Deriving Vertical Orientation from Anchor-Based Assessments

Aleksandr Chuklin[*] Maarten de Rijke

University of Amsterdam, Amsterdam, The Netherlands
a.chuklin, derijke@uva.nl

ABSTRACT

Modern search engine result pages are becoming more and more heterogeneous. This is mostly achieved by adding a special vertical results on top of traditional "general web" results. These results usually come from special sources (verticals) and the choice of verticals is different for different queries.

Vertical orientation is an important value that quantifies the user's need of having results from a particular vertical (e.g., News, Blogs, Video) on a search engine result page (SERP). It is used not just for selecting relevant verticals and positioning them on a SERP, but also for building vertical-aware click models and evaluating aggregated search performance.

In this paper we propose a way to accurately estimate vertical orientation from a limited amount of human assessments. We describe an intuitive procedure of collecting human ratings and show how these ratings can be converted to real-valued estimates of vertical orientation and further extrapolated to unseen queries with the help of machine learning.

Categories and Subject Descriptors

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

1. INTRODUCTION

Heterogeneity of a modern search engine result page is primarily achieved by adding so-called vertical results to a set of "general web" results. This paradigm is called aggregated or vertical search (see, e.g., [2]) and is widely adopted by the major search engines.[1]

While it is often said that vertical results are *added* to the general web results, in fact, adding such results pushes other documents lower in the ranking, so that users need to make additional effort to see these documents. In some cases, adding vertical results even reduces the number of general web results shown on a SERP, making them accessible only after clicking a pagination button.[2] This competition between vertical and general web results makes it important to accurately estimate their relative importance in order to make sound decisions about the placement of vertical results.

Another observation is that we need to break ties between different verticals to decide which one should be shown higher in the ranking. It is also worth noting that at least two search engines (Google and Yandex) appear to allow verticals to be present at any ranking position, not just the top, middle and bottom of the page as suggested by Ponnuswami et al. [6]. This supports the need for a real-valued vertical orientation or at least a graded value with a higher granularity.

The rest of the paper is organized as follows. In Section 2 we briefly discuss related work. In Section 3 we detail our method for collecting vertical assessments and converting them into orientation values. Section 4 is dedicated to a preliminary experiment we conducted to support our method. We conclude with a discussion in Section 5.

2. RELATED WORK

Research on result search diversification [1] as well as federated search [7] has shaped what is now called *aggregated search* and is adopted by all major search engines.

The first problem that arises when building an aggregated search system is selecting relevant verticals for a particular query [2]. Then there is the problem of item selection within a vertical (which is really just a ranking problem) and, finally, the problem of result presentation in a single aggregated SERP [6].

The term "orientation" which is central to our work was first introduced by Sushmita et al. [8]. Later Zhou et al. [10, 11] did a detailed analysis of different aspects of the vertical relevance and its relation to the vertical orientation and collection-based relevance.

There are at least three areas where vertical orientation is being used. First are the vertical selection algorithms for aggregated search (e.g., [2]). Second are the intent-aware metrics [1] that are often used for evaluating aggregated search [9]. A third application are the vertical-aware click models [4] that use orientation values for more accurate click prediction.

3. METHOD

First, we describe and motivate the procedure that we use to collect assessments. Then we move on to the usage of

[*] Part of the work was done while the first author was at Yandex Russia; now at Google Switzerland.

[1] We did a manual inspection of the SERPs returned by Bing, Google and Yandex.

[2] As of November 2014 at least for some queries submitted to Google and Yandex the total number of general web results is less then ten.

HIA'15, February 6, 2015, Shanghai, China.

the assessments and show how they can be used to obtain accurate and reusable estimates of the vertical orientation.

3.1 Collecting Assessments

As detailed in [10], when one opts for using human judges to rate verticals, there are essentially two ways of doing that. The first approach, called *inter-dependent assessment* requires presenting all verticals to raters and asking them to rank the verticals or compare to each other. Presenting more verticals at once poses a substantial cognitive load on a rater, while running multiple pairwise comparison increases the cost of collecting data.

This brings us to the second approach, which is called *anchor-based assessment*. In this method there is a single reference SERP (anchor) for each query and the raters are asked to decide on whether it will benefit from adding a particular vertical. This anchor SERP is usually a list of general web documents, but can also be a current production ranking of a search company that is considering whether to add a new vertical. As was shown by Zhou et al. [10] the ratings collected this way show moderate correlation with inter-dependent assessments.

