Evaluating Reward and Risk for Vertical Selection

Ke Zhou
University of Glasgow
Glasgow, United Kingdom
zhouke@dcs.gla.ac.uk

Ronan Cummins
National University of Ireland
Galway, Ireland
ronan.cummins@nuigalway.ie

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

Joemon M. Jose
University of Glasgow
Glasgow, United Kingdom
jj@dcs.gla.ac.uk

ABSTRACT

The aggregation of search results from heterogeneous verticals (news, videos, blogs, etc) has become an important consideration in search. When aiming to select suitable verticals, from which items are selected to be shown along with the standard "ten blue links", there exists the potential to both help (selecting relevant verticals) and harm (selecting irrelevant verticals) the existing result set.

In this paper, we present an approach that considers both reward and risk within the task of vertical selection (VS). We propose a novel risk-aware VS evaluation metric that incorporates users' risk-levels and users' individual preference of verticals. Using the proposed metric, we present a detailed analysis of both reward and risk of current resource selection approaches within a multi-label classification framework. The results bring insights into the effectiveness and robustness of current vertical selection approaches.

Categories and Subject Descriptors

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

General Terms

Measurement, Experimentation

Keywords

aggregated search, vertical selection, evaluation

1. INTRODUCTION

With the emergence of numerous vertical search engines, it is popular to present results from a set of verticals dispersed throughout the standard "general web" results (e.g. adding images results to queries about "flowers"). A key component of so-called aggregated search is vertical selection (VS), that is selecting multiple (zero to many) relevant verticals from which items are selected and presented on the search result page. Current work has focused on selecting a single relevant vertical [3], or on ranking vertical blocks that in turn are to be presented on the aggregated page [1].

When selecting suitable verticals, there exists the potential to both help (selecting relevant verticals) and harm (selecting irrelevant verticals) the existing result set. A VS system should only select a vertical when it is confident that it will benefit most users while seldom frustrating others. Existing work ([2][3][13]) evaluates VS based solely on maximising reward (e.g. the number of queries correctly classified as relating to a vertical [3]), or the average correlation with the "perfectly ranked" reference page [2]. We argue that for VS, reward must be considered in conjunction with risk. We argue that maximising the reward alone is not sufficient, and that a robust VS approach and its evaluation should focus on maximising reward while minimising risk.

We propose a new risk-aware VS evaluation metric. Rather than treating a vertical as either relevant or irrelevant given a query, as mostly done in current work [3], we propose a general framework to evaluate the reward and risk for VS on a per user basis. This is motivated by the fact that current research [12] shows that the level of inter-annotator agreement for what constitutes a 'relevant' vertical is low (users' preferred verticals are diverse). Our proposed metric is flexible as it allows systems to be evaluated across a population of users, where users may have varying levels of risk (risk-averse vs. risk-seeking) and may have varying preferences across verticals (vertical relevance is user specific). In this paper, we perform an analysis of the effectiveness of different VS approaches across these different types of user [5]. Furthermore, we present an analysis of the robustness of VS approaches across all users with various levels of risk[1].

We treat VS as a multi-label classification problem (multiple verticals are relevant to a query) and we train a set of VS systems according to different controlled risk-levels (some systems are more risk-averse than others). We then analyse these trained VS systems with varying types of user (risk-averse and risk-seeking). We hypothesise that:

- (**effectiveness**) some VS approaches are better suited to some types of users than others;

[1]An analysis of the distribution of risk-levels in the user population lies outside the scope of this work. This information could be estimated from query logs or through a survey of a sample of users.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

- (**robustness**) some VS approaches are more robust for a mixture of varying types of users than others.

Section 2 outlines our proposed risk-aware VS metric. In Section 3, we formally describe the problem of multi-label vertical classification and list the features used. In Section 4, we empirically evaluate the effectiveness and robustness of those approaches using our proposed risk-aware metric. We conclude the paper in Section 5.

2. EVALUATING REWARD AND RISK

We present our risk-aware metric for VS, which considers an entire population of users' vertical preferences for a query.

