Aligning Vertical Collection Relevance with User Intent

Ke Zhou
Yahoo Labs
London, U.K.
zhouke.nlp@gmail.com

Thomas Demeester
Ghent University
Ghent, Belgium
tdmeeste@intec.ugent.be

Dong Nguyen
University of Twente
Enschede, Netherlands
d.nguyen@utwente.nl

Djoerd Hiemstra
University of Twente
Enschede, Netherlands
d.hiemstra@utwente.nl

Dolf Trieschnigg
University of Twente
Enschede, Netherlands
d.trieschnigg@utwente.nl

ABSTRACT

Selecting and aggregating different types of content from multiple vertical search engines is becoming popular in web search. The *user vertical intent*, the verticals the user expects to be relevant for a particular information need, might not correspond to the *vertical collection relevance*, the verticals containing the most relevant content. In this work we propose different approaches to define the set of relevant verticals based on document judgments. We correlate the collection-based relevant verticals obtained from these approaches to the real user vertical intent, and show that they can be aligned relatively well. The set of relevant verticals defined by those approaches could therefore serve as an approximate but reliable ground-truth for evaluating vertical selection, avoiding the need for collecting explicit user vertical intent, and vice versa.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]
Keywords: aggregated search, federated search, vertical relevance, evaluation, user intent

1. INTRODUCTION

Due to the increasing diversity of data on the web, most of the current search engines aggregate search results from a set of vertical search engines (e.g. News, Images). This search paradigm is called aggregated search [1]. The verticals are associated with content dedicated to either a topic (e.g. "sports"), a media type (e.g. "images"), or a genre (e.g. "news"). For a given user information need, only a subset of verticals will potentially provide the most relevant results to satisfy it. For example, relevant verticals for a query such as "flowers" might include "image" and "encyclopedia" verticals. Vertical selection (VS), a subtask, deals with selecting a subset of the most relevant verticals to a given user information need and improves the search effectiveness while reducing the load of querying a large set of multiple verticals.

The relevance of a vertical could depend on the relevance of the documents within the vertical collection [3, 7] and on the user's intent (orientation) to the vertical [9]. For evaluation purposes, from the *collection* perspective, Gravano et al. [3] assumed that any collection (vertical) with at least one relevant document for a query is relevant. Powell et al. [7] refined and formalized this notion by assuming that the relevance level of the collection (vertical) depends on the number or relevant documents within. In this work, we call this the *collection-based vertical relevance*. On the other hand, from the *user intent* perspective, researchers [6, 9] found that user orientation (intent), or how oriented each vertical is to a user's information need (i.e., expectation), also plays an important role in user preference of the aggregated search page. We refer to this as the *user vertical intent*. Most of the previous work either assumes that the relevance of the vertical solely depends on the collection (i.e., its recall of relevant documents) or the user intent (the user's orientation to issue the query to the given vertical).

Although previous work [9] has shown that *user vertical intent* and *result relevance* are both correlated for influencing user experience, it fails to connect both criteria within the context of evaluation in this area. The key question is whether we could align the collection-based vertical relevance with the user vertical intent for evaluation purposes. This can be further split up into two sub-questions:

- Can the vertical relevance be derived from document relevance judgments and therefore ranked similarly to the user vertical intent (orientation)?
- Can we appropriately threshold the derived vertical rankings and ultimately align them with the binary vertical selection decision made by the users?

In this paper, we propose and test different approaches to derive the collection-based vertical relevance based on document relevance judgments as are widely used in the IR evaluation community (e.g., TREC). The approaches differ in how they quantify the relevance of a vertical and how the ultimate set of relevant verticals is derived. By conducting user studies to collect the user vertical intent, we compare those approaches on deriving collection-based vertical relevance and investigate which approach best aligns with the actual user intent.

The contributions of this paper are twofold: **(i)**. We extensively study different approaches to derive collection-based vertical relevance in the context of over a hundred heterogeneous resources (search engines). The scale and diversity

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661941.

Table 1: Summary of this work.

