Study of Relevance and Effort across devices*

ABSTRACT

Relevance judgements are essential for designing information retrieval systems. Traditionally, judgements have been judgements have been gathered via desktop interfaces. However, with the rise in popularity of smaller devices for information access, it has become imperative to investigate whether desktop based judgements are different from judgements gathered using mobiles. Recently, user effort and document usefulness have also emerged as important dimensions to optimize and evaluate information retrieval systems. Since existing work is limited to desktops, it remains to be seen how these judgements are affected by user's search device. In this paper, we address these shortcomings by collecting and analyzing relevance, usefulness and effort judgements on mobiles and desktops. Analysis of these judgements indicates that high agreement rate between desktop and mobile judges for relevance, followed by usefulness and findability. We also found that desktop judges are likely to spend more time and examine documents in greater depth on non-relevant/not-useful/difficult documents compared to mobile judges. Based on our findings, we suggest that relevance judgements should be gathered via desktops and effort judgements should be collected on each device independently.

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference format:

. 2016. Study of Relevance and Effort across devices. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 4 pages.

DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Evaluation of Information retrieval (IR) systems is dependent on document relevance. IR Systems are built to optimize for relevance, where training data consists of documents either labeled manually or derived from dwell time. However, studies [2, 11] have shown that topical relevance is not the primary factor and that '*user effort*' also affects user satisfaction. Existing work [8, 11] investigates the effect of document text and structure on effort judgements. Their findings suggest that besides relevance, the *ability to find information* i.e. *findability* in a web-page is highly correlated with user satisfaction. They showed that users prefer documents where information can be located *quickly* over documents where it takes *longer* to find relevant information. Given that now people can access the same information on the web from different devices, we posit that user's *search device* would also affect the effort required to find relevant information. For instance, small viewport of mobile and touch based input may affect how a user finds information on mobile. In this work, we investigate how document relevance and search effort vary with search device.

Information access is no longer limited to stationary desktops. With constant rise in search queries from different mediums [3, 10], it has become imperative to understand whether desktop based relevance and effort judgements can be directly used for mobiles. Recent work [7] only gathered topical relevance labels and showed that relevance labels may differ across devices. We believe that the observed differences in the labels is a function of both *topical relevance* and *effort* required to label documents on both devices. Annotators may find it more difficult to label some documents on mobile and may give up or assign incorrect label to the document. Since, the authors in [7] only elicit relevance labels, it is difficult to examine role of judging effort across devices from their judgements.

We posit that the differences in relevance labels across devices [7] is a result of both *relevance* and *user effort required to extract useful information* from web-page. We investigate these differences further by gathering judgements for relevance *and* effort on both mediums. In this work, we perform a preliminary analysis of labels obtained via crowdsourcing study on mobile and desktop. We specifically gather judgements for topical relevance, page utility [4] and effort to systematically understand the differences between mobile and desktop. For generalizability, we obtain judgements for documents of TREC Web track, a publicly available dataset.

Our work aims to further answer two research questions. First, we analyze whether judgements for relevance, page utility and effort required to find the information differ across two devices with help of judging time and annotator's actions on webpage. Secondly, we study how relevance, page utility and effort are correlated across devices. From these judgements, we observed highest agreement rate between desktop and mobile judges for relevance, followed by usefulness and findability. Second, we found that desktop and mobile usefulness labels are highly correlated, followed by relevance and findability labels. Finally, we found that desktop judges are likely to spend more time and examine documents in greater depth on non-relevant/not-useful/difficult documents compared to mobile judges.

We provide a brief overview of related work and their short comings in Section 2. We describe adopted methodology and our dataset in Section 3. We describe our findings from crowd-sourced judgements in Section 4 and summarize our conclusion in Section 6.

