Does Vertical Bring more Satisfaction? Predicting Search Satisfaction in a Heterogeneous Environment

ABSTRACT

The study of search satisfaction is one of the prime concerns in search performance evaluation research. Most existing works on search satisfaction primarily rely on the hypothesis that all results on search engine result pages (SERPS) are homogeneous. However, a variety of heterogeneous vertical results such as videos, images and instant answers are aggregated into SERPs by search engines to improve the diversity and quality of search results. In this paper, we carry out a lab-based user study with specifically designed SERPs to determine how verticals with different qualities and presentation styles affect search satisfaction. Users' satisfaction feedback and external assessors' satisfaction annotations are both collected to make a comparison regarding the perception of search satisfaction. Mouse click-through / movement data and eye movement information are also collected such that we can investigate the influence of vertical results from the perspectives of both benefit and cost. Finally, a learning-based framework is proposed to predict search satisfaction on aggregated SERPs. To the best of our knowledge, this paper is the first to analyze the effect of verticals on search satisfaction. The results show that verticals with different qualities, presentation styles and positions have different effects on search satisfaction, among which Encyclopedia verticals, as well as Download verticals, will bring the largest improvement. Furthermore, our proposed prediction framework outperforms state-of-the-art methods that are designed for search satisfaction prediction in homogeneous environment.

Categories and Subject Descriptors

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

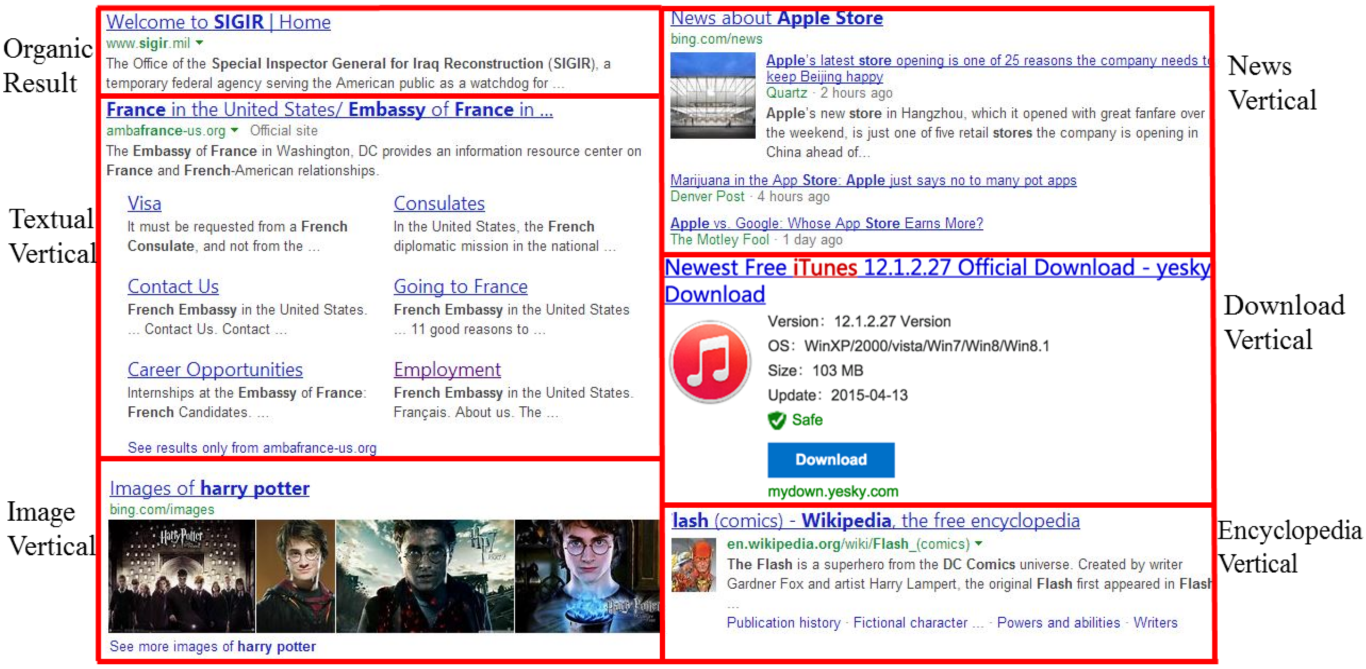
Keywords

search satisfaction; aggregated search; benefit; cost; prediction

1. INTRODUCTION

Search engine evaluation can be performed using metrics based on result relevance or alternative measures based on users' search experience. Recent studies indicate that relevance-based evaluation metrics, such as MAP and NDCG [18], may not be perfectly correlated with users' search experience (usually considered as the gold standard) [2, 17]. Therefore, search satisfaction has become one of the major concerns in search evaluation studies. Since satisfaction is a relatively subjective concept, several works [17, 34, 19] have tried to quantify users' perceived satisfaction. Furthermore, some works [1, 12, 15, 16] tried to use various types of user interactions (click-through data, mouse/eye movement information) as implicit feedback to predict search satisfaction. These existing works have achieved success on how to model users' judgements on the whole search process and how to improve search engines' ranking strategy. However, most of these works rely on the hypothesis that all results on search engine result pages (SERPs) are homogeneous, which means that all search results in an SERP share similar presentation style (one hyperlink with short snippet). However, as more and more heterogeneous vertical results (videos, images, instant answers and so on) are aggregated into modern SERPs, a user's examination and clicking behavior will be quite different [33, 35]. Because modern SERPs provide richer content than hyperlinks and short texts, the sense of fulfilling information needs during the search process may be very different from an SERP with "ten blue links". Therefore, we need to investigate users' satisfaction perception process within a heterogeneous search environment. To better describe heterogeneous SERPs, we first give a taxonomy of different search results according to their presentation styles (see Figure 1):

- Non-vertical result: One blue hyperlink with short snippet contents.
- Textual vertical: Verticals that are composed of a couple of textual snippets and blue hyperlinks. It usually provides more all-around information concerning the query topic.
- Image vertical: Verticals that are usually composed of several images grouped into one or two rows.
- News vertical: Verticals that have an image on the left and provide several hyperlinks lead to the latest news about the query topic.
- Download vertical: Download verticals are a popular type of verticals in SERPs of Chinese commercial search engines. They can provide searchers with a button to directly download the application that the query topic describes.