Tasks	ranking of verticals		set of relevant verticals	
Variants	(1). Resource Relevance	(2). Aggregation	(3). Vertical Dependency	(4). Thresholding Criterion
Collection	a. **K** (Key relevant doc recall) b. **G** (Graded precision of docs)	a. **MR** (Maximal Resource) b. **AR** (Average of Resource)	a. **D** (Dependent) b. **I** (Independent)	a. **I** (Individual vertical utility) b. **O** (Overall vertical set utility)
User Study	Fraction of majority user preference of each vertical over “General Web”		User type: 1. risk-seeking (diversity); 2. risk-medium; 3. risk-averse (relevance)	
Evaluation	Spearman Correlation		Precision, Recall and F-measure	

of the resources used has not been studied previously for our task. **(ii)**. We conduct a user study to verify that the collection-based vertical relevance derived from document judgments can be aligned with the user vertical intent. This is novel to verify that the Cranfield evaluation paradigm used in TREC could also be useful to line up with user experience [5] for evaluation in the *heterogeneous environment*.

The main elements of this paper are summarized in Table 1. We first describe different approaches to obtain both the vertical ranking and a set of relevant verticals using the collection-based document judgments (Sec. 2). We then conduct a user study (Sec. 3) to obtain the vertical ranking and relevant vertical sets from the user (as the ground-truths). Finally, we evaluate different approaches proposed in Sec. 2 and study how well they can be aligned with the user intent (Sec. 4), after which the paper is concluded (Sec. 5).

2. VERTICAL COLLECTION RELEVANCE

Formally, given a set of verticals $V = \{v_1, v_2, ...v_n\}$, the collection-based vertical relevance I_t^C derived from the collection C for topic t is represented by a weighted vector $I_t^C = \{i_1, i_2, ...i_n\}$, where each value i_k indicates the relevance of the given vertical v_k to topic t. A vertical could contain multiple resources (search engines). For example, an “image” vertical could contain resources such as Flickr and Picasa. Therefore, each vertical v_i consists of a set of resources $v_i = \{r_1, r_2, ...r_m\}$ while each resource r_j consists of a set of documents $r_j = \{d_1, d_2, ...d_k\}$. Given all the relevance judgments $rel(d_l, t)$ between any document d_l and a topic t, we aim to derive I_t^C.

Ultimately, given the collection-based vertical relevance I_t^C, we aim to threshold it in order to obtain the final binary verticle relevance vector $S_t^C = \{s_1, s_2, ...s_n\}$ where each value s_k is either 1 (indicating corresponding vertical v_k is relevant and should be selected in VS) or 0 (indicating irrelevant and should not be selected).

2.1 Approaches

We describe approaches to derive the vertical ranking, followed by methods to infer the set of relevant verticals.

2.1.1 Vertical Ranking

The strategies to derive the collection-based vertical relevance I_t^C (i.e., the vertical score) from the document relevance judgments $rel(d_l, t)$ vary in two aspects: (1). **(Resource Relevance)** the way to estimate the relevance of each resource within a given vertical; and (2). **(Vertical Relevance Aggregation)** the way to aggregate scores of the resources within the vertical to derive the vertical score.

Following previous work [2, 3], we propose two approaches to estimate **resource relevance**: (a). **K** (Key): using the recall of “key” (most relevant) documents in the resource and (b). **G** (Graded precision): the graded precision of documents in the resource [4]. The **K** approach is similar to the assumption made in Gravano et al [3] and using “key” is to reflect the relevance of the resource to return the most rel- evant results that maintain high impacts on user search experience [5]. The **G** approach is following the evaluation setup [2] made in the TREC FedWeb track 2013[1] and the essential idea is to characterize the effectiveness of each resource to “recall” relevant documents in a similar fashion as in previous work [7] when graded relevance judgments are available. Then given the estimated resource relevance scores, we test two ways to **aggregate** those scores in order to obtain the vertical scores (rankings) I_t^C: (a). **(MR)** Maximal Resource score and (b). **(AR)** Average Resource score. The **MR** approach reflects most of current web search setting that one vertical solely contains one best performing resource while the **AR** approach represents the averaged vertical performance.

By combining the different *resource relevance* and *aggregation* methods, we obtain four approaches to quantify I_t^C: **KMR**, **KAR**, **GMR**, **GAR**. Since we could also apply the same technique to the whole vertical (rather than resource), we propose another approach **GV** by using graded precision on all the documents within the entire vertical[2].

2.1.2 Relevant Vertical Set

To infer the set of relevant verticals S_t^C from the obtained collection-based vertical relevance I_t^C, we argue that the strategies could vary in two aspects: (3). (**Vertical Dependency**): assumption of whether vertical relevance is dependent on each other; and (4). (**Thresholding Criterion**): assumption of which is the criterion of thresholding. For **vertical dependency**, we tested both assumptions. By assuming (a). **(D)** Dependent, we normalize the vertical scores across all verticals following previous work [1]. By assuming (b). **(I)** Independent, we simply use the original vertical scores I_t^C.