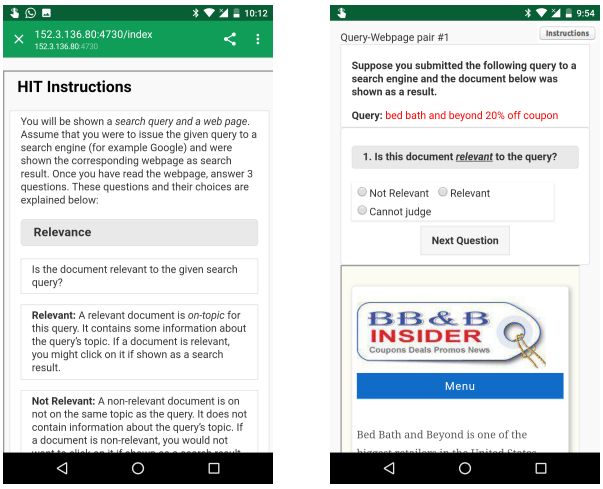
2 RELATED WORK

Our work spans multiple areas of research. We review literature that addresses crowdsourcing judgements, user behaviour on different mediums and assessor behaviour. Several studies have looked into *when* does user search for information on mobile. Existing research [10] has shown that today mobiles are used extensively to satisfy information needs. Researchers have found [1, 5] that user search logs on mobiles and desktop differ in query length, click patterns and dwell time respectively. Kamvar *et al* [3] analyze large scale query logs to distinguish between queries issued from mobile. These studies found mobile queries to be short (2.3 - 2.5 terms) and high rate of query reformulation. Small scale studies

*Produces the permission block, and copyright information

Conference'17, Washington, DC, USA

DOI: 10.1145/nnnnnnn.nnnnnnn



(a) Hit instructions (b) Labeling interface

Figure 1: Sample Mobile hit

like [6, 10] also report differences in search patterns across devices. One key result of Song *et al.* [5] studied mobile search patterns on three devices: mobile, desktop and tablets. Given significant differences between user search patterns on these platforms, their study suggested use of different web page ranking methodology for mobile and desktop.

Topical relevance has been primary focus of evaluating documents. Xu *et al.* [9, 12] conducted a study to investigate criterion that users employ to make relevance judgements. They found that topicality and novelty are the most important relevance criteria for the users, followed by understandability and reliability. However, recent work has shown [11] that besides relevance, **user effort** also affects satisfaction. Users may have to invest significant effort in reading and extracting information from a relevant document. Several parameters have been investigated [8] to characterize user effort and it has been shown that users prefer documents where it is *easier to locate* required information. Recent work [7] collected judgements for topical relevance on mobile and desktop. They showed that relevance labels may differ across devices. We believe that these differences are an outcome of effort required to label documents. In this study, we gather labels for relevance, page utility *and* effort on mobile and desktop respectively to thoroughly investigate these differences.

3 METHODOLOGY

Primary aim of this study is to collect judgements and investigate differences across devices. We designed a judging interface for mobile and desktop respectively. Mobile based judging interface is shown in Figure 1. We sampled TREC Web track queries and documents to create TREC specific evaluation dataset. We recruited annotators via the crowdsourcing platform Mechanical Turk[1].

Since clueweb12 collection is an older snapshot of www, we crawled desktop and mobile versions of URLs judged in TREC Web track. We computed cosine similarity between term vectors of clueweb12 document and crawled desktop/mobile webpage. In this study, we consider pages whose

[1]http://www.mturk.com

Table 1: Relevance label distribution

Desk↓/Mob→	NA	CJ	rel	not-rel	Total
NA	1	1	5	4	11
CJ	0	0	1	0	1
rel	12	0	65	21	98
not-rel	11	2	25	55	93
Total	24	3	96	80	**203**

desktop/mobile cosine similarity is greater than 0.80. Each TREC web query has been assigned a class on basis of its underlying information need: 'faceted', 'single' and 'ambiguous'. We construct a sample of 200 documents for 50 queries from 'single' category for query-url pairs. For an in-depth analysis, judges label each document for 1) topical relevance (*relevance*), 2) ease of finding required information (*findability*)[8] and 3) utility of the page (*usefulness*)[4] with respect to the search query. The annotation interface with instructions is available online[2]. We gather binary labels for each parameter to reduce labeling overhead on both devices. We use the following scales for each label:

- **Relevance** (*rel*): Not relevant, relevant, cannot judge (CJ).
- **Findability** (*find*): Difficult and easy.
- **Usefulness** (*use*): Not useful and useful.

