On Obtaining Effort Based Judgements for Information Retrieval

Manisha Verma
University College London
m.verma@cs.ucl.ac.uk

Emine Yilmaz
University College London
emine.yilmaz@ucl.ac.uk

Nick Craswell
Microsoft
nickcr@microsoft.com

ABSTRACT

Document relevance has been the primary focus in the design, optimization and evaluation of retrieval systems. Traditional test collections are constructed by asking judges the relevance grade for a document with respect to an input query. Recent work [44] found an evidence that *effort* is another important factor in determining document utility, suggesting that more thought should be given into incorporating effort into information retrieval. However, that work did not ask judges to directly assess the level of effort required to consume a document or analyse how effort judgements relate to traditional relevance judgements.

In this work, focusing on three aspects associated with effort, we show that it is possible to get judgements of effort from the assessors. We further show that given documents of the *same* relevance grade, effort needed to find the portion of the document relevant to the query is a significant factor in determining user satisfaction as well as user preference between these documents. Our results suggest that if the end goal is to build retrieval systems that optimize user satisfaction, effort should be included as an additional factor to relevance in building and evaluating retrieval systems. We further show that new retrieval features are needed if the goal is to build retrieval systems that jointly optimize relevance and effort and propose a set of such features. Finally, we focus on the evaluation of retrieval systems and show that incorporating effort into retrieval evaluation could lead to significant differences regarding the performance of retrieval systems.

Keywords

Information retrieval; evaluation; effort; judgements

1. INTRODUCTION

Relevance is a fundamental concept in information retrieval. Human relevance assessments are collected to form test collections. These test collections are then used for relevance-focused evaluation of retrieval systems and relevance-focused optimization of the free parameters of such systems.

Based on the assumption that relevance is the primary factor that affects user satisfaction [20, 21], evaluation using relevance judgements should be predictive of overall satisfaction of real users who are interacting with a search system. Yet, research has shown that test collection and user based evaluation may not completely agree with each other [19, 39], or agree only when there is a significant gap in terms of quality of the systems compared [1, 2].

One potential reason for this mismatch is that there are factors other than relevance which contribute to user satisfaction. Such factors would not be captured by judging scheme focussed on topical relevance. By comparing implicit feedback obtained from real users and the judgements obtained from relevance assessors, Yilmaz *et al.* [44] recently showed that the utility of a document with respect to an actual user can be different from relevance of a document. In particular, they showed that *effort* required to find or consume relevant information within a document is a significant factor that can affect user satisfaction, and can lead to a real user abandoning a document, perhaps to find another relevant document that requires less effort. This abandonment of documents was correlated with documents that are long or have a relatively high reading level. Though this work detected and analysed abandonment of some relevant documents, it did not elaborate on what constitutes effort. In particular, it did not gather human judgements of factors other than relevance, that may contribute to effort or overall satisfaction. Such judgements could be used to evaluate and optimize retrieval systems that take into account both effort and relevance.

In this work, we build on the findings of Yilmaz *et al.* [44] by first identifying factors that characterize user effort, where effort can be defined as the amount of cognitive effort it takes user to find, read and understand information in a document. Mainly, we focus on three factors that might affect the effort required to find and consume relevant information 1) easiness to find relevant information in a document, 2) readability level of the document, and 3) easiness to understand the contents of the document. We conduct experiments to obtain explicit judgements for these factors and finally analyse which of these factors are significant for user satisfaction. We show that 1) it is possible to obtain judgements from assessors with respect to all these aspects, and 2) given documents of the same relevance grade, some of these effort related factors can have a direct impact on user satisfaction. In particular, we observe that easiness to find

WSDM'16, February 22–25, 2016, San Francisco, CA, USA.

DOI: http://dx.doi.org/10.1145/2835776.2835840

relevant information in the document is a significant factor that can affect user satisfaction.