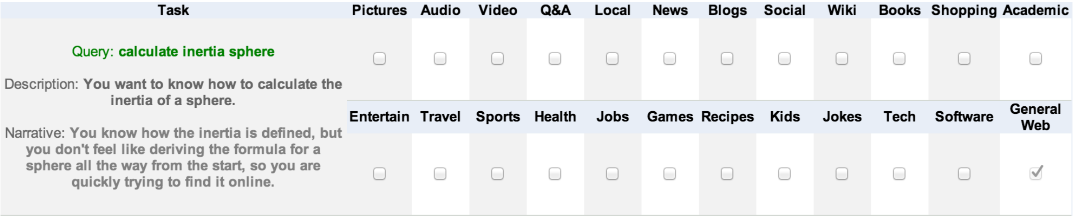
For **thresholding criterion**, we tested two different approaches. The differences between them are the criterion that the thresholding is based on: (a). **(I)** Individual vertical score: the individual vertical relevance scores; (b). **(O)** Overall relevance of the vertical set. The **I** approach basically assumes that the individual vertical requires a certain relevance to remain in the relevant vertical set while the **O** approach assumes that the relevant vertical set is required to maintain a certain percentage of relevance of the whole vertical set. By combining *vertical dependency* and *thresholding criterion*, we obtain four different approaches to infer S_t^C from I_t^C: **DI**, **DO**, **II**, **IO**.

2.2 FedWeb’13 Data

In this study, we use the TREC 2013 FedWeb track data [2]. The dataset contains 50 test topics and 157 crawled

[1] https://sites.google.com/site/trecfedweb/2013-track.

[2] We do not propose **KV** approach since practically, it outputs the same vertical ranking to **KAR** approach.



Task	Pictures	Audio	Video	Q&A	Local	News	Blogs	Social	Wiki	Books	Shopping	Academic
Query: **calculate inertia sphere** Description: **You want to know how to calculate the inertia of a sphere.** Narrative: **You know how the inertia is defined, but you don't feel like deriving the formula for a sphere all the way from the start, so you are quickly trying to find it online.**	□	□	□	□	□	□	□	□	□	□	□	□
	Entertain	**Travel**	**Sports**	**Health**	**Jobs**	**Games**	**Recipes**	**Kids**	**Jokes**	**Tech**	**Software**	**General Web**
	□	□	□	□	□	□	□	□	□	□	□	☑

Figure 1: An examplar task of the user intent study.

resources. It also categorizes the resources into different verticals and provides a set of 24 verticals, as shown in Table 2. Each vertical consists of a set of resources (search engines). For each resource, the top 10 retrieved document results are returned. The relevance judgments are made on each document with five graded relevance levels.

The 50 test topics were chosen in such a way to avoid a strong bias towards general web search engines. For the most important verticals (in terms of number or size of resources, e.g. Video, Blogs), many topics provide a significant number of relevant results. In addition, at least a few topics targeting smaller verticals (e.g., Recipes) are also selected.

Table 2: 24 verticals used in FedWeb'13: a vertical consists of a set of resources (search engines), each retrieving one unique type of documents.

Vertical	Document	Resource Count	Type
Pictures	online pictures	13	media
Audio	online audios	6	
Video	online videos	14	
Q&A	answers to questions	7	genre
Local	local information pages	1	
News	news articles	15	
Blogs	blog articles	4	
Social	social network pages	3	
Encyclopedia	encyclopedic entries	5	
Books	book review page	5	
Shopping	product shopping page	9	
Academic	research technical report	18	
Entertainment	entertainment pages	4	topic
Travel	travel pages	2	
Sports	sports pages	9	
Health	health related pages	12	
Jobs	job posts	5	
Games	electronic game pages	6	
Recipes	recipe page	5	
Kids	cartoon pages	10	
Jokes	joke threads	2	
Tech	technology pages	8	
Software	software downloading pages	3	
General Web	standard web pages	6	

3. VERTICAL USER INTENT STUDY

Given a set of verticals $V = \{v_1, v_2, ...v_n\}$, the vertical user intent I_t^U for topic t is represented by a weighted vector $I_t^U = \{i_1, i_2, ...i_n\}$, where each value i_k indicates the relevance of the given vertical v_k to topic t. To obtain I_t^U, we conducted a user study, asking assessors $U = \{u_1, u_2, ..., u_m\}$ to make binary decisions over all verticals V: $A = \{a_1, a_2, ..., a_n\}$. Therefore, we have a $m \times n$ matrix M_t for topic t. We aim to derive I_t^U by aggregating M_t and ultimately obtain a binary vector indicating the set of relevant verticals $S_t^U = \{s_1, s_2, ...s_n\}$ where each value s_k is either 1 or 0.

