyukov. An analysis of queries intended to search information for children. In *IIiX 2010*.

[8] S. Duarte Torres, D. Hiemstra, I. Weber, and P. Serdyukov. Query recommendation for children. CIKM '12, pages 2010–2014, New York, NY, USA, 2012. ACM.

[9] S. Duarte Torres and I. Weber. What and how children search on the web. CIKM '11, pages 393–402, New York, NY, USA, 2011. ACM.

[10] C. Eickhoff, P. Serdyukov, and A. P. de Vries. A combined topical/non-topical approach to identifying web sites for children. WSDM '11, pages 505–514, New York, NY, USA, 2011. ACM.

[11] E. Foss, A. Druin, R. Brewer, P. Lo, L. Sanchez, E. Golub, and H. Hutchinson. Children's search roles at home: Implications for designers, researchers, educators, and parents. *JASIST*, 63(3):558–573, 2012.

[12] K. Gyllstrom and M.-F. Moens. Wisdom of the ages: toward delivering the children's web with the link-based agerank algorithm. In *CIKM*, pages 159–168, 2010.

[13] A. Hassan, R. Jones, and K. L. Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 221–230, New York, NY, USA, 2010. ACM.

[14] D. Hiemstra. Term-specific smoothing for the language modeling approach to information retrieval: the importance of a query term. In *SIGIR*, pages 35–41, 2002.

[15] Ilad Shokouhi and L. Si. Federated search. In *FNTIR*, 2011.

[16] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*, InfoScale '06, New York, NY, USA, 2006. ACM.

[17] P. Schaer. Better than their reputation? on the reliability of relevance assessments with students. In *CLEF*, pages 124–135, 2012.

[18] M. Shokouhi, J. Zobel, F. Scholer, and S. M. M. Tahaghoghi. Capturing collection size for distributed non-cooperative retrieval. SIGIR '06, pages 316–323, New York, NY, USA, 2006. ACM.

[19] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. SIGIR '03, pages 298–305, New York, NY, USA, 2003. ACM.

[20] S. Sushmita, H. Joho, and M. Lalmas. A task-based evaluation of an aggregated search interface. In *SPIRE*, pages 322–333, 2009.

[21] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. CIKM '06, pages 94–101, New York, NY, USA, 2006. ACM.

[22] P. Thomas and D. Hawking. Evaluating sampling methods for uncooperative collections. SIGIR '07, pages 503–510, New York, NY, USA, 2007. ACM.

[23] S. D. Torres, D. Hiemstra, and P. Serdyukov. Query log analysis in the context of information retrieval for children. In *SIGIR*, pages 847–848, 2010.

[24] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *MSoDa*, pages 26–30, 2008.

[25] J. Xu, S. Wu, and X. Li. Estimating collection size with logistic regression. SIGIR '07, pages 789–790, New York, NY, USA, 2007. ACM.

[26] K. Zhou, R. Cummins, M. Lalmas, and J. Jose. Evaluating large-scale distributed vertical search. LSDS-IR '11, pages 9–14, New York, NY, USA, 2011. ACM.