Evaluating Large-Scale Distributed Vertical Search

Ke Zhou
School of Computing Science
University of Glasgow
Scotland, U.K.
zhouke@dcs.gla.ac.uk

Ronan Cummins
School of Computing Science
University of Glasgow
Scotland, U.K.
ronanc@dcs.gla.ac.uk

Mounia Lalmas
Yahoo! Research Barcelona
Barcelona, Spain
mounia@acm.org

Joemon Jose
School of Computing Science
University of Glasgow
Scotland, U.K.
jj@dcs.gla.ac.uk

ABSTRACT

Aggregating search results from a variety of distributed heterogeneous sources, i.e. so-called verticals, such as news, image, video and blog, into a single interface has become a popular paradigm in large-scale web search. As various distributed vertical search techniques (also as known as *aggregated search*) have been proposed, it is crucial that we need to be able to properly evaluate those systems on a large-scale standard test set. A test collection for aggregated search requires a number of verticals, each populated by items (e.g. documents, images, etc) of that vertical type, a set of topics expressing information needs relating to one or more verticals, and relevance assessments, indicating the relevance of the items and their associated verticals to each of the topics. Building a large-scale test collection for aggregate search is costly in terms of time and resources. In this paper, we propose a methodology to build such a test collection *reusing* existing test collections, which allows the investigation of aggregated search approaches. We report on experiments, based on twelve simulated aggregated search systems, that show the impact of misclassification of items into verticals to the evaluation of systems.

Categories and Subject Descriptors

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

General Terms

Theory

1. INTRODUCTION

Increasingly diverse content is available on the web, in terms of media (e.g. text, image, video, audio), and genre (e.g. reference/wiki, FAQs, news, blogs). Until recently each type of content was dealt with in a separate way through so-called *search verticals*, and users switched between verticals to access information of a given type. Now, there is a growing tendency to "aggregate" search results from those different large-scale distributed verticals into one single interface. This is referred to as aggregated search [10], and is implemented by most major search engines.

In order to properly evaluate those various large-scale distributed vertical search systems, there is a need for a test collection for research into the various stages involved in aggregated research. This is the main focus of this paper. A test collection for aggregated search requires verticals, each populated by items of that vertical type, a set of topics expressing information needs relating to one or more verticals, and relevance assessments, indicating the relevance of the items and their associated verticals to each topic. Constructing such a large-scale test collection from scratch is very time-consuming. Therefore, for aggregated search, there is the need to *re-use* existing (and emerging) collections to allow for evaluation in a timely fashion and with the required focus. In this way, as new, more focused verticals become available, they can be seamlessly integrated into the existing collection. It should be noted that our focus is not to replace the test collection creation methodology used in TREC, but rather utilise a similar methodology to create a practically useful, reliable, and consistent large-scale test collection for the aggregated search community. The contribution of this paper is three-fold:

- We outline the process and methodology of *reuse* to construct a test collection for aggregated search from existing test collections (Section 3).
- We build a practical test collection for aggregated search by using a SVM classifier to classify items into various types (Section 4).
- We investigate the impact of misclassification (of items into verticals) to the evaluation of systems, by experimenting with twelve large-scale distributed vertical search systems on simulated test collections with various misclassification rate (Section 5).

The remainder of the paper is as follows: Section 2 provides some background. Section 3 describes our methodology and introduces some important concepts. Section 4

LSDS-IR'11, October 28, 2011, Glasgow, Scotland, UK.

describes the stages and design decisions involved in building the test collection. Section 5 details experiments carried out to investigate the consistency of the constructed test collection. Finally, Section 6 outlines our conclusions and future work.

2. RELATED WORK

In this section we discuss research that is related to both aggregated search and test collection design.

2.1 Aggregated Search

Aggregated search is the task of searching and assembling information from a variety of sources (i.e. verticals) and placing it in a single interface [10]. Aggregated search can be compared to federated search (distributed information retrieval) and desktop search. In federated search, test collections have been developed reusing existing test collections by partitioning different text-based corpora into a number of sub-collections. The partitions [15] are generally based on topicality, publication source, date, or domain. In desktop search, test collections [9] have been created by collecting different types of information (e.g. email, web-page, office documents, etc.) for individuals.