2.1 Problem Formulation

Let $V = \{v_1, v_2, ...v_n\}$ be a set of verticals that can be selected to present along with "general web" results W, for a given query $q \in Q$. Let $V_q^{u_i}$ be a set of verticals that a user $u_i \in U$ would like to see in the result set with "general web" results for query q. These user-specific assessments can be obtained by either conducting a user study that explicitly asks users for their preferences [12] or be estimated by mining query logs [7]. We model this subjective view of vertical relevance where users' vertical preferences can be different [12]. Therefore, $V_q^{u_i} \subset V$ and $V_q^{u_j} \subset V$.

Furthermore, assume a vertical selection system s_j selects a vertical set $V_q^{s_j}$ for q. Then, for a specific user u_i, the utility of vertical search system s_j is based on both *reward* and *risk*. *Reward* is related to the number of verticals selected by s_j that user u_i deems relevant ($V_q^{s_j} \bigcap V_q^{u_i}$). While *risk* is related to the number of verticals selected by s_j that user u_i deems non-relevant ($V_q^{s_j} \bigcap (V - V_q^{u_i})$).

Furthermore, each user has his/her own estimated trade-off between reward and risk. For example, one user might be *risk-seeking* and prefers to have a page with some relevant verticals but does not mind viewing many non-relevant ones. On the contrary, another users might be *risk-averse* and prefers the page to only contain relevant verticals. Therefore, the main aim of the proposed metric is to model the trade-off between so-called *reward* and *risk* for each user u_i.

2.2 Risk-aware Metric

For a given user u_i and system s_j that returns $V_q^{s_j}$, we define the *reward* and *risk* as user-specific vertical $recall$ and vertical $fallout$ respectively as follows:

$$reward_q^{u_i}(V_q^{s_j}) = \frac{|V_q^{s_j} \bigcap V_q^{u_i}|}{|V_q^{u_i}|} \quad (1)$$

$$risk_q^{u_i}(V_q^{s_j}) = \frac{|V_q^{s_j} \bigcap (V - V_q^{u_i})|}{|\bar{V}_q^{u_i}|} \quad (2)$$

To combine the above measure and also incorporate the user's trade-off between reward and risk, we model the metric as a linear combination of reward and risk:

$$util(V_q^{s_j}, \alpha_q^{u_i}) = (1-\alpha_q^{u_i}) \cdot reward_q^{u_i}(V_q^{s_j}) + \alpha_q^{u_i} \cdot (1 - risk_q^{u_i}(V_q^{s_j})) \quad (3)$$

where $\alpha_q^{u_i}$ is a user-specific parameter that controls the trade-off between reward and risk. Setting $\alpha_q^{u_i} = 1$ leads to a risk-averse metric where returning zero irrelevant verticals would be optimal, while setting $\alpha_q^{u_i} = 0$ leads to a risk-seeking metric where returning as many relevant verticals would be optimal.

The utility of the system s_j is averaged over all $q \in Q$, and within each q is averaged over all users U. We define the utility of the system as follows:

$$Util(s_j, \alpha) = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{u_i \in U} util(V_q^{s_j}, \alpha_q^{u_i})}{|U|} \quad (4)$$

This $Util(s_j, \alpha)$ function treats all (both popular and long-tailed) queries equally and is not biased to popular queries. Although other approaches to derive utility within this framework are possible, we will leave them for future work.

At this point we have one utility metric for evaluating a VS system, accounting for reward and risk. The metric depends on the user-specific and query-specific reward-risk tradeoff parameter $\alpha_q^{u_i}$, which we need to set. In this paper, we assume that for each query q, users have the same trade-off level (α) between reward and risk. Furthermore, we assume a uniform distribution of $\alpha_q^{u_i}$ across all users. We leave the work of discovering the distribution of *risk-seeking* and *risk-averse* for future work. Using our metric we can compare vertical selection approaches for both *risk-seeking* and *risk-averse* users over a set of queries Q. Furthermore, we can measure the robustness of the VS approach over all types of users (assuming uniformity) by iterating over all values of α for all queries in Q.

3. MULTI-LABEL CLASSIFICATION

We introduce the risk-aware multi-label classification approach, followed by detailed descriptions of features used.