We conducted this user study following previous work on gathering user vertical intent [9]. Basically, two assumptions were made in guiding the assessment. Firstly, instead of asking assessors to associate an absolute score to each vertical, we asked them to make *pairwise preference assessments*, comparing each vertical in turn to the reference "general web" vertical (i.e. "is adding results from this vertical likely to improve the quality of the *ten blue links*?"). Secondly, instead of providing actual vertical results to the assessors, we only provided the *vertical names* (with a description of their characteristics presented before their assessments). Although this may not be ideal from an end-user perspective (as different assessors might have different views on the perceived usefulness of a vertical, especially as the vertical items are hidden), this assumption helped to lower the assessment burden, and yet reflects the perceived vertical user intent (orientation). We used the same FedWeb'13 data as described in Sec. 2.2. Most of the assessors are university students who were recruited to participate via a web interface. To eliminate order bias, we randomized all topics into a set of pages (with five topics per page) and provided each assessor the option to assess as many pages as he/she wished. A screenshot of one examplar task is presented in Figure 1.

In total, we collected 20 assessment sessions (i.e., assessors) with a total of 845 assessments[3]. The average number of relevant verticals per topic and per session is 2.64, with a standard deviation of 1.28. Similar to previous findings [9], the mean of inter-annotator Fleiss' Kappa is *moderate* (0.48), showing that assessors might have different preferences over the relevance of verticals, despite the clearly described query information need (as seen from the description and narrative shown in Figure 1).

To derive I_t^U, we use the fraction of majority user preferences for each vertical v_k over "General Web" as the vertical score i_k. To further obtain S_t^U, we threshold the majority user preference for each vertical i_k in I_t^U. It has been shown in our data (moderate inter-annotator Fleiss' Kappa agreement) and previous work [8] that the user's preferred number of verticals varies significantly and different users tend to have different risk-levels. By thresholding 30%, 60% and 90% of majority user preference, we obtain three different types of S_t^U, representing three types of users respectively: *risk-seeking*, *risk-medium* and *risk-averse*. The *risk-seeking* users prefer *diversity* of verticals presented (with a mean of 3.08 relevant verticals) while the *risk-averse* users are more careful when selecting verticals (with a mean of 0.52 relevant verticals): they only select verticals (as relevant) when highly confident (large fraction of user's preferences). The *risk-medium* is an average user, with a mean of 1.68 relevant verticals (following a similar distribution as shown in [1]).

4. EVALUATION

We evaluate the alignment between collection and user, on both vertical rankings and ultimate relevant vertical sets.

[3] We have made this publicly available at http://bit.ly/VJQ53B.

Table 4: Precision, Recall and F-measure of the set of relevant verticals using collection-based vertical relevance I^C from GMR approach against user vertical intent I^U with different risk-level.

User Intent	Risk-seeking (diversity)				Risk-medium				Risk-averse (relevance)			
Approaches	DI	DO	II	IO	DI	DO	II	IO	DI	DO	II	IO
Precision	0.264	0.289	**0.338**	0.271	0.303	0.330	**0.391**	0.327	0.339	0.377	**0.421**	0.381
Recall	**0.829**	0.803	0.796	0.782	**0.552**	0.523	0.516	0.520	**0.426**	0.398	0.401	0.394
F-measure	0.380	0.406	**0.446**	0.384	0.367	0.387	**0.413**	0.383	0.357	0.374	**0.379**	0.373

4.1 Vertical Ranking

Given the collection-based vertical relevance I_t^C derived from document relevance judgments and user vertical intent I_t^U obtained from user preference judgments, we evaluate whether they align with each other. We aim to evaluate five different approaches of utilizing relevance judgments for ranking verticals, as described in Sec 2.1.1. Specifically, we utilize the nonparametric Spearman Rank Correlation Coefficient as our main metric to measure the correlation between a collection-based vertical ranking and a user-based one. Since we are more concerned with highly ranked verticals (potentially relevant), we also investigate whether there are overlaps between the top-3 and top-5 ranked verticals in the collection-based and user-based rankings. The evaluation results of different approaches are shown in Table 3.