However, there are several major differences between federated search and aggregated search. Firstly, much of the research into federated search has utilised documents of the same type (i.e. mainly newswire documents), and do not investigate a truly rich and diverse information space. Aggregated search focuses on the web context, with various existing web-based vertical search engines (Blog, News, Image, etc.) available. In addition, dynamically changing content (e.g. the daily update and emergence of verticals such as twitter) is a more critical issue in aggregated search. In aggregated search, query logs can provide a rich source of user interaction data which is not applicable in the traditional federated search domain. Aggregated search deals with such diverse heterogeneous information and interaction data, that previous research into distributed test collections (e.g. those conducted on newswire collections) cannot be assumed to hold in such a scenario.

Aggregated search can be deconstructed into three areas; (1) *vertical representation*, (2) *vertical selection*, and (3) *result presentation*. For vertical representation (1), each vertical (or resource) is modelled in a specific way so that certain features of the vertical can be accessed in order to more easily identify the type and quality of the vertical for a particular topic. The test collection that we propose to build will allow the same type of evaluation as has previously been adopted in the federated search domain, but will be more appropriate in the web context. Vertical selection (2) can be treated as a multi-class query classification problem [2], and our created test collection for aggregated search can, thus, be used for training or evaluating such a classifier. Evaluating result presentation (3) consists of testing whether the right items have been retrieved from the correct verticals and, whether they are presented in the right position for the user.

A test collection for aggregated search would allow extensive investigation relating retrieval effectiveness and result presentation, and also create a valuable shared resource for the community. In particular, our methodology is based on reusing existing web collection (ClueWeb09 [1]) and multimedia collections, and the vertical partitioning reflects a realistic scenario (i.e. realistic verticals) on the web. The items contained in the verticals are of different media (e.g. image and text) and genres (e.g. Blog, News, Wiki, etc.). In addition, all the topics in our test collections simulate real information needs as they come from search engine query-logs. These, although not perfect, more accurately reflect an aggregated search scenario. It should be noted that at this stage we do not investigate the temporal nature of verticals in this paper.

2.2 Test Collections

A test collection typically consists of three parts: a set of items (often documents) to be searched, a set of information needs (stated in topics), and associated relevance judgments (referred to as qrels), stating the items that are relevant to each topic. The most time-consuming part of generating a test collection is the creation of relevance judgments. Although methods to alleviate this problem have been proposed (e.g. formally selecting a subset of the most promising items to be judged [5] or using crowd-sourcing techniques [1]), judging a set of items (often documents), and in particular, heterogeneous documents from a variety of sources, remains an extremely tedious task.

In addition, current test collections have become extremely large (e.g. Clueweb09) to reflect the much larger amount of information in many of today's retrieval scenarios. As a result, the idea of reusing test collections has been proposed. Some researchers [7] have reused an existing Q&A test collection to generate a test collection to investigate diversity in IR. Others [6] have developed means to quantify the reusability of a test collection for evaluating a different retrieval scenario than that originally built for.

Other related work of creating a test collection for aggregated search includes collecting pair-wise judgments on information originating from different verticals (vertical blocks) via crowd-sourcing [1], or inferring judgments from query-logs [12]). Our approach emphasizes the *reuse* of collections and judgments, and furthermore, leads to a reusable collection. Importantly, yet lacking in the aggregated search domain, the construction of a test collection allows different parts of an aggregated search system (i.e. vertical representation, vertical selection, and result presentation) to be systematically evaluated on a stable collection.

3. METHODOLOGY

In this section, we describe the methodology of *reuse* adopted herein to construct an aggregated search test collection. The methodology can be categorized into several steps:

1. First, we *define* the verticals that we want to investigate (Section 3.1).

2. Then, we decide which existing test collections to use and how to *simulate* verticals (i.e. by classification) (Section 3.2).

3. Thirdly, we *identify* a set of topics, from existing ones that are utilized in various evaluation forums (e.g. TREC), that could be satisfied by documents that are contained in several (one or many) simulated verticals (Section 3.3).

