The Influence of Presentation and Performance on User Satisfaction

Kanaad Pathak
kanaad.pathak@strath.ac.uk
University of Strathclyde
Glasgow, UK

Leif Azzopardi
leif.azzopardi@strath.ac.uk
University of Strathclyde
Glasgow, UK

Martin Halvey
martin.halvey@strath.ac.uk
University of Strathclyde
Glasgow, UK

ABSTRACT

Information Retrieval (IR) systems are designed to provide users with a ranked list of results based on their queries. The effectiveness of an IR system is gauged not just by its ability to retrieve relevant results but also by how it presents these results to users; an engaging presentation often correlates with increased user satisfaction. While existing research has delved into the link between user satisfaction, IR performance metrics, and presentation, these aspects have typically been investigated in isolation. Our research aims to bridge this gap by examining the relationship between query performance, presentation and user satisfaction. For our analysis, we conducted a between-subjects experiment comparing the effectiveness of various result card layouts for an ad-hoc news search interface. Drawing data from the TREC WaPo 2018 collection, we centered our study on four specific topics. Within each of these topics, we assessed six distinct queries with varying nDCG values. Our study involved 164 participants who were exposed to one of five distinct layouts containing result cards, such as "title", "title+image", or "title+image+summary". Our findings indicate that while nDCG is a strong predictor of user satisfaction at the query level, there exists no linear relationship between the performance of the query, presentation of results and user satisfaction. However, when considering the total gain on the initial result page, we observed that presentation does play a significant role in user satisfaction (at the query level) for certain layouts with result cards such as, title+image or title+image+summary. Our results also suggest that the layout differences have complex and multifaceted impacts on satisfaction. We demonstrate the capacity to equalize user satisfaction levels between queries of varying performance by changing how results are presented. This emphasizes the necessity to harmonize both performance and presentation in IR systems, considering users' diverse preferences. Ultimately, our insights can steer the evolution of more user-aligned IR systems, underscoring the balance between system performance and result presentation.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; • **Human-centered computing → Empirical studies in HCI**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0434-5/24/03
https://doi.org/10.1145/3627508.3638335

KEYWORDS

Information Retrieval (IR), User Satisfaction, Interface Layouts, Query Performance, Search Result Presentation, Empirical Study, Human-Computer Interaction, Retrieval Effectiveness, Search Interfaces,

ACM Reference Format:
Kanaad Pathak, Leif Azzopardi, and Martin Halvey. 2024. The Influence of Presentation and Performance on User Satisfaction. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3627508.3638335

1 INTRODUCTION

Information Retrieval (IR) systems, such as web search engines, aim to help users efficiently locate relevant information within vast collections of documents in response to queries. A critical aspect of an IR system's effectiveness lies in its ability to fulfil a user's information needs by retrieving documents relevant to the user's query. Retrieved documents are generally presented to users on Search Engine Result Pages (SERPs), and each result can typically be represented by a result card. A good result card aims to help the user make more effective decisions about exploring a given document by presenting on it, information such as a title, image or summary of the web page. Previous works from Kammerer and Gerjets [16], Rele and Duchowski [22], Teevan et al. [24] and Joho and Jose [15] have studied how the presentation of these result cards affects user satisfaction. The broad consensus from these analyses is that incorporating visual elements like images, links, and text summaries can strongly influence user satisfaction and perceptions of relevance.

However, it is not solely the presentation that drives user satisfaction. Users also spend their time creating queries so that the system may retrieve and present them with appropriate relevant documents for their information need. The performance of these queries is typically measured by system-side metrics such as Cumulative Gain (CG), Discounted Cumulative Gain (DCG), normalized Discounted Cumulative Gain (nDCG) etc. Work from Al-Maskari et al. [2] has explored how these metrics affect user satisfaction, finding that there is a strong correlation in most query performance metrics.

During the interaction process, users can also perform other actions such as inspecting various result cards, saving the documents behind the cards etc. These actions come with inherent costs to perform them and further work such as Azzopardi [3] have developed formal models to estimate the costs in the interaction process (such as cost to query, examine cards etc). Given this formal framework,

further research such as Azzopardi et al. [4], Morrison and Vancouver [19], Verma and Yilmaz [26] have studied how costs such as the cost to query affect user satisfaction.

Given that the presentation of the result cards can also affect user satisfaction, it is unclear how changing the presentation can affect both the system side costs (query costs) and user side costs (user satisfaction). Take, for example, two result lists with slightly differing nDCGs for a given query, presented in the same result card type (all titles). Findings from Al-Maskari et al. [2], Morrison and Vancouver [19], Verma and Yilmaz [26] would suggest that spending longer examining results for a query with a higher nDCG will lead to more satisfaction. However, if we modify the presentation of the result list with a lower nDCG to be presented with, say all titles and images (TI), users may now spend more time examining results in this list due to their changed presentation and thus feel a similar amount of satisfaction as that obtained from a result list with a higher nDCG.

Our study seeks to explore this interplay between query performance, presentation, and user satisfaction. To ground our analysis, we present findings from a crowd-sourced user study relating to an ad-hoc news search task. In our study, we examine five interface layouts with differing numbers of results per page with four distinct types of result cards. Utilizing topics, queries, and documents from the TREC WaPo 2018 corpus, participants were tasked with finding and marking relevant documents for two topics. We collected satisfaction ratings from users for each query, and subsequently for each result card layout (after they completed a topic). The primary research questions guiding our study are:

(RQ1) How do the quality of search results (as measured by query performance) and the interface layout impact user satisfaction in information retrieval tasks?

(RQ2) What are the effects of different interface layouts on user satisfaction as measured by overall satisfaction, the likability of the engine, productivity, and mental effort?