In our experiments we allow raters to see the anchor SERP, but do not show them the vertical documents, we just show a general description of the vertical. There are multiple motivations for this design choice. Firstly, we want to separate out user's vertical orientation from collection-based relevance. Even though these two notions appear to be correlated (see [11]), it is still quite likely that there is a need in the vertical which is currently poorly satisfied by the top-ranked results from the vertical. This may happen, for instance, if the vertical-specific ranking is still actively being improved at the time of collecting assessments. Secondly, as was shown by Zhou et al. [10], when assessors are allowed to see vertical results, they show higher disagreement with each other, which suggests that this additional information makes the task more difficult and ambiguous for raters.

In addition, we provide assessors a way to give us graded feedback about the vertical relevance. We do so in a very explicit and natural way, so the raters often prefer it to the binary "show / no-show" question. More concretely, we ask them the following question:

- Do you think that the current SERP would benefit from adding results from the vertical X? If yes, where should they be located? **ToP/MoP/BoP/NS** (Top of the Page, Middle of the Page, Bottom of the Page or No Show).

On the one hand, having such options saves the raters from having to make hard decisions like distinguishing marginally relevant vertical from the irrelevant ones. On the other hand, it provides them with a clearly interpretable options which we can later convert to orientation values which by itself are less clear to the raters.

Overall, our rating process is set up in such a way that it mitigates the risk of incorrect interpretation of the assessment guidelines and leaves us with only one source of potential disagreement, namely different level of vertical tolerance for different raters.

3.2 Deriving Vertical Orientation

Once we have obtained labels from the raters we want to convert them into real-valued orientation values. The idea is that the aggregated search system based on these ratings should roughly meet the expectation of the raters. For example, if a rater said that for some query the vertical should be in the middle of the page (**MoP**), our aggregated search system should place this vertical close to the middle of the page. Note, that the aggregated search system only sees the orientation values for the verticals and does not have access to the assessors' labels, i.e., it does not know for sure where on the SERP the raters wanted to see this particular vertical.

In other words, we need to find a mapping from the ratings to the orientation values $f : \{\mathbf{ToP}, \mathbf{MoP}, \mathbf{BoP}, \mathbf{NS}\} \to [0, 1]$. This mapping, when used as part of the aggregated search system, should yield a ranking that minimizes the difference from the ranking advised by the raters. We refer to the latter as the *recommended SERP*.

As an aggregated search system, we use a system that does a greedy optimization of the ERR-IA metric [1]:

$$\text{ERR-IA}(q) = \sum_{v \in V} P(v|q)\text{ERR}(q|v), \tag{1}$$

where $P(v|q)$ is a vertical orientation for the vertical v and query q, and $\text{ERR}(q|v)$ is an expected reciprocal rank [3] computed for the vertical v with vertical relevance instead of topical relevance (here we assume that "general web" is one of the verticals V).

Now we have to define a function $\rho(S_1, S_2|v)$ that would give us the distance (discrepancy) between two SERPs S_1 and S_2 in terms of how they place the vertical documents of the vertical v, i.e., how we penalize ourselves for not placing the vertical v where the raters said. Once we fix this function we can do a grid search to find an optimal mapping ($f : \{\mathbf{ToP}, \mathbf{MoP}, \mathbf{BoP}, \mathbf{NS}\} \to [0, 1]$) from assessment grades to orientation values.

Our choice of function is the following:

$$\rho(S_1, S_2|v) = \frac{|rank(v|S_1) - rank(v|S_2)|}{\min\left(rank(v|S_1), rank(v|S_2)\right)}, \tag{2}$$

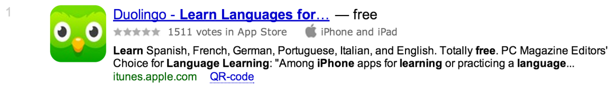
where $rank(v|S)$ is the rank of the vertical v on a SERP S if it is present there or some number bigger than S otherwise. This number depends on how much more we want to penalize for showing vertical results when they should not be shown or vice-versa. In our work we set it to $2|S|$ (twice the SERP size, where the size of the SERP is defined as the number of results it contains), effectively making a no-show equivalent to showing the vertical at the bottom of the second result page, which is accessible only after a pagination button click.