[1] http://lemurproject.org/clueweb09/

Table 1: Vertical used (simulated) in this paper

Vertical	Document	Type
Image	online images	media
Video	online videos	
Recipe	recipe page	genre
News	news articles	
Books	book review page	
Blog	blog articles	
Answer	answers to questions	
Shopping	product shopping page	
Discussion	discussion thread from forums	
Scholar	research technical report	
Reference/Wiki	encyclopedic entries	
General web	standard web pages	

4. Furthermore, we discuss how existing *relevance assessments* can be used correctly so that the aggregated collection remains reliable.

3.1 Defining Verticals

In web search, a vertical is associated with content dedicated to either a topic (e.g. "finance"), a media type (e.g. "images") or a genre (e.g. "news")[2]. In this paper, we are mainly concerned with the latter two types, which is less well-studied than the former (e.g. topic-focused distributed collections have been studied in federated search [16]). Consistent with existing web search engines, we consider the verticals listed in Table 1. These verticals can be simulated by existing test collections (mainly web-based and multimedia collections), as we show in Section 4. The last vertical in Table 1, "general web", consists of the standard web search pages, that form the majority of search results [10]. It is to these results that results from other verticals are added, if relevant [2].

3.2 Simulating Verticals

For the purposes of building our aggregated search collection, two main types of existing test collections are available. The first type of collection are those that could be used in their entirety, to simulate a vertical. The second type of collection are those that need to be decomposed into parts, each of which could be used to simulate a vertical, or part thereof. Examples of the latter include large-scale web collections, comprised of documents that are not only standard web documents, but of various genres (e.g. news, wiki, blogs, etc). Documents in such a collection are more problematic as they need to be classified into a genre, and then added to the corresponding vertical.

3.3 Identifying Topics

Now we must identify a subset of the topics (from all available topics) that could reflect concrete search scenarios in aggregated search. Following [2], this subset should consist of approximately 1/4 of the topics for which only the "general web" vertical is of high *vertical intent*, and 3/4 for which more than one vertical (including "general web") is of high *vertical intent*. At this stage, we must clarify the concept of "*vertical intent*" when referring to a vertical. We define two criteria to determine the *vertical intent* of a vertical:

1. **Topical relevance**, i.e. the vertical should contain at least one topically relevant document (i.e. it should be capable of satisfying the user's need in a topical manner).

2. **Vertical orientation**, i.e. the degree to which a specific type of information, originating from one specific vertical, satisfies a user's information need (e.g. images are highly oriented to the topic "photographs of flowers").

We state that a topic has a high *vertical intent* to a specific vertical only when both criteria are satisfied. Therefore, to identify a set of usable topics, we must first identify verticals that contain at least one relevant item for a topic. Then, we must identify if those verticals have a high *vertical intent* for each of the queries.

3.4 Reusing Existing Relevance Assessments

Reusing existing relevance assessments is one of the most problematic areas when it comes to creating an aggregated search collection. As topics for one simulated vertical, typically do not overlap with topics from another, it is difficult to collect a large set of topics that span multiple verticals. To avoid a situation whereby whole verticals have not been assessed (for relevance) for particular topics, we have used one large collection of heterogeneous documents (ClueWeb). We do, however, use two different media type collections (e.g. image and video). These are of a different media type and the majority of ClueWeb topics do not have corresponding relevance assessments available within these collection. This is somewhat problematic. However, one way to minimise this impact is to manually judge the vertical intent of each query, and then perform relevance assessments only on collections that might be useful for that query. In this work, this incompleteness is minimised and we assume that the image and video verticals have a very low *vertical intent* for the queries originating from the ClueWeb query set.

4. AGGREGATED SEARCH TEST COLLECTION

In this section, we describe the actual construction of our test collection. We will describe our document classification approach (Section 4.1), topic identification method (Section 4.2) and statistics of the created test collection (Section 4.3), respectively.