2 BACKGROUND

Several factors can impact user satisfaction with an information retrieval (IR) system. These factors can be thought of as costs and be measured system-side or user-side to determine the overall effectiveness of an IR system. The system side costs include standard IR effectiveness metrics such as precision, cumulative gain (CG), discounted cumulative gain (DCG) and normalized DCG (nDCG). Where, on the user side the costs can include the formulation of queries, the display of results, time to browse, user satisfaction etc. On the system side, studies such from Al-Maskari and Sanderson [1] and Al-Maskari et al. [2] have examined how IR effectiveness measures such as precision, CG, DCG and nDCG affect user satisfaction, and found that CG and precision have a better correlation with user satisfaction as compared to nDCG. The correlation with nDCG was weak due to having limited judgements. On the system side, experimentation has been conducted independent of the presentation of results. Studies on the user side have also considered the effect of presentation on user satisfaction, mainly along three dimensions. In the first dimension, studies such as Rele and Duchowski [22] and Kammerer and Gerjets [16] tell us that there are two main layouts in which results can be optimally presented to maximize user satisfaction. The first is a single list, and the second one is a grid layout. However, it is worth noting that both of these studies have found conflicting results on which layout is better. The second dimension explores the relationship between different result card formats and user satisfaction. Work done by Bota et al. [8], Dziadosz and Chandrasekar [11], Joho and Jose [15], Teevan et al. [24], Tombros and Sanderson [25] has explored standard cards which contain information such as the URL, text and images. The first two dimensions (layout and presentation format) come with an important trade-off on the space and utility of each result item, which is explored in the third dimension. For example, if results are presented as a list of ten blue links versus if we present them with an image and some summary text, the results will occupy different amounts of space on the screen, and then satisfaction to the user will largely depend on the visual appeal and informativeness of the result [17, 18, 23]. Models developed on economic search theory by Azzopardi and Zuccon [5] have been proposed which provide a framework to estimate the costs associated with user behaviour, and studies such as Jansen et al. [14], Verma and Yilmaz [26] have measured how costs associated with search, such as the length of the query, the number of viewed documents and clicked snippets affect user satisfaction. However, both on the system side and the user side, these costs have been studied independently of each other. That is to say; user satisfaction has not been studied in the context of the presentation of results and also standard IR metrics such as nDCG. The interplay between IR performance and presentation is still not well understood.

3 METHODOLOGY

To explore how query performance and result card layouts influence user satisfaction, we conducted a between-subjects study using simulated ad-hoc search tasks [7]. To position the information-seeking process within a structured context, participants were presented with a series of pre-picked queries, which were grouped into three categories based on their nDCG@10: low, medium, and high. Each category contained two queries. The task involved participants engaging in an exploratory search session, examining various queries and documents to find and pinpoint relevant examples within relevant documents related to the given topic. All documents were indexed and retrieved using our custom-built system, ensuring consistency across searches and presentation of results. The between-group variable in our study was defined by five distinct interface layouts. These layouts prominently featured cards consisting of titles, images, and summaries of news articles.

3.1 Collection and System

For this study, we used the TREC Washington Post Corpus (WaPo) collection from the TREC Common Core 2018 track [1]. The WaPo collection consists of 608,180 news articles and blog posts published between January 2012 and August 2017 categorized into 50 topics for information retrieval tasks. This collection provides a diverse range of topics for analysis and experimentation, allowing us to explore the effectiveness of our proposed approach across different topical themes. We used ***Whoosh***[2] (a pure python search engine

[1] https://trec-core.github.io/2018/

[2] $pip install whoosh==2.7.4

library) with BM25 ($b = 0.75$, $k_1 = 1.2$) to index and retrieve documents for a given query. We presented results on a SERP, as shown in Figure2(b). Our SERP view consisted of result cards in presentation formats of two major news sources (The Washington Post and Google News).

We chose five different types of interface layouts to show the participant, with the four different result cards shown in Figure 1.

(1) Title + Image + Summary [TIS]
(2) The Washington Post Style, Title + Image + Summary [TIS WaPo],
(3) Google News Style, Title + Image[TI]
(4) Title only [T]
(5) Random, a combination of the four above.

We consider our viewport to have a fixed amount of space (6 columns using bootstrap column widths and 12 rows, computed using approximately 100px per row). Thus, the total number of results shown on the page depended on the type of card and the number of rows it occupied. For example, on a single result page layout with our page constraint, there could either be approximately 12 T, 2 TIS cards, 6 TI Google Cards or 4 TIS WaPo cards.

We used the same CSS stylesheet as the Washington Post website to display all the results. The titles were set at a font size of 14pt, and the summaries were set at a font size of 12pt. Since the Washington Post website summaries consisted of the leading 250 characters of the news result (at the time of the experiment), the result summaries we displayed contained the leading 250 characters as well.

3.2 Search Topics and Tasks

From the 50 topics available in the TREC WaPo collection, we selected four topics for our study:

(1) **Topic 341:** Airport Security,
(2) **Topic 363:** Transportation Tunnel Disasters,
(3) **Topic 367:** Piracy at Sea and,
(4) **Topic 408:** Tropical Storms.

These topics were chosen based on their inclusion of at least 120 TREC Relevance Judgements and a minimum of 60 relevant documents with associated article images available for download. To maintain consistency in the presentation of result cards, we re-scaled each image to ensure uniform sizing.

Participants were instructed to find and save – different and relevant documents that they felt suited the relevance criteria for the given topic by exploring as many queries as necessary. For example, in topic 408 (see instructions in Figure 2(a),(b) and (c) on the left side), participants were asked to find a number of different tropical storms that caused widespread destruction and loss of life. Examples requested for the other topics were:

- Topic 341 the airport and security measures employed;
- Topic 363 the name of the tunnel and the cause of the disaster and,
- Topic 367 instances of piracy where vessels were boarded.

We generated 6 queries per topic using the techniques outlined in [18], and then stratified the resulting queries into three tiers based on their nDCG scores. Specifically, the queries were grouped into low (0.1-0.2), medium (0.2-0.6), and high (0.6+) nDCG categories.

3.3 Measures

We split the dependent variables in our study into three main categories: (a) search behaviours, (b) search experience and (c) performance:

Search Behaviours: To provide insights into user search behaviours we logged the number of...