We allowed annotators to skip documents that they did not want to judge to reduce spurious labels in the dataset. We payed MTurk annotators 0.06 cents for annotating a single document. Each document was annotated by 3 judges and each judge was required to label at least 10 documents to get payed. This was to ensure that only annotators interested in the task completed it. Annotators that had acceptance rate of >95% and had completed over 5000 HITs could attempt our task on Mechanical Turk. We tracked mouse movements and touch events on both devices via Javascript.

4 RESULTS

In total, we obtained labels for 203 TREC Web documents for 44 queries by 90 and 42 judges on desktop and mobile respectively. Each query-document pair was labeled by three judges. We elicit labels from judges for 3 aspects: relevance, usefulness and findability. We begin by analyzing label distribution of each dimension.

4.1 Label Distribution

Since, each document was labeled by three judges, we use majority vote to compute the final label on both mediums. Documents with no majority vote are marked NA. Distribution of majority relevance, findability and usefulness labels is given in Table 1, Table 2 and Table 3 respectively. Overall, we obtain similar distribution of relevance labels on mobile and desktop. We observe that more documents have no majority on mobile than desktop. We obtain a similar distribution for usefulness labels on both devices with slightly more documents with no majority label on mobile. However, there is slightly more variation in findability labels, where more number of documents are *difficult* (85) on mobile than desktop where only 57 documents got labeled *difficult*.

We compute inter-rater agreement using Krippendorff's Alpha (α) and Cohen's kappa (κ) on binary judgments given in Table 4. Since three judges labeled each document, we report average (and standard deviation) Cohen's kappa over

[2]http://128.16.12.66:4730/index, batch:nxaa, workerid:userid

Table 2: Findability Label Distribution

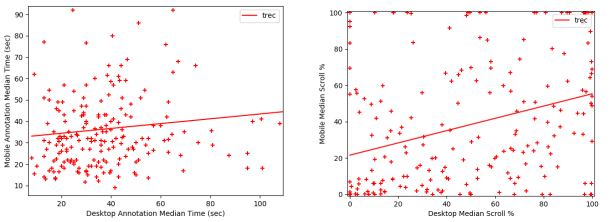
Desk↓/Mob→	CJ	easy	difficult	Total
CJ	1	1	4	6
easy	12	79	49	140
difficult	5	30	22	57
Total	18	110	85	203

Table 3: Usefulness label distribution

Desk↓/Mob→	NA	CJ	use	not-use	Total
NA	0	1	5	7	13
CJ	1	0	0	0	1
use	8	0	64	23	95
not-use	15	4	15	60	94
Total	24	5	84	90	203

Table 4: Inter-rater agreement on binary judgments

	Desktop		Mobile		M/D	Random
	α	κ	α	κ	κ	κ
Useful	0.47	0.46 (0.02)	0.44	0.43 (0.02)	0.38	0.11 (0.06)
Rel	0.55	0.56 (0.02)	0.44	0.43 (0.02)	0.41	0.12 (0.05)
Find	0.22	0.23 (0.04)	0.11	0.11 (0.03)	0.17	0.06 (0.07)



(a) Judging Time on Desktop and Mobile **(b) Scroll depth on Desktop and Mobile**

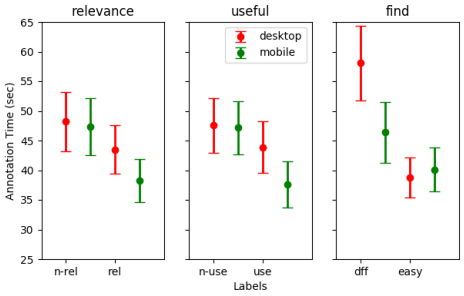
Figure 2: Scroll Depth and Judging time on desktop and mobile

50 trails. For each trial, we randomly sample two labels for each query-url pair and compute cohen's kappa. We also report agreement within device (desktop and mobile), across devices (M/D) and agreement computed between randomly chosen desktop and mobile labels. It is worth noting that (M/D) is computed using document's majority label for mobile and desktop. Random Agreement (column 4 in Table 4 is the average agreement between randomly choose label from desktop and mobile respectively. We observed the high agreement rate for relevance on desktop, followed by usefulness and findability respectively. In mobile, we obtain similar agreement rates for relevance and usefulness but least agreement for findability. Agreement rate between mobile and desktop is also largest for relevance followed by usefulness and findability. Random agreement between mobile and desktop is expected to be lower than others. However, chance agreement of relevance labels is the highest amongst all other labels.