4.1 Document Classification

Table 2 lists the collections and topics used in our aggregated collection. The majority of our documents come from the ClueWeb collection and, therefore, need to be classified into a specific vertical. We now describe the classification approach used for this web-based collection. Genre classification is not new in the community [13] and our aim is to demonstrate the feasibility of the approach (rather than thoroughly investigating how to improve genre classification). Our classification can be categorized into two steps:

1. Classifying the unlabeled documents using a *machine learning genre classifier.*

2. Increasing the accuracy of the classification of documents from Step 1 using existing vertical search engines.

[2] A topic-focused vertical may contain documents of various types, standard web pages, images, reviews, etc.

Table 2: Description of collections, topics and qrels used

Collection name	Type	Num of docs	Tracks (number of topics used)	Num of topics
ClueWeb09(B)	general web	50,220,423	TREC Web 2009-2010 (100) TREC Million Query 2009 (685)	785
ImageCLEF	image	670,439	ImageCLEF Photo Retrieval (178) ImageCLEF WikiMM (45)	223
TRECVID	video	1,253[3]	TRECVid Search (268)	268
Total		50,892,115		1,276

Table 3: Genre classification – Confusion Matrix

Vertical	Recipe	News	Book	Blog	Answ	Shop	Disc	Schol	Ref	Web
Recipe	**0.74**	0.00	0.02	0.04	0.01	0.00	0.01	0.00	0.00	**0.18**
News	0.00	**0.60**	0.00	**0.05**	0.01	0.01	0.00	0.00	0.00	**0.33**
Book	0.01	0.01	**0.53**	0.04	0.00	**0.10**	0.00	0.00	0.01	**0.30**
Blog	0.02	0.03	**0.06**	**0.57**	0.03	0.02	0.02	0.00	0.01	**0.24**
Answ	0.02	0.02	0.00	0.00	**0.72**	0.00	0.02	0.01	0.00	**0.21**
Shop	0.02	0.03	**0.07**	**0.07**	0.02	**0.58**	0.04	0.01	0.00	**0.16**
Disc	0.00	0.02	0.01	0.03	**0.05**	0.02	**0.78**	0.00	0.00	**0.09**
Schol	0.00	0.00	0.01	0.00	0.00	0.00	0.00	**0.81**	0.04	**0.14**
Ref	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	**0.89**	**0.08**
Web	0.00	0.03	0.01	0.04	0.00	**0.05**	0.02	0.00	0.00	**0.85**

For Step 1 (machine learning classification), we use a two-stage classifier. The first stage filters out pages that do not occur within the domains that are known to be associated with a vertical (e.g. www.allrecipes.com for the recipe vertical), and the second step uses a SVM classifier with other features of the page. In the first stage, we filter out the low-quality websites/domains for each vertical, since in our preliminary experiments those "poor quality" websites have been empirically shown to contribute a lot to the misclassification. We constructed this filter by using a website ranking service *Alexa*[4]. For each vertical, we manually find the top 100 ranked domains (e.g. www.allrecipes.com) that exist in our collection, and only web pages from those domains are candidate documents to be classified.

In the second stage, we use a multi-class SVM[5], which is known to perform well for this genre classification task, on these candidate documents. Both textual and structural features are used in the SVM. The textual features include the term-frequencies in various parts of the web document (URL, title, meta-data and full document), genre-based symbols (e.g. the "?" symbol contained in help documents), and named-entity features. Structural features include *html* tag frequency (e.g. list, form, image tag count), links (e.g. number of outlinks). The SVM classifier was trained using five-fold cross validation where the training and testing data consisted of approximatively two hundred manually labeled documents of each genre, and one thousand manually labelled documents of the "general web" genre. Note that all those training web pages exist in the collections we use (i.e. ClueWeb09 B). We have found that approximately 25% of the pages in the collection can be labeled as a non-web vertical by utilizing this two-stage classifier method.

To boost the accuracy of the classification provide in step 1, we submitted the "titles" of all of the classified pages to existing vertical search engines. Therefore, for each document (web page) from the classified set, we submit its "title" to all the corresponding state-of-the-art vertical search engines [6] by using the strict matching retrieval function (i.e. the exact title has to appear in the document.). Then, if the URL of the document (unique identifier) appears in the top 20 results (empirically shown to be sufficient), we relabelled the page with the corresponding vertical. We have found that 18.9% of the classified pages (i.e. those already classified into verticals) were re-classified using this method. This relabelling step (step 2), only affects about 18.9% of the initially labelled documents, but improves the accuracy of the classifier by over 10%.