(1) ...queries clicked
(2) ...pages viewed
(3) ...documents viewed
(4) ...documents saved

For relevant documents saved, we instructed the participant to record the relevant bits of the document into a text field that popped up if the user clicked on the "relevant" button on the document view page and saved it to our database. From the interaction logs, we could also compute the following time-based measures, including the time spent...

(1) ... to complete the task
(2) ... per result card (snippet)
(3) ... on a relevant document
(4) ... on a non-relevant document

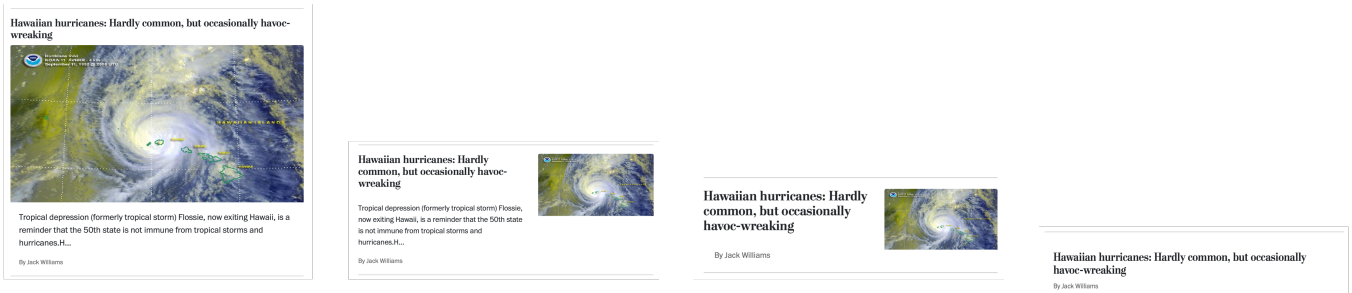
The relevance and non-relevance of a document were obtained using the TREC WaPo Qrels for the retrieved documents. One thing to note is that, in our document index, we only indexed documents which had TREC relevance judgements. For time spent on a snippet, we use aggregated mouse hover times as a proxy for eye gaze [9, 12, 13, 20, 21] computed with a modified lightweight JavaScript code [6].

Search Experience: We measured the search experience of the participant through a user satisfaction score. We collected user satisfaction at two levels: (a) the query level (collected after changing a query) and (b) the interface level (collected after every topic/task). For (a) query satisfaction, we collected data using a 6-point Likert scale by asking participants how satisfied they were with the results for that given query (with 1 being very dissatisfied to 6 being very satisfied). For (b) interface satisfaction, we asked participants whether they ...

(1) ...felt **productive** using the system
(2) ...found the interface layout to be **mentally taxing**
(3) ...found the interface layout to be **engaging**
(4) ...found the interface layout to be **distracting**
(5) ...were **satisfied** with the interface layout,

on a 6-point Likert scale with 1 being strongly disagree and 6 being strongly agree.

Performance: By using the TREC Common Core 2018 relevance judgements, we were also able to provide an estimate of search performance at the (a) system side and (b) user side. On the system side, for each query that was submitted by a participant, we evaluated the query's nDCG@10 and Total gain on the Page (see §4.2 for further detail). For the user-side performance measures, given all of the documents that participants clicked on and saved, we could use the aforementioned relevance judgements as ground truth, allowing us to compute the accuracy of a participant's searching ability. This was summarised as the proportion of correctly identified relevant items saved (i.e., documents that are identified as relevant in the relevance judgements) vs. the total number saved.



(a) Title+Image+Summary (TIS): ~6 rows **(b) Title+Image+Summary (TIS WaPo): ~3 rows** **(c) Title+Image (TI Google): ~2 rows** **(d) Title Only (T): ~1 row**

Figure 1: Example of the different result card types, with an approximation of the number of rows each card type occupies.

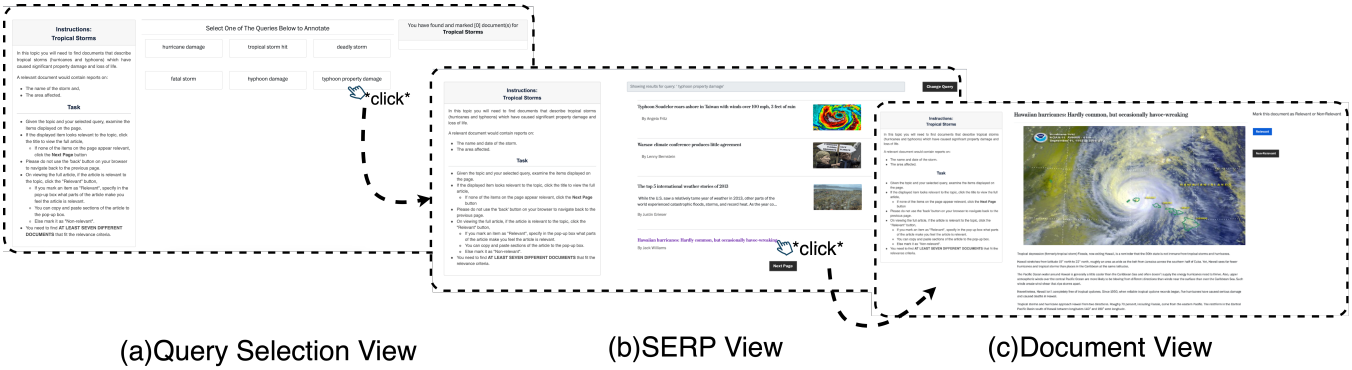


Figure 2: An example of the user interface presented to participants for collection of annotations. Sub-figure (b) shows an example of a SERP layout with a random arrangement of cards.