3.1 Risk-aware Classification Approach

The approach to classification consists of two phases: testing and training. We separate 56 queries (conforming to a real-world distribution of verticals [3]) as a training set. This is used for determining a threshold γ (see below). We use the remaining dataset (264 queries) for testing the approaches.

We use a thresholding approach to select verticals. For a set of verticals $V = \{v_1, v_2, ...v_n\}$ with scores $X^{s_j} = \{x_1, x_2, ...x_n\}$ (generated by a vertical selection approach s_j) and a threshold γ, we denote $V_{x_i>\gamma}^{s_j}$ as the set of verticals with each vertical v_i whose score $x_i > \gamma$. If no vertical has $x_i > \gamma$, then $V_{x_i>\gamma}^{s_j} = \emptyset$. Note that each vertical score x_i is obtained by normalising across all vertical scores.

In essence, the vertical scoring functions of each VS approach is adapted to multi-label vertical selection by selecting the top-k verticals where k is decided by a threshold γ. The threshold is trained on the training set. If no verticals receive a score greater than the threshold, no verticals are deemed relevant for that query.

With respect to the risk-aware training, for a given vertical selection approach s_j with scores over all verticals $X^{s_j} = \{x_1, x_2, ...x_n\}$, we train a set of systems $S_j = \{s_j^{\alpha_1}, s_j^{\alpha_2}, ...s_j^{\alpha_m}\}$ where each system varies in its reward-risk trade-off operating point (by setting different training objective functions with different α, and obtaining corresponding γ), i.e. some of the systems are trained to be more risk-averse whereas others to be more risk-seeking. The optimal threshold γ^* for a given system s_j^α (with reward-risk trade-off α) is trained as follows:

$$\gamma^* = argmax_\gamma \; Util(V_{x_i>\gamma}^{s_j^\alpha}, \alpha) \quad (5)$$

Therefore, for each feature (vertical selection approach s_j), we iterate α and obtain a set of systems S_j.

3.2 Features

We investigate a number of resource selection approaches (CORI [4], Clarity [6], GAVG [8], ReDDE [10], CRCS(l) [9], CRCS(e) [9]) as features for multi-label VS approaches. We use each feature individually for training and aim to compare them. While these approaches derive evidence from the same source (sampled vertical representation), they model different aspects of the sources under consideration. CORI, Clarity and GAVG model the similarity between the query and the source, whereas ReDDE, CRCS(l) and CRCS(e) model the collection's average document score in a full-dataset retrieval (all sources together).

3.2.1 CORI

CORI adapts INQUERY's inference net document ranking approach to collection. Here, all statistics are derived from sampled documents rather than the full collection.

3.2.2 Clarity

Clarity is a retrieval effectiveness prediction algorithm that measures the similarity between the language of the top ranked documents and the language of the collection, estimated using the Kullback-Leibler divergence between the query θ_q and the collection language model θ_{v_i}.

$$Clarity_q(v_i) = \sum_{w \in v_i} P(w|\theta_q) log_2 \frac{P(w|\theta_q)}{P(w|\theta_{v_i})} \quad (6)$$

3.2.3 Geometric Average

GAVG issues the query to a centralized sample index, one that combines document samples from every vertical, and scores vertical v_i by the geometric average query likelihood from its top m sampled documents.

$$GAVG_q(v_i) = \left(\prod_{d \in topm} P(q|\theta_d) \right)^{\frac{1}{m}} \quad (7)$$

3.2.4 ReDDE

ReDDE scores a target collection based on its expected number documents relevant to the query. It derives this expectation from a retrieval of an index that combines documents sampled from every target collection. Given this retrieval, *ReDDE* accumulates a collection score $ReDDE_q(v_i)$ from its document scores $P(q|\theta_d)$, taking into account the difference between the size of the original collection N^{v_i} and a sampled set size N^{samp}.

$$ReDDE_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d \in topm} I(d \in v_i) P(q|\theta_d) \quad (8)$$

where $I(.)$ is a indicator function.