4.2 Judging time and Examination depth

We now compare the overall distribution of judging time of each annotator on Mobile and Desktop. Figure 2a shows the overall distribution of time it took any judge to label the same document on mobile and desktop. We also compare the percentage of document examined by any assessor on mobile and desktop before submitting the judgments. Note that y-axis reports the percentage of URL examined on mobile and x-axis depicts the percentage of same URL examined on desktop. The scroll percentage distribution is shown in Figure 2b. Here, judges examine more content on mobile as compare to desktop[3].

Figure 3: Judging Time on Mobile/Desktop



We observe that median judging time on desktop is weakly associated with median judging time on mobile. Pearson's correlation ρ between desktop and mobile judging time is 0.12 (p-val < 0.01). However, in Figure 2b we see stronger correlation between median scroll depth on mobile and desktop. Pearson's correlation ρ between desktop and mobile examination depth is 0.28 (p-val < 0.01), higher than that of judging time correlation.

We also examine whether relevance, usefulness and effort labels differ on basis of judging time and examination depth. We plot mean judging time and examination depth with 95% confidence intervals for each label in Figure 3 and Figure 4 respectively.

One key observation is that judges take more time to judge non-relevant/not-useful and difficult documents on mobile and desktop respectively. However, we find that desktop judging time distribution of relevant (useful) and non-relevant (not-useful) are not statistically different. On the contrary, mobile judging time of relevant (useful) and non-relevant (not-useful) documents is significantly different. Mobile judging time of relevant (and useful) documents is also significantly lower than desktop judging time which is in line with previous findings in [7].

We observe a different trend for findability labels. We found that mobile judges are much less likely to spend time examining difficult documents in comparison to desktop judges. We attribute this difference to limited input capabilities and touch interaction on mobile devices. Desktop judges can be more thorough as they can easily interact with a webpage via keyboard and mouse. Similar conclusion can be drawn from scrolling/swiping behavior in Figure 4. We find that desktop judges examine significantly more content than mobile judges regardless of document's relevance/usefulness/effort label.

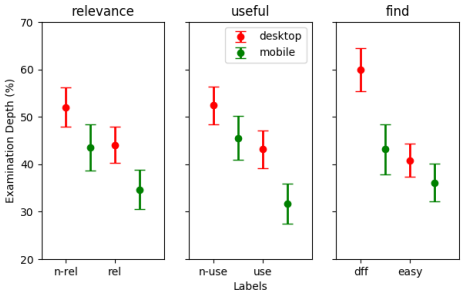
Figure 4 shows that judges examine non-relevant/not-useful/difficult documents in greater depth than relevant/useful/easy documents on both devices. This is expected, as judges would need to read difficult documents more carefully and thoroughly to find required information. This would result in higher examination depth on non-relevant/useful/difficult

[3]Percentage is computed with respect to document length rendered on desktop/mobile screen to remove effect of screen size on calculation.

Table 5: Correlation between all judgements

Desk↓/Mob→	relevance	usefulness	effort
relevance	0.23**	0.33**	0.21**
usefulness	0.28**	0.34**	0.20**
effort	0.16*	0.13	0.08

Figure 4: Examination Depth on Mobile/Desktop



documents. We observe a significantly large difference in examination depth of difficult pages between mobile and desktop, in combination with judging time information, indicates that mobile judges may be less patient in looking for query specific information.