3.4 Procedure

Participants were recruited from the online crowd-sourcing platform Prolific [3]. Participants were also pre-screened based on their first language; all participants indicated a native speaker proficiency in English at the time of undertaking the experiment. This was done to maintain consistency across the participant's ability to carry out the task accurately. Four different pages were created on Prolific to fill participants for each topic. Each page contained the link to complete the task using the topic specified for that page. Before participating in the study, participants were presented with an on-screen information sheet detailing the procedure of the study. They were required to provide their informed consent before proceeding with the study. Upon successful completion of the study, participants received the equivalent of USD$7 for their time, which falls in line with minimum payment requirements. Each participant was randomly allocated one of the five layouts when they began the experiment.

The goal of the experiment was to complete a news search task based on a given topic. Participants were asked to find and mark documents relevant to one of the selected topics by exploring a set of pre-defined queries as described in § 3.2. When the participants began the experiment, they were presented with a list of six queries in a 3x2 grid that corresponded to the topic they selected. The order in which these queries were presented was randomized. An example of this query selection grid can be observed in Figure 2(a). Participants were instructed to choose any query to inspect and explore the associated results with that query to find the relevant documents. Participants were asked to evaluate the relevance of the documents based on the criteria provided on the left of the screen in a floating instruction box. An example of this floating instruction box can be seen in Figure 2(a),(b) and (c). These instructions were continuously visible during the process of the experiment. Once a participant picked a query, they were shown all the documents associated with that query in one of the layouts, in the style of a SERP.

The ordering of the relevance of results was random in all layouts. In figure 2(b) we can see on the SERP how results were presented for a random layout, we can observe the results presented as TI, T, TIS WaPo and T. Pagination was made available via a button at the bottom of the screen to move to the next set of results for a query. The participant could click on any result card to inspect the document behind it in further detail. Upon inspecting a card, the full contents of the document were displayed on a new page, this can be seen from Figure 2(c). If a participant inspected a result card and found the document relevant, they were asked to provide instances of the document that made it relevant in a pop-up text area. Participants were asked to provide at least one instance per relevant document.

[3]https://www.prolific.co

When a participant moved between queries of the same topic, we collected the query satisfaction. In the query selection view, on the right side, we displayed the titles of the documents that the participant had marked as relevant, along with what section they marked within that document so that participants could quickly glance at their task progression. The participant needed to inspect at least two queries and find seven different relevant documents before finishing the topic. When participants finished one topic we collected the interface satisfaction as described in § 3.3, and then the second topic was randomly assigned to them (from the pool of three remaining topics) with the same result layout as the first topic.

3.5 Participant Demographics

Participation was completely remote, with the researchers not interacting with any participant in any capacity. Participants directly interacted with the web application designed to collect interaction data.

The user study involved 164 participants, most of whom fell in the age range of 20 to 40 years old, with a mix of students (27) and non-students (137). The majority of participants were employed, with 122 reporting full-time or part-time work, while the remaining participants were not engaged in paid work, such as homemakers, retired, or disabled individuals.

3.6 Ethics Approval

A departmental review board approved the study (ethics no 2027). We strictly followed ethical guidelines and ensured that every participant gave informed consent. All participants received a thorough explanation of the study's procedures, potential risks, their rights, and the option to leave at any point. The consent form also provided a link to the ethics application approval.

4 EXPERIMENTAL RESULTS

4.1 Summary of Search Behaviours

Comprehensive data analysis examined differences in task completion rates, interaction times, and other user metrics, such as the number of queries, clicks, and time spent across various interface layouts. Welch's ANOVAs was used to assess whether significant differences existed between the conditions and the measures under investigation. The primary effects were analyzed at a significance level of $\alpha = 0.05$. Pairwise Games-Howell tests were utilized for post-hoc analyses. For the reported tests, the F-score, p-value, and effect size η_p^2 are presented to two decimal places. The ranges of η_p^2 values correspond to small (< 0.06), medium (0.06 - 0.14), and large (> 0.14) effect size [10]. The ± values reported in the tables denote the mean and standard deviation.

Table 1 reports the average search behaviours of users for each interface layout, detailing the number of actions performed per topic, per query. Incorporated in this analysis is the accuracy measure, highlighting how well participants identified relevant documents from the non-relevant (i.e., the proportion of relevant documents saved versus the total number saved).

Considering the varied interface layouts, there is evident consistency in user behaviours. Across the board, for any topic, participants on average clicked to view 3 to 4 queries. Notably, participants examined on average only a single page for every query they issued. This is despite the fact that users could examine more pages within the same query. For every query viewed, participants clicked and viewed between 4 to 5 documents. They saved about 3 of the viewed documents, and out of these, they correctly identified around 2 as relevant. The accuracy of judgements fluctuated between 0.75 to 0.83.

We found that with the TIS WaPo layout (when all results were presented with TIS WaPo cards) participants were able to more accurately identify and mark relevant documents, achieving a peak accuracy of about 0.83, which was significantly more than other layouts ($F(4,441.837) = 2.51$, $p = 0.04$, $\eta_p^2 = 0.01$). This is possibly due to TIS WaPo cards providing useful information in the form of a summary that helped users to click and accurately mark them as relevant. However, it is interesting to note that this was significantly higher than the TIS layout, in which the result cards contained the same information but occupied more space. We hypothesize that this occurs due to the ability to view more cards containing summaries within the same space, potentially expanding the context window of users viewing the result cards. However, on average, per query and topic, we found no statistically significant differences in the search behaviours of participants across any layout.

Due to the synthetic nature of our queries and the controlled nature of our study, we hypothesize that these behaviours may be specific to our study and that in a more naturalistic search scenario, where users can type out queries, they may tend to issue queries differently to find relevant information. This could further impact other factors such as the time spent examining documents and inspecting pages on the SERP.

Table 2 offers a comprehensive look at the average timings for the search behaviours (in seconds) participants took for various actions during their search sessions. Firstly, in general, we observe that there is no statistically significant difference between the times that users took to complete the task (topic). Participants took on average approximately 20 minutes to annotate a topic. The time spent on a snippet in a layout was computed as the amount of time users spent hovering over results. We found significant differences between all layouts ($F(4,9579.212) = 34.306$, $p < 0.001$, $\eta_p^2 = 0.01$). We found no notable differences in the time required to read and make a decision for a relevant or non-relevant document for any given interface layout. Participants spent an average of 43 seconds to read the document and decide the relevance.