Welcome to SIGIR | Home
www.sigir.mil ▾
The Office of the Special Inspector General for Iraq Reconstruction (SIGIR), a temporary federal agency serving the American public as a watchdog for ...

France in the United States/ Embassy of France in ...
ambafrance-us.org ▾ Official site
The Embassy of France in Washington, DC provides an information resource center on France and French-American relationships.

Visa
It must be requested from a French Consulate, and not from the ...

Consulates
In the United States, the French diplomatic mission in the national ...

Contact Us
French Embassy in the United States. ... Contact Us. Contact ...

Going to France
French Embassy in the United States ... 11 good reasons to ...

Career Opportunities
Internships at the Embassy of France: French Candidates. ...

Employment
French Embassy in the United States. Français. About us. The ...

See results only from ambafrance-us.org

Images of harry potter
bing.com/images
See more images of harry potter

News about Apple Store
bing.com/news
Apple's latest store opening is one of 25 reasons the company needs t keep Beijing happy
Quartz · 2 hours ago
Apple's new store in Hangzhou, which it opened with great fanfare over the weekend, is just one of five retail stores the company is opening in China ahead of...

Marijuana in the App Store: Apple just says no to many pot apps
Denver Post · 4 hours ago

Apple vs. Google: Whose App Store Earns More?
The Motley Fool · 1 day ago

Newest Free iTunes 12.1.2.27 Official Download - yesky Download
Version: 12.1.2.27 Version
OS: WinXP/2000/vista/Win7/Win8/Win8.1
Size: 103 MB
Update: 2015-04-13
✓ Safe
Download
mydown.yesky.com

lash (comics) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Flash_(comics) ▾
The Flash is a superhero from the DC Comics universe. Created by writer Gardner Fox and artist Harry Lampert, the original Flash first appeared in Flash
...
Publication history · Fictional character ... · Powers and abilities · Writers

Figure 1: Different presentation styles of vertical results on SERPs (we use the results of Bing as examples)

- Encyclopedia vertical: Verticals that provide an image as well as some textual information at the same time and thus can provide more comprehensive information to the query topic.

From Figure 1, we can see that the appearances of vertical results can be rather different from non-vertical results and may provide information in a completely different way. Previous works [33, 23] showed that such vertical results may have a strong effect on user behavior. These findings inspired us to investigate the effect of verticals on user satisfaction. To shed light on this question, we construct a lab-based search engine system with specifically designed heterogeneous search result pages. We provide SERPs with verticals that vary in qualities, presentation styles and positions to users to see how their satisfaction is affected. Users' explicit satisfaction feedback, as well as mouse click-through / movement data and eye movement data, are collected so that we can make a detailed analysis of how vertical results affect users' search satisfaction from the perspective of both benefit and cost. To avoid the subjectivity of user satisfaction feedback, we also invite external assessors to annotate the satisfaction scores of the users' search sessions to make a comparison. Finally, we propose a learning-based framework to predict search satisfaction on aggregated search result pages. The results show that our proposed method outperforms state-of-the-art methods, which are not specifically designed for heterogeneous search pages.

Our contributions in this paper include: (1) To the best of our knowledge, we are among the first to study the effect of vertical search results on search satisfaction. (2) With rich information collected by an experimental search engine system, we make a deep analysis of the effect vertical results have on both users' and external assessors' satisfaction judgements in a benefit-cost framework. (3) We propose a learning-based prediction framework to predict the search satisfaction of SERPs with heterogeneous vertical results and demonstrate its effectiveness by comparing with state-of-the-art methods.

The rest of this paper is organized as follows: Related works are reviewed in Section 2. Our lab-based search engine system and corresponding data collection process are presented in Section 3. The effect of different verticals on search satisfaction are shown in Section 4 and a deeper analysis in the benefit-cost framework is shown in Section 5. Section 6 introduces our satisfaction prediction framework and discusses its effectiveness. The paper's conclusions are presented in Section 7.

2. RELATED WORK

2.1 Search Satisfaction Prediction

The concept of satisfaction was first proposed by Su et al. [30] and was defined as "the fulfillment of a specified desire or goal" by Kelly [21]. To evaluate a search system, satisfaction can be considered as regarding not only to the whole search experience but also to some specific aspects [31], such as the precision or completeness of search results, response time and so on. Because search satisfaction is a subjective concept that is difficult to measure, some existing works collected explicit feedback directly from users as the ground truth of search satisfaction, such as Guo et al. [15] who predicted user satisfaction with mouse movement information, and Feild et al. [12] who predicted user frustration using query-logs. In addition to explicit search satisfaction feedback, some works [16, 17] have also tried to recruit external assessors to restore the users' search process and make satisfaction annotations according to their own opinions. However, recent research in [32] showed that external annotations may not be a good estimator of users' self-judgements, and a number of works (e.g, [19, 20]) have started using the benefit-cost framework to analyze the satisfaction judgement process of users. In this framework, both the benefit factors (result relevance) and search cost (effort) users spend are used to estimate satisfaction. In this work, we follow the benefit-cost framework to make a deep comparison between search satisfaction from users and external assessors in a heterogeneous environment.

2.2 Behavior Modeling on SERPS with Verticals

Recently, more and more heterogeneous search results have been aggregated into search result pages to promote users' search experiences. There are also a number of existing works which are focused on this kind of federated search. Among them, most prior works focused on predicting which verticals are relevant to a query (vertical selection). Diaz et al. [10] first carried out a system to collect news dynamically and aggregated them into web search results. Arguello et al. [4, 5] showed that query logs will be useful for selecting relevant verticals. Zhou et al. [35] presented an approach that considers both reward and risk within the task of vertical selection.

Because the presentation styles of different verticals may be rather different, a user's browsing behavior may be changed when an SERP becomes more and more heterogeneous. Some existing studies tried to analyze a user's new behavior pattern on a heterogeneous SERP: Wang et al. [33] found that different verticals may create examination biases on users' search behavior. They suggested that images and videos will attract a user's attention more than other search results. Liu et al. [23] further showed three three behavior effect in federated search, namely, the vertical attraction effect, the examination cut-off effect and the examination spill-over effect. Navalpakkam et al. [26] also showed that a knowledge graph will also influence a user's attention distribution on SERPs.