On analysis of judging time and examination depth information, in conjunction with agreement rates in Table 4 (column M/D) we believe that desktop judges are more thorough in labeling documents than mobile judges. They are likely to spend more time and examine the document in greater depth before assigning any label. Mobile judges, however, due to device and interface limitations, may be less patient in labeling a document with respect to a search query. However, low agreement of findability between mobile and desktop judges clearly elicits the need of device specific effort judgements.

5 LABELS CORRELATION

To investigate this further we can compute correlations between desktop and mobile relevance, usefulness and findability labels. Our hypothesis is that weak correlation of labels across both devices would indicate a disagreement between desktop and mobile judges while a stronger correlation would reflect higher agreement between judges.

We present the correlation between each label across devices in Table 5. Statistically significant entries are marked with * (p-val < 0.05) or ** (p-val < 0.01) respectively. Table 5 clearly indicates that relevance and usefulness have higher correlation than relevance and findability labels. Weak correlation between effort labels across devices suggests that effort labels differ across devices and should be gathered on per-device basis.

6 DISCUSSION AND CONCLUSION

Information retrieval (IR) is no longer limited to desktops. Today, users increasingly rely on IR systems to find information on devices such as mobiles or tablets. While prior work exists on designing, exploiting or evaluating IR systems across devices, little work has been done to investigate affect of user's search device on relevance judgements. Relevance judgements lie at the heart of IR systems and existing algorithms need large scale labeled data to be effective. Until recently, judgements have been collected via desktop interfaces. However, with advent of different devices, we need to investigate whether relevance judgements differ across devices.

Recently, researchers have also suggested to gather usefulness and effort based judgements along with relevance to train more effective systems. We believe that these judgements would be affected by user's search device. Existing work only analyzes these judgements for desktops. In this work, we address the above shortcomings in that we systematically collect and analyze device specific judgements for TREC dataset for three factors: relevance, usefulness and effort. Our analysis indicates three key trends. First, we observed highest agreement rate between desktop and mobile judges for relevance, followed by usefulness and findability. Second, we found that desktop and mobile usefulness labels are highly correlated, followed by relevance and findability labels. We also observed higher correlation between relevance and usefulness labels than between relevance and findability labels. Finally, we found that desktop judges are likely to spend more time and examine documents in greater depth on non-relevant/not-useful/difficult documents compared to mobile judges.

Based on our findings, we suggest that relevance judgements should be gathered via desktops as desktop judges are more patient and thorough in assessing a webpage compared to mobile judges. However, effort based judgements should be collected on each device independently to account for affect of device specific properties on labels.

REFERENCES

[1] K. Church and B. Smyth. Understanding the intent behind mobile information needs. In *Proc. IUI*. ACM, 2009.
[2] J. Jiang, D. He, D. Kelly, and J. Allan. Understanding ephemeral state of relevance. In *CHIIR '17*, 2017.
[3] M. Kamvar, M. Kellar, R. Patel, and Y. Xu. Computers and iphones and mobile phones, oh my!: A logs-based comparison of search users on different devices. In *Proc. WWW*. ACM, 2009.
[4] J. Mao, Y. Liu, K. Zhou, J.-Y. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo. When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 463–472, New York, NY, USA, 2016. ACM.
[5] Y. Song, H. Ma, H. Wang, and K. Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *Proc. WWW*, 2013.
[6] C. Tossell, P. Kortum, A. Rahmati, C. Shepard, and L. Zhong. Characterizing web use on smartphones. In *Proc. SIGCHI*. ACM, 2012.
[7] M. Verma and E. Yilmaz. Characterizing relevance on mobile and desktop. In *European Conference on Information Retrieval*, pages 212–223. Springer, 2016.
[8] M. Verma, E. Yilmaz, and N. Craswell. On obtaining effort based judgements for information retrieval. WSDM '16. ACM, 2016.
[9] Y. C. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7), 2006.
[10] J. Yi, F. Maghoul, and J. Pedersen. Deciphering mobile search patterns: A study of yahoo! mobile search queries. In *Proc. WWW*. ACM, 2008.
[11] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: an analysis of document utility. In *CIKM '14*. ACM, 2014.
[12] Y. Zhang, J. Zhang, M. Lease, and J. Gwizdka. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14. ACM, 2014.