4.2 RQ 1: How do the quality of search results (as measured by query performance) and the interface layout impact user satisfaction in information retrieval tasks?

We ran two ordered models on 1,398 observations from 164 participants to scrutinize the association between query performance, presentation and user satisfaction. We were concerned with examining two main metrics to measure query performance. (1) nDCG@10 and (2)Total gain of the first result page.

Within our models, we also explored several interaction effects, such as the interplay between the topic and interface layout, the sequence in which the topic was completed (referred to as "topic order", meaning if the topic was completed as the first or second),

Table 1: Search behaviours, with the mean number of actions performed per user, per topic, per query. Here, Q denotes Queries, $Docs$ denotes documents. R and $\bar{R}$ denote relevant and non-relevant. Highest accuracy values are bolded

Interface Layout	#Q	#Docs viewed	#Pages	Documents Saved		Accuracy
				#$Docs$	#R $Docs$	
a. TIS	4.12±2.19	4.71±4.03	1.05±0.22	2.71±2.24	2.19±1.69	0.79±0.27
b. TIS WaPo	4.34±3.00	4.65±3.73	1.12±0.39	3.31±2.43	2.69±2.01	**0.83±0.22**
c. TI Google	4.17±1.83	4.94±5.14	1.01±0.00	2.97±2.25	2.33±1.65	0.79±0.26
d. T	3.91±2.33	5.22±4.67	1.05±0.24	2.94±2.15	2.40±1.74	0.75±0.28
e. Random	4.17±2.04	4.63±4.31	1.02±0.13	2.89±2.96	2.38±2.21	0.78±0.25

Table 2: Average timings for various search behaviours actions during the study, per user, per topic, per query. The timing data is in seconds. Asterisks (*) denote a significant difference between all groups($p < 0.05$)

Interface Layout	Task	Time per ...		
		Snippet	R Doc	$\bar{R}$ Doc
a. TIS	1345.23 ± 670.13	2.25 ± 1.23*	44.13 ± 41.90	37.92 ± 30.28
b. TIS WaPo	1469.89 ± 1138.38	2.09 ± 1.25*	52.34 ± 45.74	36.94 ± 33.55
c. TI Google	1442.04 ± 931.42	1.82 ± 1.18*	41.24 ± 36.48	34.44 ± 36.26
d. T	1519.73 ± 1027.33	1.95 ± 1.18*	42.36 ± 40.77	52.29 ± 65.62
e. Random	1367.29 ± 882.38	2.11 ± 1.27*	42.97 ± 43.16	36.00 ± 29.78

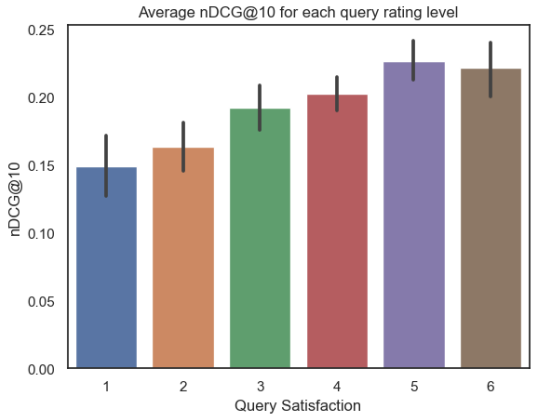


Figure 3: The relationship between query satisfaction and nDCG@10

and the relationship between the interface layout and query performance. Equation 1 shows the independent variables in our ordered model alongside the coefficients β. For analyses focusing on the total gain on the first page, we adjusted the equation by substituting the β_1 coefficient.

$$\begin{aligned} Y_{ij} = \beta_0 &+ \beta_1(\text{nDCG@10})_{ij} + \beta_2(\text{Topic Order})_{ij} \\ &+ \beta_3(\text{Topic ID})_{ij} + \beta_4(\text{Interface Layout})_{ij} \\ &+ \beta_5(\text{Topic ID} \times \text{Topic Order})_{ij} \\ &+ \beta_6(\text{Topic ID} \times \text{Interface Type})_{ij} \\ &+ \beta_7(\text{Interface Type} \times \text{nDCG@10})_{ij} \\ &+ b_{0j} + (1|\text{user})_j + \epsilon_{ij} \end{aligned} \tag{1}$$

As we can observe from the top half of Table 3, we found a significant positive relationship between nDCG@10 and user satisfaction, which can also be observed from Figure 3. We observed no interaction effects between the nDCG@10 and the interface layout which could have affected the query satisfaction. This signifies that a poorer nDCG of a query cannot increase user satisfaction to match the same level as that of a higher nDCG if we change the presentation of results. However, we observed a significant effect on query satisfaction when users attempted Topic 408 with the Random layout as the second topic.

For the total gain on the first page, we examined the effectiveness of each query within the context of the first page of results, utilizing the metric NDCG@k. Recognizing the dual significance of result relevance and quantity, we used a 'total gain' measure for the first page of results. This measure was calculated by multiplying the NDCG@k score, which evaluates the relevance of the documents on the first page, by the number of results (k) displayed on that page. By doing so, this 'total gain' measure accounts for both the quality and quantity of results, providing a more holistic assessment of query performance on the first page of results for across multiple queries. This allows us to factor in the varying number of results displayed by different interface layouts, and understand how these layouts perform not just in terms of relevance per document (as captured by NDCG@k), but also in terms of total relevance gain for the user across multiple queries. We can also observe this similarly positive relationship for the total gain from Figure 4

The second ordered model was run with the formula defined in Equation 1, but with the β_1 parameter being substituted for the total gain on page 1. The results from this second model (as shown in the bottom half of Table 3) showed that the interaction effects between the total gain and the interface layouts consisting of TI Google and TIS WaPo cards played a significant effect ($p < 0.05$)

Table 3: Results of the Ordered Model analysis on query satisfaction, where p-value was statistically significant for the β parameter. The category differences were all significant.