Traditional search result evaluation metrics may also become inappropriate when dealing with federated search pages. Various diversity aware IR metrics have been proposed [7, 8, 28], which may be adjusted to evaluate heterogeneous result pages. Zhou et al. [35] introduced the concept of vertical orientation and instantiated a suite of metrics for evaluating aggregated search pages. Markov et al. [25] proposed two vertical-aware metrics based on user click models for federated search and demonstrated its effectiveness.

Despite of these existing works, how users perceive satisfaction on aggregated search pages still remains uninvestigated. In this paper, we incorporate vertical information into SERPs and follow the benefit-cost framework to analyze the effect of vertical results on search satisfaction. We also propose a learning-based satisfaction prediction method and demonstrate the effectiveness of vertical information.

3. DATA COLLECTION

In this section, we describe the lab-based search engine system used in our work and show the process of how we collect search satisfaction scores as well as search interaction data, such as mouse and eye movements.

3.1 Experiment Procedure

To investigate the effect of verticals on search satisfaction, we construct a lab-based search engine system to collect user behavior data as well as satisfaction scores from both users and external assessors. The entire experiment procedure is shown in Figure 2, from which we can see that four types of information are collected during the procedure: (1) mouse movements and click-through information, (2) eye movements, (3)users' satisfaction scores and (4) external assessors' satisfaction annotations.

Before the experiment, each participant should first go through a calibration process as required by the eye tracker to make sure that reliable eye movement information is collected. The eye tracker we use in our work is Tobii X2-30. Each participant will be asked to complete 30 search tasks one by one within 1 hour during our experiment. The procedure of the experiment is shown in Figure 2. Before each task, they will first go through the search queries and corresponding explanations to make sure they know the task clearly. Then, he/she will be guided to a pre-designed SERP where query and search results are fixed. The participant should examine the search results provided by our system and click a button on the top right corner to end the task either if the search goal is completed or he/she becomes disappointed with the results. During such process, his/her mouse movement information and click-through data were logged by injected JavaScript on the SERPs, and eye movement information is also logged by the eye tracker. Each time the participant finishes a search task, he/she will be required to label a 5-point satisfaction score to the search session, where 5 means the most satisfactory and 1 means the least. Then, they will be guided to continue to the next search task. All participants are required to finish two warm-up search tasks first to become familiar with the experiment process.

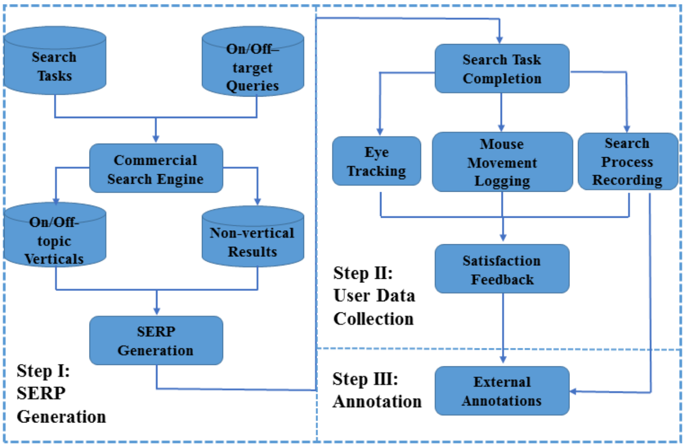


Figure 2: Data Collection Procedure

3.2 Search Tasks and SERP Generation

To investigate the effect of vertical results on satisfaction, we sample a large number of search queries based on the search logs from a major commercial search engine. We use such queries to organize our search tasks just to make sure that our experimental SERPs are consistent with the practical scenario. We selected 30 specific search tasks with corresponding on/off-topic verticals and non-vertical results crawled from the commercial search engine. The queries we use in our experiment are neither long-tailed nor hot ones to avoid unnecessary biases.

The SERPs we present to users vary in three aspects that may have effects on users' search behavior and satisfaction judgements:

- Quality: We investigate whether the quality of vertical result will affect search satisfaction. We use a subset of the terms from the original query or add a few new items to generate an off-target query and submit it to the search engine to get the off-topic vertical. Because the new query just overlapped a subset of the original one, the vertical results we obtained are usually irrelevant to the original query but appear to be quite similar to the on-topic verticals. This strategy has also been adopted by a number of existing works to generate irrelevant

Table 1: Examples of Search Tasks and Manipulated Off-target Queries to Retrieve Verticals

Vertical Presentation Style	Original Query	Off-target Query
Textual	poems describing spring rain	poems describing rain
	ancient Greek architectural style	ancient Greek
Encyclopedia	covering the sky (novel)	covering the sky (game)
	the 9th zone (movie)	the 9th zone (novel)
Image	nike basketball shoes	nike football shoes
	pictures of wine cabinet	pictures of cupboard
Download	iTunes download	iTools download
	Renren desktop app download	Weibo desktop app download
News	ebola virus mutation	news of ebola virus
	Chinese city competitiveness	Chinese enterprise competitiveness

verticals [3, 6, 33]. Table 1 shows some examples of the search tasks and corresponding queries used to crawl on/off-topic verticals in our experiment.

- Position: Existing work show that position bias affects user attention in federated search engine [33], thus may affect search process and satisfaction judgement. So we take this factor into consideration in our experiment. Vertical results will be randomly placed at position 1, 3 and 5 of the result lists, respectively.
- Presentation styles: We further evenly separate out search tasks into five groups according to the vertical's presentation style to see if different types of verticals will have different effect on satisfaction. Thus, the 30 search tasks will contain all 5 types of verticals in Figure 1, namely, textual vertical, image vertical, news vertical, download vertical and encyclopedia vertical results.

To investigate the effect of the above three factors, we generate seven different SERPs for each search task. The first SERP is a non-vertical one, which means that there are only ten non-vertical results shown in the result page. These non-vertical results are crawled from the same commercial search engine, and kept the original orders unchanged. The remaining six SERPs are composed of one vertical result and nine non-vertical results(the last organic result from the non-vertical SERP is excluded). Two verticals are of the same presentation style but with different quality on these six SERPs, including an on-topic vertical for three of them and an off-topic one for the other three. Each vertical is inserted at three different positions of the organic result list: the first rank, the third rank and the fifth rank. Thus, we obtain seven pages for a search task: 2 quality types × 3 position ranks + 1 organic. Therefore, we generate 210 (30 search tasks × 7 pages) SERPs in total.

