Assessing and Predicting Vertical Intent for Web Queries

Ke Zhou[1], Ronan Cummins[2], Martin Halvey[1], Mounia Lalmas[3], and Joemon Jose[1]

[1] University of Glasgow, Glasgow, UK
{zhouke, halvey, jj}@dcs.gla.ac.uk
[2] National University of Ireland, Galway, Ireland
ronan.cummins@nuigalway.ie
[3] Yahoo! Research Barcelona, Barcelona, Spain
mounia@acm.org

Abstract. Aggregating search results from a variety of heterogeneous sources, i.e. so-called verticals [1], such as news, image, video and blog, into a single interface has become a popular paradigm in web search. In this paper, we present the results of a user study that collected more than 1,500 assessments of vertical intent over 320 web topics. Firstly, we show that users prefer diverse vertical content for many queries and that the level of inter-assessor agreement for the task is *fair* [2]. Secondly, we propose a methodology to predict the vertical intent of a query using a search engine log by exploiting *click-through* data, and show that it outperforms *traditional* approaches.

1 Introduction

With the emergence of numerous vertical search engines, it is becoming popular to present results from a set of specific verticals dispersed throughout the standard "general web" results (e.g. adding image results to the *ten blue links* for the query "pictures of flowers"). In this paper, the concept of *vertical intent* is defined to reflect the perceived usefulness of a vertical from the user's perspective, without regard to the quality of the vertical results (e.g. an image vertical will still be one intent for the query "pictures of flowers" even if all the results from the vertical are irrelevant). The contributions of this paper are two-fold: (1) we study the agreement between user judgments of vertical intent, and (2) we develop a novel approach to predict vertical intent using query-logs.

2 Identifying Vertical Intent

Given a set of verticals $V = \{v_1, v_2, ...v_n\}$, the vertical intent I_t for topic t is represented by a weighted vector $I_t = \{i_1, i_2, ...i_n\}$, where each value i_k indicates the importance of the given vertical v_k to topic t. Commonly, I_t is treated as a binary vector [1], where each element indicates whether or not the given

vertical is *intended* by the query. To obtain I_t, we conducted a user study, asking assessors $U = \{u_1, u_2, ...u_m\}$ to make binary decision 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 make two assumptions in guiding the assessment. First, instead of asking assessors to associate an absolute score to each vertical, we ask 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 provide the *vertical names* (with a description of their characteristics). 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 eases the assessment burden, and reflects the perceived vertical intent which we are aiming to study.

We used a pre-existing aggregated search test collection [4] as the main source queries, documents, and verticals. This collection models eleven representative verticals commonly used on the web (i.e. image, video, recipe, news, books, blog, answer, shopping, discussion, scholar, wiki) and contains 320 web topics.

Assessors were anonymous online respondents who participated freely via a web interface when contacted through a number of mailing lists. We provided the assessors with the names of all the verticals and details of their unique characteristics (e.g. an "image" vertical might provide more visually attractive results) with an exemplar information need and results collected from Google search engine. The main criteria for labeling a vertical as intended (by a query) was if the annotator believed it would be beneficial to add the items from the vertical to the *ten blue links*. To eliminate order bias, we randomized all 320 topics (i.e. title and description) into a set of pages (with five topics per page) and provided each assessor the options to assess as many pages as he/she wished.

For measuring users' agreement over verticals, we report inter-annotator agreement in terms of Fleiss' Kappa [2] (denoted by K_F), which corrects for agreement due to chance. Furthermore, for deriving the binary vertical intent vector I_t from M_t, we use a threshold approach using three threshold values of 50%, 75% and 100% (e.g. if 75% of the assessors agree that a specific vertical v_j is a vertical intent, then we label i_j as 1, otherwise 0).

We collected 75 assessment sessions (i.e. assessors) with a total of 1,515 assessments. The average assessments per session varied from 5 to 120. The mean of the number of relevant verticals per topic per session is 2.06, with a standard deviation of 1.09. With topics (231/320) possessing more than four assessments, the distribution of derived vertical intents I_t is presented in Table 1. The number in the "web-only" column shows the number of topics where the assessors have not assigned any relevant verticals (i.e. "general web" results is all that

Table 1. Distribution of Number of Topics Assigned to Various Vertical Intents (with Different Assessors' Majority Preference)

Majority Preference	Img	Vid	Recipe	News	Book	Blog	Ans	Shop	Disc	Schol	Wiki	Web-only	Total Qrys
50%	41	13	7	22	25	22	38	4	38	11	139	30	209 (90.4%)
75%	16	5	4	9	5	6	4	0	10	2	73	9	114 (49.4%)
100%	4	0	3	0	0	2	0	0	1	0	29	1	38 (16.5%)

should be presented). It can be observed that many topics possess vertical intents (with various assessor majority preference level) thus confirming that users prefer diverse vertical content for many queries.