3.2.5 CRCS

Like *ReDDE*, *CRCS* issues the query to a centralized sample index and scores a collection according to an accumulation of a more refined estimation of document score. Specifically, the document score for *CRCS(l)* and *CRCS(e)* are estimated by a linear or a negative exponential weighting according to its presented position respectively.

$$CRCS(l)_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d_j \in topm} (m - j) \quad (9)$$

$$CRCS(e)_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d_j \in topm} \alpha \cdot exp(-\beta \cdot j) \quad (10)$$

where $\alpha = 1.2$ and $\beta = 2.8$ in our setting.

4. EXPERIMENTS

Our experiments aim to investigate various resource selection approaches under our risk-aware multi-label classification framework. We report the data used in the experiments first, followed by the main experimental results on both *effectiveness* and *robustness*.

4.1 Data

The user-specific preferred vertical ground-truth information of each query ($V_q^{u_i}$) is obtained by only providing the vertical names (with a description of their characteristics) and asking a set of assessors to make pairwise preference assessments, comparing each vertical in turn to the reference "general web" vertical [12]. We used an existing web test collection [11] to obtain the vertical representations used for the vertical selection approaches. The verticals used and the distribution of majority user preferred verticals (more than 50% of the users preferred the vertical to "general web") for all queries for the collection are described in Table 1.

4.2 Evaluating VS Approaches

4.2.1 Effectiveness

A VS approach s_j trained on a given user risk-level α is tested on the corresponding type of user (with same α). An approach is *effective* if prediction of relevant verticals $V_q^{s_j}$ can satisfy users of that type (i.e. high $Util(s_j, \alpha)$).

The main evaluation results on effectiveness for single-feature (each resource selection approach) classifier runs are shown in Figure 1. When only reward is considered ($\alpha = 0$), all of the approaches perform comparably. However, when risk is considered ($\alpha > 0$), we observe that in general, ReDDE performs consistently better than the other approaches. From a 2-tailed paired t-test ($p < 0.05$), we find that ReDDE is significantly better than GAVG and CRCS(e) at $\alpha = 0.3, 0.4$, CRCS(l) at $\alpha = 0.3$, Clarity and CORI at $\alpha = 0.2, 0.3, 0.4, 0.5$. Of the VS approaches tested CRCS(l) and ReDDE are more risk-aware (when $\alpha > 0.3$ for example). However, when favouring reward (low α), GAVG and ReDDE achieve higher results. CORI and Clarity are, on average, the worst approaches across many values of α.

We also empirically observe that different approaches perform differently for a range of queries whereas some of them hinder/increase the performance of more queries than the other when applying vertical selection. The percentage of benefited and hindered queries conforms to the training setting of the reward-risk trade-off. This demonstrates the need for current VS approaches to be more risk-aware.

In conclusion, comparably, ReDDE and CRCS(l) achieve the best performance on *effectiveness* in those settings, mostly with a large range of queries benefited and a small amount hindered.

4.2.2 Robustness

Rather than evaluating on one single type of user, robustness of VS approach is measured over all types of users (assuming uniformity) by iterating over risk-level α for all queries.

Table 1: Distribution of Number of Queries Assigned to Majority User Prefered Verticals

Verticals	Image	Video	Recipe	News	Book	Blog	Answer	Shop	Discuss	Scholar	Wiki	Web-only	Total Qrys
Qry num	41	13	7	22	25	22	38	4	38	11	139	141	320