After those two steps, all the documents in the ClueWeb B collection have been classified as either belonging to a vertical or the default "general web" vertical. Table 3 shows the confusion matrix for our genre classification (remembering that these results are generated from 200 manually labelled documents from each vertical using five-fold cross validation). The right-to-left diagonal shows the percentage of correctly classified documents of each type. We achieve an average accuracy of 70.7% (varying from 53% to 89%). Importantly, most mis-classifications are placed into one vertical (i.e. the "general web"). This is not surprising as "general web" is the default genre [10]. This should not affect our work as documents from the "general web", as is the case for major search engines, form the high majority of search results [10]. In addition, the overall misclassification remains low and it is comparable to state-of-the-art ([13, 8]). Furthermore, this classification reflects the real scenario of vertical creation on the web. Regardless, our experimental section (Section 5) will revisit the impact of this classification process.

4.2 Identifying Topics

In section 3.3 we defined vertical intent as being related to both *topical-relevance* and *vertical orientation*, and therefore, we must identify a set of topics that are associated to multiple verticals that contain both of these criteria.

[3] Each video is a mixture of many events/shots (normally more than one hundred) that can be further segmented.

[4] www.alexa.com

[5] http://svmlight.joachims.org/svm_multiclass.html

[6] We use Google News, Blog, Recipe, Shopping, Book, Answer, Discussion, Scholar, and Wiki.com for Reference Search.

4.2.1 Identifying topics associated with multiple topically relevant verticals

First, we wish to identify topics for which topically relevant documents exist in multiple verticals. This is not problematic for the ClueWeb B collection as we have automatically classified documents into different verticals, and therefore, relevant documents for a topic will be classified into different verticals. However, for the multi-media collections (i.e. image and video), we must identify topics that are statements of the same, or a very similar, information need, as those that exist in the ClueWeb topic set [7]. Therefore, we represent each topic as a weighted vector of its *title* terms (i.e. using $tf \cdot idf$) and the cosine similarity is then used to compare topics. Any pair of topics for which the cosine similarity is above a threshold γ are candidate topics. We then manually judged all candidate topics, using the *description* and the *narrative* fields. This yielded two video topics that had a similar information need to those in the ClueWeb B topic set.

4.2.2 Identifying topics with high vertical-orientation

To determine this topic set, first, we make an assumption that highly oriented verticals for a topic should contain above a certain *threshold* of relevant items. Therefore, to define a threshold for each vertical, we analyse a query log (i.e. the AOL log [11]). We identified a set of queries in this log that were highly orientated to a particular vertical (v_i). We identified queries that were highly orientated to a vertical by finding queries with an explicit vertical label (e.g. if the term "recipe" or "recipes" appeared in the query "pork chops recipe" we deemed it a recipe query). We also used the main sub-query, created by removing the vertical label (e.g. "pork chops"), as a highly orientated query. The vertical labels for each vertical are obtained from human annotation while the objective is to ensure a highly accurate classification of query's vertical intent, meanwhile covering a wide range of queries. For example, for "recipe" vertical, we used the term "recipe" and "recipes", whereas for "image" vertical, we used the term "image", "images", "img", "picture", "pictures", "photo", "photos", "pics". For classifying the clicked documents, we use the same approach on URLs of the documents [8], (e.g. if the term "recipe" or "recipes" occurs in the URL, we consider the document is associated with the "recipe" vertical). Then for each query, we then calculated the fraction of clicks that linked to pages in that vertical (v_i), compared to the number of total clicks for that query. These fractions were then averaged over all queries that were identified as highly *orientated* to a vertical. Given that a click is a noisy estimation of relevance, this fraction gives us an estimation of the number of relevant documents that must be in a vertical before the vertical is deemed highly-orientated. Finally, for our simulated collection, a vertical was deemed *highly-orientated* when it contained over this threshold of relevant documents. Using this process, 243 queries were found (in the ClueWeb topic sets) that had multiple vertical intents. We compared these vertical intents with those of two human annotators for a subset of queries and found a high degree (60%) of overlap[9].