The mean of the user agreement K_F of vertical intents over all topics with more than four assessments is 0.37, which is considered *fair* agreement. For those topics, 14.7% have agreement considered at least *substantial* ($0.4 < K_F < 1$), 37.2% considered *fair* ($0.2 < K_F < 0.4$), and 48.1% considered *slight* or *poor* ($K_F < 0.2$). For a large number of queries, the agreement is not particularly high, as assessors might have different preferences over the number of intents for verticals or due to the ambiguous nature of the query. However, this analysis allows us to study different sets of queries where we have varying levels of agreement of perceived usefulness for a vertical. As 75% assessors' majority preference is neither too noisy (50%) or stringent (100%), and it more realistically conforms to the real-world vertical intent distribution [1], we select the corresponding vertical intent set for further experiments.

3 Predicting Query Vertical Intent Using Query-logs

Having collected labels that pertain to the intended verticals for a set of queries, our next step is to automatically predict these vertical intents. This is different from the resource selection task in federated search where the prediction of query vertical intent requires determining the specific number of intended verticals (zero to many). Therefore, given the topical relevance assessments over items within verticals, we aim to investigate if verticals with a high intent for a topic contain above a certain *threshold* of relevant items. The research questions we want to answer are: "is the recall of topically relevant items within each vertical (traditional resource selection criteria [3], denoted as "Trad") sufficient in predicting vertical intent?" and "is our newly developed query-log approach (denoted as "QLog") able to make more accurate vertical intent predictions?".

For our query-log approach (i.e. using the large scale AOL query-log), we use *click-through data* to infer a threshold for each vertical, above which the vertical is assumed to have a high intent for a query. Our approach consists of several steps. First we identify a set of queries in this query-log that we deem to possess particular verticals of a high intent. We identified those queries 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 query for which the 'recipes' vertical has a high intent). We also used queries that are particular sub-queries of those found using the previous step. For example, we also assumed the query "pork chops" was a 'recipe' query if it appeared in the query-log. We also used a number of humanly annotated variants of the vertical labels to identify queries (e.g. for queries relating to 'image' verticals, we used the terms 'image', 'images', 'img', 'picture', 'pictures', 'photo', 'photos', 'pics'). Secondly, we classified *all the clicked documents* for those queries into verticals using a similar approach on the URLs of those *clicked documents*. For example, the URL "http://www.recipes.com/chicken" has four terms, "www", "recipes",

Table 2. Comparison of Various Approaches on Vertical Intent Prediction

	QLog	Trad(1)	Trad(2)	Trad(3)	Trad($Perfect$)
F-measure	0.602	0.553▼	0.437▼	0.363▼	0.547▼

"com" and "chicken" and is therefore classified as belonging to the "recipe" vertical because it matches the one of that vertical's name variants. Finally, we *infer the threshold* by calculating 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 are then averaged over all queries identified to have that vertical intent. 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 of high intent to the topic. Using this dynamic threshold, a vertical is deemed of high intent to a query, when the vertical contains over the threshold of relevant documents calculated above.

We tested "QLog" approach over our topics and measured the performance of various approaches using the F-measure. We compare the "QLog" approach with "Trad" at various vertical intent thresholds (fixed threshold 1, 2 and 3, and "Perfect", i.e. setting a query-specific threshold to the number of intended verticals for each topic). Significance with respect to "QLog" was tested using a 2-tailed paired t-test ($p < 0.05$) on topics (denoted by ▼). Table 2 shows the results. We can observe that using the recall of relevant items within verticals performs considerably well in predicting vertical intents. Therefore, the fraction of topically relevant items in a collection is related to vertical intent. Furthermore, our proposed query-log approach performs consistently better than traditional resource selection approaches by inferring a dynamic threshold using query-logs.

References

1. J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo.: Sources of evidence for vertical selection. In SIGIR09, 315-322, 2009.
2. J. Fleiss.: Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5):378-382, 1971.
3. J. C. French, A. L. Powell.: Metrics for evaluating database selection techniques. In World Wide Web 3(3): 153-163, 2000.
4. K. Zhou, R. Cummins, M. Lalmas, and J. Jose.: Evaluating large-scale distributed vertical search. In LSDS-IR workshop in CIKM11, 9-14, 2011.