Beta Parameter	Coeff.	SE	z-value	p-value	95% CI	
β_1(nDCG@10)	2.3015	0.897	2.566	0.010	0.543	4.06
β_5(TOPIC 408 × Order = 2)	0.8355	0.289	2.891	0.004	0.269	1.402
β_6(TOPIC 408 × Interface Layout = Random)	1.5771	0.452	3.490	<0.001	0.691	2.463
β_1(Total Gain on Page 1)	0.2002	0.079	2.542	0.011	0.046	0.355
β_3(TOPIC 408 × Interface layout = Random)	1.5491	0.451	3.434	0.001	0.665	2.433
β_5(TOPIC 408 × Order = 2)	0.8471	0.289	2.936	0.003	0.282	1.413
β_7(Interface Layout = TIS WaPo × Total Gain on Page 1)	0.5173	0.186	2.778	0.005	0.152	0.882
β_7(Interface Layout = TI Google × Total Gain on Page 1)	0.3024	0.152	1.990	0.047	0.005	0.600

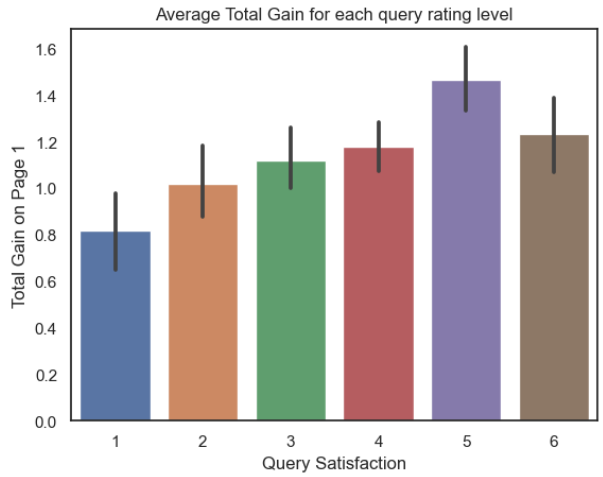


Figure 4: The relationship between query satisfaction and Total Gain on Page 1

on the query satisfaction. Same as with our first ordered model, we observed a significant effect on query satisfaction when users attempted topic 408 with the random layout as the second topic. Our findings from this model essentially indicate that for total gain on the first page, the presentation of results can affect user satisfaction (i.e., by modifying the interface layout, a layout with a lesser total gain on the first page can attain user satisfaction comparable to a layout with a higher total gain.)

Our results on the link between nDCG@10 and user satisfaction diverge slightly from past studies, such as Al-Maskari et al. [2], which found only weak ties between nDCG and satisfaction[4]. We identified strong linear correlations between nDCG@10 and query satisfaction. Additionally, we noted distinct gains on the first page for two layouts, TI Google and TIS WaPo, revealing an interplay between result presentation, total gain, and query satisfaction. In conclusion, while nDCG@10 effectively predicts user satisfaction, no direct linear relationship exists between result presentation and user satisfaction for metrics like nDCG@10. However, metrics like total gain on the first page do influence presentation and satisfaction.

[4]We also compared other metrics from the Al-Maskari et al. [2] study such as precision and CG and confirmed that precision and CG are strongly correlated to user satisfaction ($p < 0.05$) but the interface layout did not affect the user satisfaction.

4.3 RQ 2: What are the effects of different interface layouts on user satisfaction as measured by overall satisfaction, the likability of the engine, productivity, and mental effort?

Looking at Table 4, we see the average satisfaction scores at the interface satisfaction for each aspect we considered. When we directly compare the layouts based on these individual metrics, the Welch ANOVA test reveals that there is no statistically significant difference between them. Since the differences might be more subtle or complex, to gain a better understanding, we used a MANOVA test.

Our analysis revealed significant differences across the different layouts. For the test statistics, including Wilks' lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root, we found $F(5, 318) = 131.647, p < 0.001$. The observed effect sizes (η_p^2) ranged from medium (0.065 for Wilks' lambda and 0.135 for Pillai's trace) to large (0.414 for both Hotelling-Lawley trace and Roy's greatest root).

Given these differences exist, we try to separate the contributing components to each interface layout via Linear Discriminant Analysis (LDA). The coefficients from the LDA, which are provided in Table 5, represent the standardized contribution of each user satisfaction metric to the discriminant of the interface layouts.

From Table 5, the LDA coefficients underscore that variations in interface designs subtly impacted user perceptions and experiences, culminating in different satisfaction levels, productivity perceptions, and cognitive demands. For instance, the T layout was predominantly associated with high overall satisfaction (0.296) and cognitive load (0.106). In contrast, the random layout interface layout was characterized by higher overall satisfaction (0.312) and lower cognitive load (-0.167), but lower productivity (-0.341), revealing a potential trade-off between user satisfaction and perceived productivity.

The explained variance ratios from the LDA show the proportion of variance captured by each discriminant function. Specifically, the first discriminant function accounts for approximately 54.8% of the variance, highlighting its significance in distinguishing between

Table 4: Results of Interface Satisfaction. No statistically significant differences were found between any of the measures for a given interface layout.

Interface Layout	Felt Productive	Mentally Taxing	Liked Engine	Distracting	Overall Satisfaction
TIS	3.71±1.47	3.52±1.51	3.78±1.24	2.95±1.31	3.94±1.37
TIS WaPo	4.05±1.58	3.19±1.31	4.00±1.22	2.81±1.34	3.92±1.35
TI Google	4.06±1.18	3.01±1.40	3.8±01.12	2.79±1.27	4.07±1.17
T	4.00±1.13	3.32±1.30	3.81±0.97	2.91±1.06	4.16±1.05
Random	3.78±1.33	3.07±1.38	3.85±1.13	2.9±1.40	4.10±1.17

the interface types. This is followed by the second, third, and fourth functions, which capture 26.4%, 16.6%, and 2.2% of the variance, respectively. Based on these ratios of the discriminants, Figure 5 shows a visualization of these two discriminants in separating the different interface layouts.