4.3 Test Collection Composition

In this section, we describe the obtained collection. Table 4 shows the document statistics of our aggregated test collection in terms of verticals defined. In total, we have more than 50 million documents. General web documents are prevalent, thus mimicking aggregated search scenarios. A total of twelve verticals are simulated and many are common to the usual 16 verticals[10] found in current search engines (Google, Yahoo and Bing). We have simulated many of the verticals that are prevalent in web search engines and we have simulated more verticals than some search engines. The choice of those verticals are restricted to the collections we possess. However, as we have already mentioned, when a new collection (e.g. microblog) is available, we can use the same approach to add that collection and reuse available relevance judgments.

Table 4: Document statistics of the aggregated search collection (verticals)

Verticals	Recipe	News	Books	Blogs	Answer	Shopping
Ratio	0.3%	3.0%	1.4%	3.8%	0.6%	1.6%

Verticals	Discussion	Scholar	Reference	Image	Video	Web
Ratio	1.1%	0.1%	12.6 %	1.3%	0.0%	74.2%

Statistics relating to the final set of topics and qrels are shown in Table 5. In total, 320 topics are available for testing, which is larger than the minimum recommended number of topics in other areas of IR (i.e. 50) [17]. Also, 69.7% of the topics have two vertical *intents* and 6.2% of topics have three or more vertical intents. These statistics are comparable to those from [2], obtained from real data. The distribution of topics per vertical is shown in Table 6, which also conforms to that of [2].

Table 5: Statistics of topics and qrels

Statistics	number/ratio
number of topics	320
average rel docs per topic	26.0
average rel verticals per topic	1.83
ratio of topics with only "general web" intent	24.1%
ratio of topics with two vertical intents	69.7%
ratio of topics with more than two vertical intents	6.2%

5. EXPERIMENTS

We used a classifier to assign documents to verticals and, therefore, some documents may be incorrectly assigned to a vertical. We need to assess the impact of this. We now describe an experiment carried out to evaluate the effect that document misclassification has on our newly created test collection. We create different versions of the test collection,

[7] For example, the topic "find a shot of golden gate bridge" for the video collection and "golden gate bridge" for the web collection.

[8] We considered that terms were any maximal sequence of alphanumeric characters. For example, the URL "http://www.bbc.com/image" has four terms, "www", "bbc", "com" and "image".

[9] These were preliminary experiments and a full evaluation of this process is planned.

[10] 16 common verticals include News, Image, Video, Blog, Discussions, Answer, Reference, Maps, Books, Updates, Scholar, Shopping, Financial, Local Listings, Weather, Web.

Table 6: Percentage of topics assigned to each vertical

Verticals	recipe	news	books	blogs	answer	shop	disc	scho	ref	image	video	web only
percentage	3.8%	4.1%	3.8%	5.3%	4.7%	5.6%	0.3%	0.0%	54.7%	0.0%	0.6%	24.1%

Table 7: System ranking correlation for different misclassification rates

misclassified	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
correlation	0.96	0.95	0.94	0.91	0.91	0.93	0.89	0.84	0.80	0.86

where the only difference is that we intentionally mis-classify a certain percentage of the documents. Then, having created these modified collections, we investigate whether the ordering (based on an effectiveness metric) of a number of different aggregated search systems is preserved (or at least correlated) when run on these different versions of the test collection (i.e. collections that have various levels of misclassification). If the ordering of the systems is preserved, or highly correlated, we can conclude that the effect of misclassified documents on our created collection is minimal.

5.1 Simulating misclassified documents

We create several "misclassified" collections where we reassign the documents into incorrect verticals. For topics that have at least one vertical intent (excluding "general web"), for each specific vertical, we distribute a percentage of its documents uniformly into the remaining incorrect verticals. We iterate this process across all verticals (excluding "general web"). Therefore, according to different misclassification rates" (from 5% to 50%), we create a set of "misclassified" test collections. We also create another test collection (called "random") by randomly assigning documents into verticals. This corresponds to a random classification of documents with regard to the vertical contents.