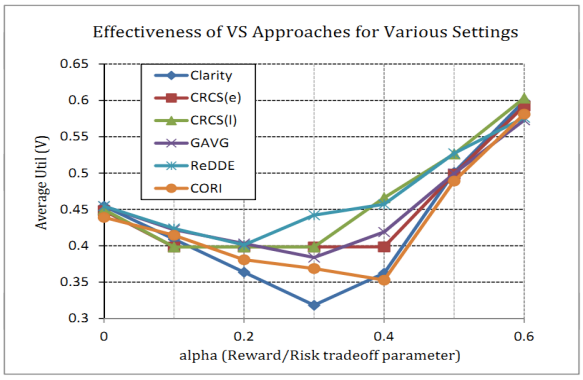

Figure 1: Comparing Effectiveness for Various Vertical Selection Approaches

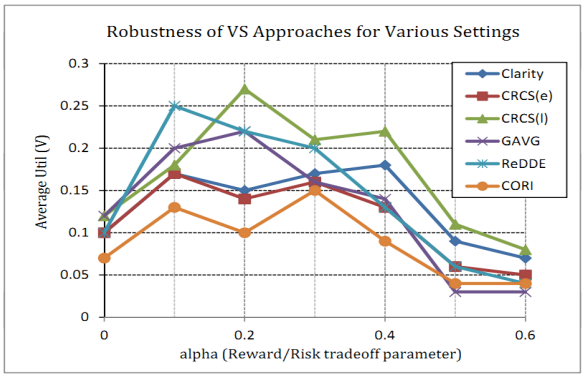

Figure 2: Comparing Robustness for Various Vertical Selection Approaches

The main evaluation results on robustness are shown in Figure 2. Firstly, we can observe a general trend that VS approaches that balance the trade-off between reward and risk perform better than the ones that considers solely reward or solely risk. This is not surprising since VS approaches that solely maximise reward frustrate most of users that are *risk-averse*. On the contrary, only minimising risk could degrade user experience for users that are *risk-seeking*.

Secondly, it can be observed that in general, CRCS(l) perform more robust than other approaches. From a 2-tailed paired t-test ($p < 0.05$), we find that CRCS(l) is significantly more robust than all other approaches when $\alpha >= 0.4$. When $\alpha < 0.4$, CRCS(l) is significantly better than CORI at $\alpha = 0.0, 0.1$, GAVG, Clarity, CRCS(e) and Clarity at $\alpha = 0.2, 0.3$, ReDDE at $\alpha = 0.2$. Of the VS approaches tested, CRCS(l) and Clarity are more risk-aware (when $\alpha > 0.4$ for example). However, when favouring reward (low α), GAVG and ReDDE achieve higher results. CORI and CRCS(e) are, on average, the worst approaches across many values of α. We can conclude that CRCS(l) achieve the best performance on *robustness* in our settings.

5. CONCLUSIONS AND FUTURE WORK

This paper incorporates a risk-aware evaluation of vertical selection approaches in a multi-label classification framework. We propose a novel multi-label vertical selection evaluation metric that incorporates both rewards and risks. We present a detailed empirical analysis of both effectiveness and robustness of current vertical selection approaches. We demonstrate that ReDDE is the most effective VS approach and CRCS(l) is the most robust.

Future work might include investigating more VS approaches (e.g. query-log based) in this multi-classification framework and study their robustness. Further investigations on real-world user risk-level distribution for this evaluation would provide more insights. A detailed analysis on how this novel risk-aware metric correlate with user satisfaction would further verify its fidelity.

Acknowledgments This work was supported partially by the EU LiMoSINe project (288024). Any opinions, findings, and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsors.

6. REFERENCES

[1] J. Arguello, F. Diaz, and J. Callan. Learning to aggregate vertical results into web search results. *CIKM* 2011.
[2] J. Arguello, F. Diaz, and J. Callan. A methodology for evaluating aggregated search results. *ECIR* 2011.
[3] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. *SIGIR* 2009.
[4] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. *SIGIR*, 1995.
[5] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. *CIKM* 2011.
[6] S. Cronen-Townsend, Y.Zhou, and W.B.Croft. Predicting query performance. *SIGIR*, 2002.
[7] 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. *WSDM*, 2011.
[8] J. Seo and B.W. Croft. Blog site search using resource selection. *CIKM* 2008.
[9] M. Shokouhi. Central-Rank-Based Collection Selection in Uncooperative Distributed Information Retrieval. *ECIR* 2007.
[10] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. *SIGIR* 2003.
[11] K. Zhou, R. Cummins, M. Lalmas, and J.M. Jose. Evaluating large-scale distributed vertical search. *CIKM Workshop LSDS-IR* 2011.
[12] K. Zhou, R. Cummins, M. Halvey, M. Lalmas and J. M. Jose. Assessing and Predicting Vertical Intent for Web Queries. *ECIR* 2012.
[13] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating aggregated search pages. *SIGIR* 2012.