The idea behind such a distance function is that we want to penalize more for misplacing verticals higher in the ranking and we use an inverse rank of the result as a penalty decay.

All in all, the procedure is the following:

1. Using human raters collect vertical assessment labels $a(v, q)$ for a set of queries $q \in Q$ and the vertical v; $a(v, q) \in \{\mathbf{ToP}, \mathbf{MoP}, \mathbf{BoP}, \mathbf{NS}\}$.

2. For each mapping from assessment labels to orientation values $a(v, q) \mapsto P(v|q)$ (4 numbers), compute the per query average distance (2) between S and S' where S is the *recommended SERP* with the vertical v inserted at $a(v, q)$ and S' is the SERP returned by the aggregated search system that uses $P(v|q)$ as vertical orientation.



1

Duolingo - Learn Languages for… — free

★★★★★ 1511 votes in App Store iPhone and iPad

Learn Spanish, French, German, Portuguese, Italian, and English. Totally **free**. PC Magazine Editors' Choice for **Language Learning**: "Among **iPhone** apps for **learning** or practicing a **language**...

itunes.apple.com QR-code

Figure 1: A vertical results coming from the Mobile Applications vertical.

3. Return the mapping that yields the smallest average distance $\rho(S, S')$.

4. PRELIMINARY EXPERIMENT

The original motivation for this problem came from the aggregated search system that was used at Yandex as of 2012 and required real-valued input for the vertical orientation. A new vertical of Mobile Applications (see Figure 1) was planned to be launched and we wanted to assign adequate weighting to this vertical in order to properly blend it with the other verticals as well as general web results.

We followed a procedure similar to the one described above to obtain the mapping from assessments to vertical orientation values and further extrapolated it to the unseen queries using machine learning (see, e.g., [2]). The size of the training set was 800 queries where half of the labels were negative (**NS**); the feature set consisted of 24 features. The resulting aggregated search system was compared to the production system (that did not have this new vertical) using A/B-testing [5]. After two weeks of comparison with 1% of the users the new system was shown to have a statistically significantly[3] lower abandonment rate than the control group.

This experiment was just a first verification of our method and, although it was shown to have practical value, it still demands a more comprehensive evaluation.

5. DISCUSSION AND CONCLUSION

In this short paper we discussed a way to collect human assessments for vertical orientation and suggested a way to convert these graded labels to the real-valued orientation values. We performed a preliminary experiment that proved the legitimacy of our approach. As the next step we plan to compare our method to simple baseline mappings using some publicly available aggregated search dataset.

Since click modeling was mentioned as one of the applications that requires accurate vertical orientation values, we plan to compare performance of the vertical-aware click models [4] and show that our method improves the predictive power of the models.

Another possible evaluation approach is to evaluate vertical orientation values directly, e.g., by looking at the query flow and see how often users click on a particular vertical or explicitly ask for it. One may also employ a bigger number of human raters and compare their aggregated binary vertical preference to the orientation values obtained using our method.

Some questions also require further research, such as effects of visual saliency [8] or result inter-connection and redundancy [10]. To a certain extent we eliminated the first problem by not showing the results to the raters and the second problem is not specific to aggregated search. We do believe, however, that both questions need to be studied in the context of aggregated search in more detail.

Acknowledgements. This research was partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.930, the Center for Creation, Content and Technology (CCCT), the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.

REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, 2009.

[2] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *SIGIR*, 2009.

[3] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, 2009.

[4] A. Chuklin, K. Zhou, A. Schuth, F. Sietsma, and M. de Rijke. Evaluating intuitiveness of vertical-aware click models. In *SIGIR*, 2014.

[5] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1), July 2008.

[6] 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 *WSDM*, 2011.

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

[8] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. Factors affecting click-through behavior in aggregated search interfaces. In *CIKM*, 2010.

[9] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating aggregated search pages. In *SIGIR*, 2012.

[10] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Which vertical search engines are relevant? Understanding vertical relevance assessments for web queries. In *WWW*, 2013.

[11] K. Zhou, T. Demeester, D. Nguyen, D. Hiemstra, and D. Trieschnigg. Aligning vertical collection relevance with user intent. In *CIKM*, 2014.

[3] A Mann-Whitney U test with $\alpha = 0.01$ was used.