In our experiment, each participant will go through all 30 search tasks but with different SERP settings. We adopt a Graeco-Latin square design to ensure that each SERP condition has the same opportunity to be shown to users.

3.3 Participants

We recruited 35 participants (aged 18-25) for the data collecting process. All participants are college students and have a variety of self-reported search engine usage experiences. Their majors vary, from biology, economics, social science to engineering. We did not invite computer science or electrical engineering students because they may be too familiar with the use of search engines and cannot represent ordinary search engine users. Each user completed the 30 search tasks and were paid 10 US dollars.

3.4 External Annotation

Considering the fact that the process of satisfaction judgement may be subjective and different users may have different opinions, we recruited several external assessors to label the satisfaction scores of users' search sessions. All these assessors had worked in a commercial search engine company for at least one year and can be regarded as professionals in judging search performance and satisfaction. We exported the video of our participants doing their search tasks from the eye tracker, which had recorded the whole search process as well as eye/mouse movements and click-through information. We split the video into sessions and excluded the part where users labeled satisfaction scores. We showed these videos of search sessions to assessors so that they can fully restore the original searchers' search experience and make a reasonable annotation. The assessors were also asked to give a 5-point satisfaction score so that the satisfaction scores from the two resources will be comparable. They were paid 10 US dollars for annotating every 60 search sessions. Each search session is annotated by two assessors, and the KAPPA coefficient of their annotations is 0.48, which means a moderate agreement according to Cohen [9].

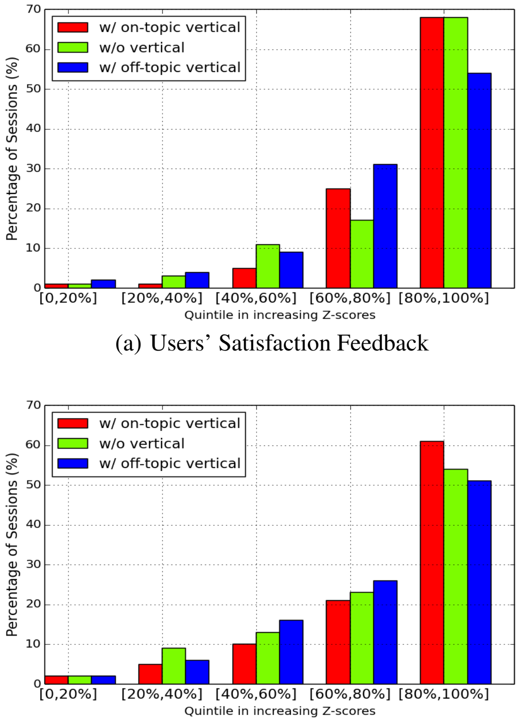
4. EFFECT OF VERTICALS ON SATISFACTION

4.1 Effect of Vertical Quality

We collected 768 search sessions in total because some participants fail to pass the calibration of the eye tracker. With these collected data, we try to compare the difference between the satisfaction scores from actual users and external assessors. Considering that satisfaction judgement may be quite subjective and different users may have different opinions, we regularize the satisfaction scores labelled by each user/assessor to Z-scores according to equation 1, where sat_i is one particular satisfaction score given by one user/assessor and $Avg(Sat)$ is the average of all satisfaction scores he/she labelled. $Var(Sat)$ in this equation refers to the variance of the satisfaction scores of this user/assessor. We should note that in this way, a Z-score may be below zero, which may be difficult to understand. Thus, we add the same δ_1 to all Z-scores of users' satisfaction (and the same δ_2 to all Z-scores of the external assessors) in this section to normalize the minimum value to zero to avoid confusion and to maintain the relative differences at the same time.

$$Z\text{-}score_i = \frac{sat_i - Avg(Sat)}{Var(Sat)} + \delta_{1/2} \quad (1)$$

Figure 3 shows an overview of the effect of vertical qualities on



(a) Users' Satisfaction Feedback

(b) External Assessors' Satisfaction Annotation

Figure 3: Satisfaction Distribution Based on Verticals with Different Qualities

satisfaction scores from two different resources. Different colors show satisfaction scores on pages with on/off-topic verticals or without verticals. We can see that both users and assessors tend to give a high satisfaction score, which indicates that commercial search engines generally provide promising results for these non-long-tailed queries. We can also see from Figure 3 that both users and assessors tend to be less satisfied when the vertical is off-topic because the percentage of sessions with the highest Z-scores (80%-100%) is comparatively lower in the off-topic case, although the difference is not very remarkable (68% for SERPs without verticals, 54% for SERPs with off-topic verticals in user satisfaction and 54%, 51% in assessors' annotations, relatively). Moreover, the percentage of sessions with low Z-scores (0%-60%) is also lower in the on-topic case (6% users and 17% assessors) than that in non-vertical case (15% users and 24% assessors), which indicates that both users and assessors tend to be more satisfied if there are on-topic verticals on SERPs.

4.2 Effect of Verticals with Different Presentation Styles

Table 2 shows the effect of verticals with different presentation styles on satisfaction scores from both users and external assessors. The values shown in the second, third and fourth columns are the average Z-scores of all the pages' verticals with the corresponding presentation style and quality (non-vertical, on/off-topic vertical). Values shown in parenthesis are the differences compared with that on pages without verticals. Values in the last column shows the improvement in Z-scores from pages with off-topic verticals (in the fourth column) to those with on-topic verticals (in the third column).