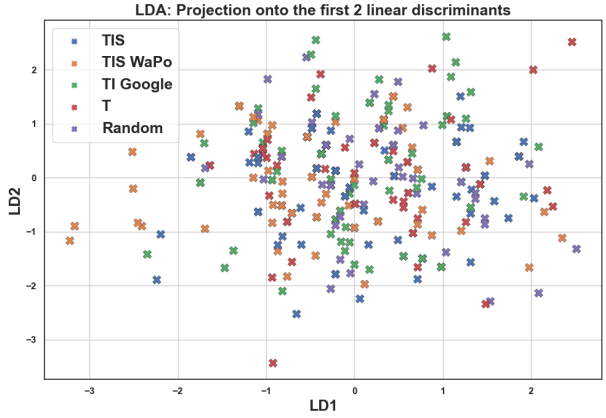


Figure 5: Visualization of the first two Linear Discriminants (LD1 and LD2), for different interface layouts

Our assessment reveals that although satisfaction metrics are interconnected, they do not completely linearly differentiate the interface layouts and that there are small overlaps between the layouts (even though some clustering is observed), as seen from Figure 5. While the layouts exhibit distinct characteristics, their differences are not solely driven by individual satisfaction metrics. Instead, a collective, non-linear interaction of these metrics influences the differences observed across interface layouts.

In our study, we have found some interesting insights. Based on the findings from RQ1, it is evident that nDCG@10 acts as a robust predictor for user satisfaction at the query level, with interface layout being influential when considering total gain on page 1. With RQ2, our exploration extends into understanding how these layouts influence user satisfaction when users complete a task (session level). While there exist differences in layout preferences and perceptions, our analysis using LDA revealed that the connection between satisfaction metrics and interface layout satisfaction is intricate and layered, deviating from a straightforward relationship. The layouts did not differ on any one specific metric of user satisfaction. These nuanced differences uncovered by LDA demonstrate that users' satisfaction with interface layouts is multi-factorial, influenced by various combinations of satisfaction metrics. By integrating the insights from both research questions, we discern that optimizing user satisfaction in IR systems is not solely about enhancing query performance or refining the presentation of results. It requires a harmonious synchronization of both elements, considering the subtle intricacies in user preferences and satisfactions, offering a pathway to building more user-centric and adaptive Information Retrieval systems.

5 CONCLUSION & FUTURE WORK

In this paper, we explored the correlation between query performance, specifically marked by nDCG@10 scores, presentation and user satisfaction, with a user study consisting of 164 participants in an ad-hoc news search task. We aimed to bridge the gap between query performance, presentation and user satisfaction, venturing beyond independent studies such as Al-Maskari et al. [2], Joho and Jose [15], Kammerer and Gerjets [16], Morrison and Vancouver [19], Rele and Duchowski [22], Teevan et al. [24], Verma and Yilmaz [26] to encapsulate the nuances of presentation impact. Our analysis revealed a strong and significant correlation between nDCG@10 scores and user satisfaction at the query level, deviating in findings from Al-Maskari et al. [2], where only a weak correlation was observed. However, we observed no direct relationship between the presentation (interface layouts), user satisfaction and query performance (with nDCG@10). Signifying that, while interface modifications impact user interactions and perceptions, they do not intrinsically augment the effectiveness of the queries for metrics such as nDCG@10, however, it does lead to changes with respect to other metrics such as the total gain on the first page. This means that with respect to presentation, the number of results and the space they occupy play a role in user satisfaction. Despite the absence of a direct correlation between interface layouts and query performance, the presentation can still impact user satisfaction metrics—such as productivity, cognitive load, likability, distraction, and overall satisfaction, falling in line with all previous work such as Joho and Jose [15], Kammerer and Gerjets [16], Morrison and Vancouver [19], Rele and Duchowski [22], Teevan et al. [24], Verma and Yilmaz [26] which reports that users perceive different result cards in different ways. We further assert that the differentiation in user satisfaction across interface layouts is complex, stemming from a multi-factorial combination of user satisfaction metrics. It is imperative to acknowledge that the interface's structure holds substantial weight in shaping user satisfaction, even though it does not directly impact query performance. This research, therefore, serves as a catalyst for a more nuanced understanding of the intricate dynamics

Table 5: Coefficients of the Linear Discriminant Analysis (LDA) for distinguishing between different interface layouts based on the features captured in the interface feedback. Each row represents the coefficients for a specific interface type.

Interface Layout	Felt Productive	Mentally Taxing	Liked Engine	Distracting	Overall Satisfaction
TIS	-0.180	0.209	0.030	-0.105	0.133
TIS WaPo	0.260	-0.013	0.349	-0.001	-0.559
TI Google	0.162	-0.120	-0.215	-0.003	-0.033
T	0.016	0.106	-0.237	-0.002	0.296
Random	-0.341	-0.167	0.014	0.119	0.312

between search performance, result presentation, and user satisfaction. It underscores the importance of interface layouts, stressing the role it plays in altering user interaction and satisfaction without directly altering search performance metrics such as the nDCG@10. Currently, our findings are limited in their applicability to other domains and tasks due to the controlled nature of the study with limited topics, queries and interfaces. Therefore, in future endeavours, we intend to broaden the scope of our study by incorporating diverse tasks, document collections, and topics, aiming to assess the generalizability of our findings across varied contexts and scenarios. We also set the scene for further exploration into the differentiation of result card layouts based on different user satisfaction metrics. This study thus provides a foundation for further exploration into the intricate interplay between system performance, presentation, and user satisfaction.