Several trends can be observed. Firstly, all the collection-based approaches have a moderate correlation (0.6-0.7) with the user-based vertical ranking. We also study whether this correlation is statistically significant (against random) by performing a permutation test. We found that the correlation for all the five approaches are statistically significant (with $p < 0.05$). Note that the performance difference between different approaches is marginal while the approaches using the graded precision metric outperforms the others. Secondly, we observed that there tends to be some overlap between the top ranked verticals from both collection-based and user-based vertical rankings, albeit moderate (0.4-0.5). However, it is interesting to see that when using simple metrics on document-based relevance judgements, around half of the top-ranked verticals are aligned with the user intent.

In summary, our experiments suggest that collection-based vertical relevance can be utilized as an approximate surrogate for measuring user's vertical intent, and vice versa.

4.2 Vertical Relevant Set

We study whether the obtained set of relevant verticals after thresholding is aligned with the ones derived from the user perspectives. For simplicity, we only present results on thresholding with one collection-based vertical ranking approach **GMR** (Graded precision of Maximal Resource) since we found similar results across all those different approaches.

As we have mentioned, we defined three types of ground-truths, representing three different types of users: risk-seeking users prefer a large set of diverse verticals, while the risk-averse users prefer selecting verticals only when they are most relevant, and risk-medium are in between. We test different thresholding approaches for these user settings.

Table 3: Spearman correlation and overlap of top-k verticals between vertical rankings from collection-based vertical relevance I^C and user intent I^U.

Approaches	KMR	KAR	GMR	GAR	GV
Correlation	0.656	0.659	0.664	0.689	0.671
Overlap (5)	0.496	0.492	0.516	0.532	0.508
Overlap (3)	0.340	0.340	0.327	0.400	0.353

For each thresholding approach, its numerical threshold was determined based on iterative data analysis, such that the maximum number of relevant verticals for any test topic could not exceed five. In addition, since almost all the FedWeb'13 test topics target verticals, we also make sure that at least 80% of the topics have at least one relevant vertical using the selected threshold.

The results are shown in Table 4. We observe similar performance trends for different thresholding approaches under different user settings (risk-level). **II** (Independent Individual) thresholding approach performs best in terms of precision and F-measure while **DI** (Dependent Individual) thresholding approach generally would achieve better recall. Generally, an F-measure of around 0.4 could be achieved by mapping the estimated collection-based relevant vertical set with the users' relevant (intended) vertical set. Although not particularly high, this still shows that vertical collection relevance could be aligned relatively well with users' vertical intent and therefore this could serve as a surrogate of ground-truths for evaluating vertical selection.

5. CONCLUSIONS

In this paper, we propose a set of different approaches to utilize document judgments to derive the set of relevant verticals. We evaluate the effectiveness of those approaches by correlating with the user vertical intent obtained from a user study. We found that collection-based vertical relevance can be aligned relatively well with users' vertical intent. This implies that we could reliably use document relevance judgments to evaluate vertical selection for capturing user intent in heterogeneous federated web search, and vice versa.

6. REFERENCES

[1] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *SIGIR '09*.

[2] T. Demeester, D. Trieschnigg, D. Nguyen, D. Hiemstra. Overview of the TREC 2013 Federated Web Search Track. In *TREC'13*.

[3] L. Gravano, H. Garcia-Molina and A. Tomasic. Precision and recall of GlOSS estimators for database discovery. In IEEE Parallel and Distributed Information Systems, 1994.

[4] J. Kekalainen and J. Kalervo. Using graded relevance assessments in IR evaluation. In *JASIST'02*.

[5] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *SIGIR'10*.

[6] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. Factors affecting click-through behavior in aggregated search interfaces. In *CIKM '10*.

[7] A. Powell and J. French. Comparing the performance of collection selection algorithms. In *ACM TOIS'03*.

[8] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating reward and risk for vertical selection. In *CIKM'12*.

[9] K. Zhou, R. Cummins, M. Lalmas and J.M. Jose. Which Vertical Search Engines are Relevant? In *WWW'13*.