Table 2: Effect of Verticals with Different Presentation Styles on Satisfaction (* indicates statistical significance at $p < 0.1$ level,** indicates statistical significance at $p < 0.05$ level)

	w/o vertical	w/ on-topic vertical	w/ off-topic vertical	on-off difference
Users' Satisfaction Feedback				
Textual	5.15	5.10 (-0.05)	4.95 **(-0.20**)**	+0.15*
Image & Textual	4.46	4.99 **(+0.53**)**	4.67 (+0.21)	+0.32**
Image	5.17	5.07 (-0.10)	4.58 **(-0.59**)**	+0.49**
Download	4.75	5.25 **(+0.50**)**	4.60 (-0.15)	+0.65**
News	4.43	4.34 (-0.09)	4.38 (-0.05)	-0.04
External Assessors' Satisfaction Annotation				
Textual	3.75	3.64 (-0.11)	3.45 **(-0.30*)**	+0.19*
Image & Textual	3.18	3.62 **(+0.44**)**	3.13 (-0.05)	+0.49**
Image	3.34	3.59 **(+0.25*)**	3.18 (-0.16)	+0.41**
Download	2.85	3.58 **(+0.73**)**	3.31 **(+0.46**)**	+0.27**
News	3.10	2.73 **(-0.37*)**	3.02 (-0.08)	-0.29*

A number of interesting findings can be concluded from Table 2: 1) Image verticals do not really bring more satisfaction to users. This is partly because information obtained from image verticals can usually also be easily obtained from non-vertical results. It is worth noting that off-topic images may result in a remarkable decline because irrelevant images are usually conspicuous and annoying. The fact that images bring more satisfaction to assessors may be because that assessors care more about search effort [22] and on-topic images may sometimes provide an instant answer to the query task, which will save a lot of time. 2) The encyclopedia vertical, as well as the download vertical, will bring more satisfaction for both users and assessors, and there is no significant decline even when the vertical result is irrelevant, which means such two kinds of vertical results are worth inserting into SERPs to improve the page quality. 3) Both on-topic and off-topic news verticals have no significant effect on users' satisfaction. On-topic news verticals will bring a significant drop in satisfaction for assessors, which may be because relevant news verticals may attract users to click them, leading to another search result page (in our experimental environment), which may be considered as a waste of time. 4) It is worth noting that encyclopedia verticals and news verticals have similar appearances (see Figure 1) but have completely different effects on satisfaction scores. This may be because encyclopedia verticals can provide users a more structured information with figures and texts describing the search target. In contrast, the figure provided by news verticals may be not so closely related to the search target and other non-vertical results can also provide as rich information as news vertical results.

Table 2 also shows some subtle differences between users and external assessors. On-topic image verticals will improve satisfaction scores from external assessors but have no significant effect on those from users, which may be because assessors will be more satisfied if the search task is finished in short time periods while users will probably be satisfied as long as their search need is met in not a very long time. Moreover, assessors will be dissatisfied when there are on-topic news verticals on SERPs while users will not, which indicates that assessors may be stricter with the wasted time caused by news verticals than users. All these findings indicate that assessors may care more about search effort, which is in line with the findings in [22].

4.3 Effect of Verticals at Different Positions

We further investigate the effect of verticals at different positions on satisfaction scores from both user's and assessor's perspectives. The results are shown in Table 3.

Table 3: Effect of Ranking Positions of Verticals on Satisfaction (* indicates statistical significance at p < 0.1 level,** indicates statistical significance at p < 0.05 level)

	w/o vertical	w/ on-topic vertical	w/ off-topic vertical	on-off difference
Users' Satisfaction Feedback				
Rank 1	4.79	5.06 **(+0.27**)**	4.43 **(-0.36**)**	+0.63**
Rank 3	4.79	4.93 (+0.14)	4.63 (-0.16)	+0.29**
Rank 5	4.79	4.87 **(+0.08*)**	4.85 (+0.06)	+0.02
External Assessors' Satisfaction Annotation				
Rank 1	3.24	3.48 **(+0.24*)**	3.09 (-0.15)	+0.39**
Rank 3	3.24	3.46 **(+0.22**)**	3.31 (+0.07)	+0.15
Rank 5	3.24	3.37 (+0.13)	3.27 (+0.03)	+0.10

Verticals are placed in three different positions, namely, rank 1, rank 3 and rank 5. Values in Table 3 are organized in a similar form with those in Table 2. Different rows show average Z-scores on pages with verticals at different positions. We can see that on-topic verticals bring significant improvement to satisfaction for both users and assessors when placed at rank 1. The effect is not so significant when verticals are inserted at rank 3 or rank 5. This encourages us to insert on-topic verticals at a very top position so that they may help improve the search experience of users. Meanwhile, an off-topic vertical reduces users' satisfaction the most significantly if inserted at rank 1. Off-topic verticals at rank 3 and 5 do not exert any significant influence.

5. EFFECT OF VERTICALS ON BENEFIT AND COST

According to our experimental results and [19], users' search satisfaction may be greatly affected by the benefit they obtain from the search result page and the cost during the information searching process. In this section, we first discuss some evaluation metrics from the perspective of benefit and cost and then investigate how verticals take effect in these metrics. In this way, we try to obtain a deeper insight into the effect of vertical results.

5.1 Estimation of Users' Benefit and Cost

5.1.1 Metrics with Cumulative Gain

We use cumulative gain (sCG) and normalized discounted cumulative gain (NDCG@N, N=3, 5, 10 in our case) [18] as a measure of search result page outcomes, which is also used in [19].

$$sCG = \sum_{r_i \in SERP} Rel(r_i) \tag{2}$$

$$nDCG = \sum_{r_i \in SERP} \frac{Rel(r_i)}{log(i+1)} (n = 3, 5, 10) \tag{3}$$

In Equation 2 and 3, r_i is the ith result on the corresponding SERP and $Rel(r_i)$ is its relevance score. We invite three professional assessors from a commercial search engine company to label a four-point-scaled relevance score for all query-result pairs in our experiment. The KAPPA coefficient of their annotation is 0.73, which can be characterized as a substantial agreement.

5.1.2 Metrics Based on Fixation Data

With the eye movement information we collected during our experiment process, we can exactly figure out users' examined result list and thus measure the search benefit from users' perspective. We use $Rlist$, which represents the examined result list obtained from the eye movement data, to replace the list of all results on SERPs used in equation 2 and 3 to obtain another group of metrics for search benefit. Notice that it is different from metrics in section 5.1.1 as we only sums up the relevance scores of those results which have been exactly examined by users instead of all results on SERPs.