ACKNOWLEDGMENTS

We want to thank the reviewers for their insightful suggestions and feedback and all the participants who took part in the study. The work reported here is funded by the DoSSIER project under the European Union's Horizon 2020 research and innovation program, Marie Skłodowska-Curie grant agreement No. 860721

REFERENCES

[1] Azzah Al-Maskari and Mark Sanderson. 2010. A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology* 61, 5 (5 2010), 859–868. https://doi.org/10.1002/ASI.21300

[2] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07* (2007), 773–774. https://doi.org/10.1145/1277741.1277902

[3] Leif Azzopardi. 2011. The economics in interactive information retrieval. *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2011), 15–24. https://doi.org/10.1145/2009916.2009923

[4] Leif Azzopardi, Diane Kelly, and Kathy Brennan. 2013. How Query Cost Affects Search Behavior. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (7 2013), 23–32. https://doi.org/10.1145/2484028.2484049

[5] Leif Azzopardi and Guido Zuccon. 2015. An analysis of theories of search and search behavior. In *ICTIR 2015 - Proceedings of the 2015 ACM SIGIR International Conference on the Theory of Information Retrieval.* https://doi.org/10.1145/2808194.2809447

[6] Nilavra Bhattacharya. 2021. Record User Interactions on your Webpages: A tutorial. https://medium.com/@nilavra/60ccc19f0516

[7] Pia Borlund and Peter Ingwersen. 1997. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation* 53, 3 (1997), 225–250. https://doi.org/10.1108/EUM0000000007198/FULL/XML

[8] Horațiu Bota, Ke Zhou, and Joemon M. Jose. 2016. Playing your cards right: The effect of entity cards on search behaviour and workload. *CHIIR 2016 - Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval* (3 2016), 131–140. https://doi.org/10.1145/2854946.2854967

[9] Mon Chu Chen, John R. Anderson, and Myeong Ho Sohn. 2001. What can a mouse cursor tell us more? Correlation of eye/mouse movements on web browsing. *Conference on Human Factors in Computing Systems - Proceedings* (2001), 281–282. https://doi.org/10.1145/634067.634234

[10] Jacob Cohen. 1973. Eta-squared and partial eta-squared in fixed factor anova designs. *Educational and Psychological Measurement* 33, 1 (4 1973), 107–112. https://doi.org/10.1177/001316447303300111/ASSET/001316447303300111.FP.PNG{_}V03

[11] Susan Dziadosz and Raman Chandrasekar. 2002. Do thumbnail previews help users make better relevance decisions about web search results?. In *SIGIR Forum (ACM Special Interest Group on Information Retrieval).* https://doi.org/10.1145/564437.564446

[12] Qi Guo and Eugene Agichtein. 2010. Towards predicting web searcher gaze position from mouse movements. *Conference on Human Factors in Computing Systems - Proceedings* (2010), 3601–3606. https://doi.org/10.1145/1753846.1754025

[13] Jeff Huang, Ryen W. White, and Susan Dumais. 2011. No clicks, no problem: Using cursor movements to understand and improve search. *Conference on Human Factors in Computing Systems - Proceedings* (2011), 1225–1234. https://doi.org/10.1145/1978942.1979125

[14] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management* 36, 2 (3 2000), 207–227. https://doi.org/10.1016/S0306-4573(99)00056-4

[15] Hideo Joho and Joemon M Jose. 2006. A comparative study of the effectiveness of search result presentation on the Web. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 3936 LNCS. https://doi.org/10.1007/11735106{_}27

[16] Yvonne Kammerer and Peter Gerjets. 2010. How the interface design influences users' spontaneous trustworthiness evaluations of web search results: Comparing a list and a grid interface. In *Eye Tracking Research and Applications Symposium (ETRA).* https://doi.org/10.1145/1743666.1743736

[17] Diane Kelly and Leif Azzopardi. 2015. How many results per page? A study of SERP size, search behavior and user experience. *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (8 2015), 183–192. https://doi.org/10.1145/2766462.2767732

[18] David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. 2017. A study of snippet length and informativeness behaviour, performance and user experience. *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (8 2017), 135–144. https://doi.org/10.1145/3077136.3080824

[19] Elizabeth W Morrison and Jeffrey B Vancouver. 2000. Within-person analysis of information seeking: The effects of perceived costs and benefits. *Journal of Management* 26, 1 (2000), 119–137. https://doi.org/10.1177/014920630002600101

[20] Florian Mueller and Andrea Lockerd. 2001. Cheese: Tracking mouse movement activity on websites, a tool for user modeling. *Conference on Human Factors in Computing Systems - Proceedings* (2001), 279–280. https://doi.org/10.1145/634067.634233

[21] Vidhya Navalpakkam and Elizabeth F Churchill. 2012. Mouse Tracking: Measuring and Predicting Users' Experience of Web-based Content. (2012).

[22] Rachana S Rele and Andrew T Duchowski. 2005. Using eye tracking to evaluate alternative search results interfaces. In *Proceedings of the Human Factors and Ergonomics Society.* https://doi.org/10.1177/154193120504901508

[23] Nirmal Roy, David Maxwell, and Claudia Hauff. 2022. Users and Contemporary SERPs: A (Re-)Investigation Examining User Interactions and Experiences. *SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* 11, 22 (7 2022), 2765–2775. https://doi.org/10.1145/3477495.3531719

[24] Jaime Teevan, Edward Cutrell, Danyel Fisher, Steven M Drucker, Gonzalo Ramos, Paul André, and Chang Hu. 2009. Visual snippets: Summarizing web pages for search and revisitation. In *Conference on Human Factors in Computing Systems -*

Proceedings. https://doi.org/10.1145/1518701.1519008

[25] Anastasios Tombros and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)* (1998). https://doi.org/10.1145/290941.290947

[26] Manisha Verma and Emine Yilmaz. 2017. Search costs vs. User satisfaction on mobile. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10193 LNCS (2017), 698–704. https://doi.org/10.1007/978-3-319-56608-5{_}68/FIGURES/8