5.2 Simulating aggregated search systems

We generate twleve (2 × 3 × 2) aggregated search systems by combining different variants of each component. For *vertical representation*, we use one complete and one incomplete representation that uses query-based sampling [3]. For *vertical selection*, we experiment with three existing methods, CORI, ReDDE and CRCS(e) ([4, 16, 14]). For simplicity, we select the top two ranked verticals (not including "general web") for each query. For *result presentation*, we implemented two retrieval systems, a "good" (i.e. BM25) and a "bad" (i.e. a simple cosine similarity with a tf term-weighting function) ranking function.

To test whether the right subset of documents from each vertical have been identified, we select the top 7 documents from the "general web" vertical, top 5 documents from the first ranked vertical , and top 3 from the second ranked vertical. We used α-NDCG [7] (with the default setting of α as 0.5) as a performance metric, and modelled verticals of high vertical intent as sub-topics.

5.3 Results

We ran the 12 systems on the the collections with various levels of misclassification (with five iterations for each level). We used a subset of topics that had at least two vertical intents. The average Spearman rank correlation between the performance of the systems on the different collections is shown in Table 7. We can see that in general there is a high correlation between the systems even if misclassification increases to 30%. A random classification of documents (not shown in Table 7) leads to a moderate correlation of 0.573.

6. CONCLUSIONS

We describe a method for creating a large-scale aggregated search test collections by reusing existing test collections. We have demonstrated that by identifying topics from existing test collections, a sufficient number (320) of topics with multiple vertical intents can be collected. In addition, through simulation we have showed that aggregated search approaches can be properly evaluated even if there are inherent misclassification within the verticals. Future work includes rigorous testing of evaluation measures that incorporate many aspects of aggregated search system performance.

7. REFERENCES

[1] J. Arguello, F. Diaz, J. Callan, and B. Carterette.: A methodology for evaluating aggregated search results. In ECIR11, pages 141-152, 2011.

[2] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo.: Sources of evidence for vertical selection. In SIGIR09, pages 315-322, 2009.

[3] J. Callan and M. Connell.: Query-based sampling of text databases. ACM Trans. Inf. Syst., 19:97-130, 2001.

[4] J. P. Callan, Z. Lu, and W. B. Croft.: Searching distributed collections with inference networks. In SIGIR95, pages 21-28, 1995.

[5] B. Carterette, J. Allan, and R. Sitaraman.: Minimal test collections for retrieval evaluation. In SIGIR06, pages 268-275, 2006.

[6] B. Carterette, E. Gabrilovich, V. Josifovski, and D. Metzler.: Measuring the reusability of test collections. In WSDM10, pages 231-240, 2010.

[7] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon.: Novelty and diversity in information retrieval evaluation. In SIGIR08, pages 659-666, 2008.

[8] I. Kanaris and E. Stamatatos.: Learning to recognize webpage genres. IP&M, 45:499-512, 2009.

[9] J. Kim and W. B. Croft. : Building pseudo-desktop collections. In SIGIR Workshop on the Future of IR Evaluation, pages 39-40, 2009.

[10] V. Murdock and M. Lalmas.: Workshop on Aggregated Search. SIGIR Forum, 42:80-83, 2008.

[11] G. Pass, A. Chowdhury, and C. Torgeson.: A picture of search. In InfoScale06, 2006.

[12] A. K. Ponnuswami, K. Pattabiraman, Q. Wu, R. Gilad-Bachrach, and T. Kanungo.: On composition of a federated web search result page: using online users to provide pairwise preference for heterogeneous verticals. In WSDM11, pages 715-724, 2011.

[13] M. Santini.: Automatic identification of genre in web pages. Phd Thesis, University of Brighton, Brighton (UK), 2007.

[14] M. Shokouhi.: Central-rank-based collection selection in uncooperative distributed information retrieval. In ECIR07, pages 160-172, 2007.

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

[16] L. Si and J. Callan.: Relevant document distribution estimation method for resource selection. In SIGIR03, pages 298-305, 2003.

[17] J. Zobel.: How reliable are the results of large-scale information retrieval experiments? In SIGIR98, pages 307-314, 1998.