We set the examination threshold to 200 milliseconds, which is recommended in [24, 29]. We also tried a number of other thresholds ranging from 100 ms to 1000 ms, and the results were quite similar. We regard those results with an eye fixation time of more than 200 ms, as examined, and thus we obtain the users' examined result list.

5.1.3 Metrics with Users' Cost

The metrics used to evaluate the cost users spend while examining the search result page can be obtained from two different ways. The first group of metrics, namely, search dwell time, maximum clicked rank and number of clicks, is obtained from users' mouse movement log and click-through data. These three metrics are widely adopted in existing search satisfaction works [16, 19] and are demonstrate to be effective at measuring search cost. The second group of metrics is based on the eye movement data. We can obtain the number of examined results as well as the length of the examined result sequence (note that a user may examine one particular result more than once, and in such case, it will be counted for multiple times). The examination threshold used here is still 200 ms.

5.2 Effect of Verticals on Search Benefit and Cost

In this section, we investigate how verticals affect search satisfaction following the benefit-cost framework. Due to the restriction of space, we only select out some typical metrics discussed in Section 5.1 as an example, including (**SCG**), defined by Equation 2, length of examined result sequence(**ESL**), described in Section 5.1.4, and SCG of examined results (**ESCG**), defined by Equation 7. We take these metrics as examples because they are representative, reflecting the quality of SERP (SCG) or the examination behavior of users (ESL and ESCG). The situations of the other metrics are similar with these selected ones in general.

Table 4: Correlation of Measures with Satisfaction(all correlation values are statistically significant: $p < 0.05$)

		SCG (Benefit)	ESCG (Benefit)	ESL (Cost)	SCG / ESL (Benefit / Cost)	ESCG / NER (Benefit / Cost)
non-vertical pages	Pearson	0.27	-0.42	-0.51	0.18	0.31
	Kendall	0.12	-0.17	-0.19	0.20	0.14
with on-topic verticals	Pearson	0.16	-0.42	-0.49	0.36	0.30
	Kendall	0.09	-0.26	-0.30	0.28	0.24
with off-topic verticals	Pearson	0.21	-0.39	-0.48	0.31	0.28
	Kendall	0.12	-0.22	-0.28	0.28	0.19

Table 4 shows the correlations between these evaluation metrics and user satisfaction on SERPs with on/off-topic verticals or without verticals. The results show that satisfaction has a weak positive correlation with SCG, a weak negative correlation with ESL, and a relatively strong positive correlation with and SCG/ESL. Such findings are similar with that in [19]. It is worth noting that there is a moderate negative correlation between ESCG and satisfaction, which is different from that between SCG and satisfaction. This is reasonable because SCG is just a metric that measures the result quality of a certain SERP and has nothing to do with users' examination behavior. Usually, the better the SERP is, the more satisfied the user will feel. However, ESCG calculates the information gained from the results examined by the user, and a higher ESCG may come from the examination of more search results, which usually means more cost and thus will probably result in less satisfaction. The correlation between satisfaction and ESCG per examined result (ESCG / ESL) is positive, which means a user will be more satisfied if he can get more information with less effort. The results in Table 4 also show that the correlations for different types of verticals share similar trends but there are still some subtle differences. We will provide some insight by showing detailed distributions of these metrics across different types of verticals.

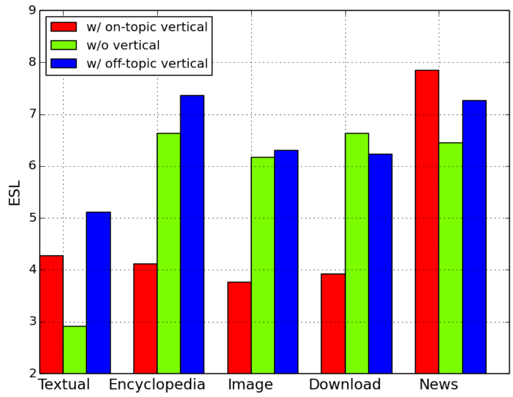


Figure 4: Distribution of ESL across verticals with different presentation styles

Figure 4 shows the effect of verticals on ESL, which is obtained from the eye movement data and can be a signal of search cost. From this figure, we can see that on-topic image, encyclopedia and download verticals can reduce the number of examined results remarkably, which may be because such vertical results can usually provide instant answers or a direct download link. An on-topic textual vertical or news vertical will not reduce the search cost, which may imply that such type of verticals can hardly improve users' search efficiency. We assume that this is because a textual vertical is usually just a combination of information from several perspectives. Additionally, a news vertical just leads the user to another list of news search results, which may not be satisfying. It is worth noting that nearly all types of off-topic verticals will increase the number of examined results, which indicates that we should be careful not to put low quality verticals on SERPs to reduce uesrs' cost.

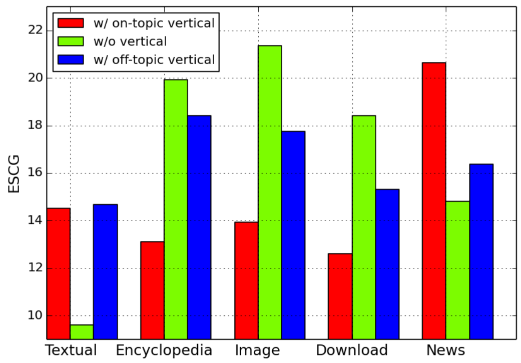


Figure 5: Distribution of ESCG across verticals with different presentation styles

Figure 5 shows the effect of verticals on ESCG, which is regarded as a metric of search benefit. The results show that ESCG is lower when there is an on-topic vertical result of encyclopedia, image and download. This is reasonable because a relevant vertical may be good enough to finish the search task and thus may reduce the number of examined results, which may then lead to a decline in ESCG. Relevant textual and news verticals may cause remarkably high SCG because they may not be useful to users most of the time and the high relevance of such results may become a waste. The results in Figure 5 indicate that ESCG may not be a positive estimator for satisfaction because the it is also affected by search cost. To verify this, we further develop another metric by dividing ESCG by ESL; the result is shown in Figure 6.

The results in Figure 6 show that for the case of on-topic download verticals, ESCG/ESL is remarkably higher that on SERPs with off-topic verticals and without verticals, which means on-topic download verticals help improve users' search satisfaction and is in line with the findings in Section 4.2. The differences between ESCG/ESL on SERPs with other four kind of verticals and without verticals are not significant, which is slightly different from the findings in Section 4.2. This may imply that a

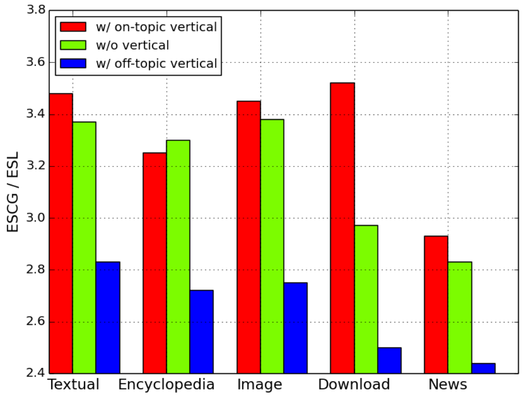


Figure 6: Distribution of ESCG / ESL across verticals with different presentation styles

evaluation metric defined as benefit divided by cost is still not perfect to model user satisfaction and a more appropriate fitting function is needed. We leave this for future work. Nevertheless, it is easy to find that all types of off-topic verticals will result in a significant reduction in ESCG/ESL, which implies adding off-topic verticals to SERPs may not be a good idea.

Findings in this section show that user satisfaction can be studied in the benefit-cost framework and that sometimes metrics generated by dividing benefits with costs can better estimate user satisfaction, which is the same with the findings in homogeneous environment [19]. Meanwhile, more suitable metrics are needed in the future to better model search satisfaction. In the next section, we will predict satisfaction scores from both users and external assessors with the collected information and compare the predictive power of features from different resources.

6. SATISFACTION PREDICTION FOR SERPS WITH VERTICALS

Although there are plenty of existing studies [1, 12, 15] in the prediction of search satisfaction, none of them take the existence of vertical results into consideration. According to the findings in Sections 4 and 5, vertical results have important impacts on the satisfaction judgements for both users and external assessors. Verticals also affect the cost and benefit of users while completing Web search tasks. Therefore, it is necessary to investigate how we can predict the satisfaction of users while facing SERPs with verticals. Recent studies in [19] and findings in Section 5 encourage us to predict satisfaction with features in a benefit-cost framework. The difference between our proposed method and the existing solutions in [19] is that we also focus on the effect of verticals in addition to other interaction behaviors.

Table 5 shows the vertical-aware feature sets adopted in the prediction of search satisfactions. There are two major information sources. One is information related to SERP itself as well as mouse log information, which can both be be easily collected at large scale. All these featues we used here are directly related to vertical results and we denote them as Click-Through features. The other is eye movement information, which is proven to have strong ability in predicting search performance [6, 11, 13, 27] but will be hard and expensive to collect in practice. We denote them as Eye-Tracking features. Recent studies [14] also show that other interaction behaviors, such as mouse movements, can be good substitutes for eye tracking information. We try to compare the predictive power of these two different sources of information to see to what extent we can predict search satisfaction.

Table 5: Vertical-aware Features for Predicting Satisfaction

Feature	Feature Description
Click-Through Features	
v_style	the presentation style of the vertical result, integer variable, ranging from 0 to 5, refers to 6 types of verticals (including pages without verticals)
v_position	the rank position of the vertical result, values can be 1, 3, 5 and 0 (0 is used when there is no vertical result)
arr_time	$(t\_ve - t\_s)/(t\_e - t\_s)$, where t_ve refers to the time when mouse arrives at the vertical result, t_s refers to the time when search session starts and t_e refers to the time when search session ends
if_click	whether the vertical result is clicked
click_time	$(t\_v - t\_s)/(t\_e - t\_s)$, where t_v refers to the time when the vertical result is clicked
aft_time	$(t\_e - t\_fv)/(t\_e - t\_s)$, where t_fv refers to the time when the user finished examining the vertical result. This feature will be assigned as $(t\_e - t\_s)$ if no vertical result is clicked.
hover_time	the mouse hover length (in second) on the vertical result
v_dwell_time	dwell time (in second) on the vertical result landing page
if_other_click	this feature will be $True$ if there is any other result click after the vertical result is clicked, otherwise it will be $False$
Eye-Tracking Features	
exam_num	number of examined results
exam_seq_len	the length of the examined result sequence
fix_arr_time	$(e\_ve - t\_s)/(t\_e - t\_s)$, where e_ve refers to the time when eye fixation arrives at the vertical result
fix_time	eye fixation length (in second) on the vertical result

The data set described in Section 3 is adopted in the prediction with five-fold cross validation. The learning algorithm in the prediction process is ridge regression, which is widely used in prediction tasks with continuous values. This prediction model has a penalty regarding the size of coefficients, which may help avoid the problem of over-fitting. We implement the satisfaction prediction method in Guo et al.[15] as the baseline method (**Baseline_1**) because it is a state-of-the-art method based on fine-grained mouse behavior data for predicting web search success. The predictive model in [19] is also used here to be a baseline (**Baseline_2**) because the features are extracted in a benefit-cost framework and can estimate graded search satisfaction more accurately than most existing works in homogeneous environment. Note that in our experiment, there is only one query in a search task. So any feature that is related to multi-queries is not included in the implementation. We use the Normalized Root Mean Square Error (NRMSE, ranging from 0 to 1, smaller values

Table 6: Comparison of Different Methods for Predicting Search Satisfaction

Features	NRMSE (Predicting User Satisfaction)		NRMSE (Predicting External Annotation)	
	Baseline_1	Baseline_2	Baseline_1	Baseline_2
original	0.235	0.233	0.130	0.140
original + click-through	0.23	0.225	0.135	0.147
original + eye-tracking	0.226	0.205	0.128	0.129
original + click-through + eye-tracking	0.222	0.199	0.131	0.133

mean better prediction) to evaluate the model performance as in most continuous value regression tasks [19]. The prediction performance of different feature groups is shown in Table 6.

The results in Table 6 show a number of interesting findings: 1) The prediction performance for annotations from external assessors is much better than those from users for all types of feature combinations. This probably means that the annotations from the assessors' side are more objective and consistent with each other. 2) The involvement of vertical-aware features significantly improves the prediction performance in both types of feature types. This shows the effectiveness of vertical information in predicting search satisfaction on SERPs with heterogeneous vertical results. 3) Among the two sources of features, we can see that eye-tracking features perform better than page-log features, and we achieve the best performance when all feature groups are used for predicting user satisfaction. 4) Methods based on Baseline_1 [15] perform better when predicting external annotations while methods based on Baseline_2 [19] perform better when predicting user satisfaction. This indicates that features in the benefit-cost framework can make better use of users' search information and can estimate user satisfaction more accurately. 5) Finally, we can also realize that when predicting external satisfaction annotations, the improvement of our proposed features is not as remarkable as that when predicting user satisfaction. This probably reflects the fact that our proposed method is more helpful when predicting user satisfaction and may be more suitable for practical applications.

7. CONCLUSIONS

Satisfaction prediction is an important research issue in the evaluation of search engine performance, and satisfaction studies on federated search pages have not been pursued. In this paper, we carry out a lab-based user study with specifically designed aggregated search result pages to see how verticals affect search satisfaction. We collect satisfaction scores from both users and external assessors to make a comparison because the concept of satisfaction is quite subjective. We find that on-topic Encyclopedia verticals as well as Download verticals will bring significant improvement to search satisfaction and that even off-topic ones will not result in a significant decline. Good Image verticals may not bring too much improvement but irrelevant ones will bring significant dissatisfaction. Most users will not care about News verticals because they only provide another list of search results. Good news verticals sometimes will even bring negative effects because a click leading to another SERP may be considered a waste of time. Verticals will have the largest influence when they are presented at the top of a page. As the position of verticals becomes lower, the effectiveness will decline to a large extent. With the rich information collected in our experimental system, we demonstrate that a benefit-cost framework will be useful when analyzing satisfaction. We find that vertical results will significantly affect both result benefits and search costs. We also conclude that a metric of dividing benefit by cost will be a more appropriate estimator of user satisfaction. Finally, we proposed a learning-based framework to predict search satisfaction from both users and external assessors. We verify that with features related to verticals, we can predict search satisfaction much better than our baseline methods, which are effective in homogeneous environment. We demonstrate that features extracted from the perspective of both benefit and effort can greatly improve the user satisfaction prediction model. We also conclude that satisfaction scores from external assessors are easier to predict probably because professional assessors are usually more objective while judging satisfaction. Interesting directions for future work include understanding and predicting satisfaction on SERPs with multi-verticals. Moreover, incorporating the effect of verticals into user behavior models will also be a challenge.

8. REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: a game for modeling different types of web search success using interaction data. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 345–354. ACM, 2011.

[2] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between ir effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 773–774. ACM, 2007.

[3] J. Arguello and R. Capra. The effects of vertical rank and border on aggregated search coherence and search behavior. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 539–548. ACM, 2014.

[4] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 315–322. ACM, 2009.

[5] J. Arguello, F. Diaz, and J.-F. Paiement. Vertical selection in the presence of unlabeled verticals. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 691–698. ACM, 2010.

[6] G. Buscher, S. T. Dumais, and E. Cutrell. The good, the bad, and the random: an eye-tracking study of ad quality in web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 2010.

[7] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 621–630. ACM, 2009.

[8] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM, 2008.

[9] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[10] F. Diaz. Integration of news content into web results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 182–191. ACM, 2009.

[11] S. Djamasbi, A. Hall-Phillips, and R. R. Yang. Serps and ads on mobile devices: An eye tracking study for generation y. In *Universal Access in Human-Computer Interaction. User and Context Diversity*, pages 259–268. Springer, 2013.

[12] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 34–41. ACM, 2010.

[13] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *SIGIR'04*, pages 478–479. ACM, 2004.

[14] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. In *CHI'10*, pages 3601–3606. ACM, 2010.

[15] Q. Guo, D. Lagun, and E. Agichtein. Predicting web search success with fine-grained interaction data. In *CIKM'12*, pages 2050–2054. ACM, 2012.

[16] Q. Guo, R. W. White, S. T. Dumais, J. Wang, and B. Anderson. Predicting query performance using query, result, and user interaction features. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 198–201, 2010.

[17] S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 567–574. ACM, 2007.

[18] K. Järvelin, S. L. Price, L. M. Delcambre, and M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *Advances in Information Retrieval*, pages 4–15. Springer, 2008.

[19] J. Jiang, A. H. Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. 2015.

[20] J. Jiang, D. He, and J. Allan. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. 2014.

[21] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1—2):1–224, 2009.

[22] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *SIGIR'15*. ACM, 2015.

[23] Z. Liu, Y. Liu, K. Zhou, M. Zhang, and S. Ma. Influence of vertical result in web search examination. In *SIGIR'15*. ACM, 2015.

[24] L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, 59(7):1041–1052, 2008.

[25] I. Markov, E. Kharitonov, V. Nikulin, P. Serdyukov, M. de Rijke, and F. Crestani. Vertical-aware click model-based effectiveness metrics. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1867–1870. ACM, 2014.

[26] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 953–964. International World Wide Web Conferences Steering Committee, 2013.

[27] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530. ACM, 2007.

[28] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1043–1052. ACM, 2011.

[29] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. ACM, 2000.

[30] L. T. Su. Evaluation measures for interactive information retrieval. *Information Processing & Management*, 28(4):503–516, 1992.

[31] L. T. Su. A comprehensive and systematic model of user evaluation of web search engines: Ii. an evaluation by undergraduates. *Journal of the American Society for Information Science and Technology*, 54(13):1193–1223, 2003.

[32] S. Verberne, M. Heijden, M. Hinne, M. Sappelli, S. Koldijk, E. Hoenkamp, and W. Kraaij. Reliability and validity of query intent assessments. *Journal of the American Society for Information Science and Technology*, 64(11):2224–2237, 2013.

[33] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. Incorporating vertical results into search click models. In *SIGIR'13*, pages 503–512. ACM, 2013.

[34] H. Wang, Y. Song, M.-W. Chang, X. He, A. Hassan, and R. White. Modeling action-level satisfaction for search task satisfaction prediction. 2014.

[35] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating reward and risk for vertical selection. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2631–2634. ACM, 2012.