Given the evidence that effort in document consumption can impact user satisfaction, retrieval systems should be optimized for effort together with relevance, and evaluation mechanisms should incorporate effort together with relevance. For this purpose we propose a set of features that could be used in an effort-aware ranking system. Analysis shows that some of the features which are significant for retrieving low effort documents are not captured when focusing solely on relevance. This suggests that part of the work in incorporating effort in retrieval optimization is to add new features. We also analyse the effect of incorporating effort into retrieval evaluation. We analyse systems submitted to TREC Web Track Adhoc task 2012-2014, and show that even though the top systems show similar performance in terms of relevance, these systems tend to perform quite differently when effort is considered.

We begin by explaining related work on relevance estimation and effort in information retrieval in Section 2. We elaborate on effort parameters and related user studies in Sections 3, 4 and 5 respectively. Our experiments and findings are reported in Section 6. Evaluation of TREC web track submissions is described in Section 7. We conclude with our findings and future work in Section 8.

2. RELATED WORK

There is a large body of work characterizing relevance and effort. We briefly discuss how effort is defined, measured and incorporated in information retrieval. Since we also study what factors are associated with effort, we briefly cover factors used to assess document relevance. Finally, we cover literature from other areas that highlight the importance of parameters we choose to capture effort.

Effort definitions and measurement

Existing literature defines and studies effort differently during various stages of information searching and gathering. For example, Gwizdka *et al.* [16] study user search effort where effort is only defined as number of documents visited while searching. However, in this work we study what factors best capture effort while consuming a *single* document. Similarly, Carterette *et al.* [6] study existing evaluation measures on basis of user model which estimates the effort a user must put forth to achieve a certain utility from ranked list. Smucker *et al.* [35] propose to evaluate search effectiveness on the basis of time spent and information gained (TBG) as user scans ranked document list. While TBG uses time spent on a document to implicitly measure effort, we collect explicit judgements of effort for a document. Recently, Ferro *et al.* [14] proposed a measure that evaluates the effectiveness of a system based on the effort it takes to retrieve desired information. They compare and contrast gain versus effort required to scan search results. While existing effort-based evaluation mechanisms [6, 14, 16, 35] account for user's interaction with a list of documents, we explicitly capture effort per *relevant document* regardless of its *position* in search results.

Some work has also studied user effort in searching or judging a document. Villa *et al.* [40] conducted a study that looked at the relationship between document length and both judging effort and accuracy. They concluded that accuracy is not affected by document size but judging longer documents required more effort. In this work, we use document length not as definition of effort but one of the features to predict effort. They also found that relevant documents require most effort to judge (significant differences were found for mental demand, physical demand, and effort). Our work would be useful in identifying and labelling such documents. Judging effort has also been studied for images, in [17] authors found that while image size did not affect judging accuracy, it significantly impacted the judging time.

Factors characterizing relevance

Since document relevance is of prime importance in evaluation, significant amount of work investigates what factors are important in assessing relevance and whether they remain constant or evolve with user's interaction with search results. Schamber *et al.* [33] concluded that relevance is a multidimensional concept that is affected by several factors.

Xu *et al.* [43] conducted a study to investigate criterion that users employ to make relevance judgements. They proposed that topicality, novelty, reliability, understandability, and scope characterize relevance. They found that topicality and novelty are the most important relevance criteria for the users, followed by understandability and reliability. Zhang *et al.* [45] also showed that understandability and reliability did not affect relevance. They also found understandability did not explain relevance judgements as completely as novelty and topicality did. Some work also studies what factors are important for judging relevance when search task and session are taken into account. Borlund [4] studied how user's notion of relevance changes with session time. Taylor's work [37] with two longitudinal studies investigated association between the search process and 15 different relevance criterion. They found both structure and understandability became more important to subjects during later search stages and are pre-requisite to positive relevance judgements. We use web page oriented features to capture structure and language specific readability measures to capture understandability in our experiments. Above mentioned work [4, 12, 31, 33, 43, 45] states that relevance has several components and factors like structure, reliability and topicality affect relevance judgements. Given that these parameters determine relevance, with this work, we conduct further analysis of factors that distinguish two *relevant* documents. We investigate whether factors such as understandability, findability etc. can be useful in distinguishing two documents of *equal* relevance grade. Importance of such factors will greatly differ when a relevant document is compared with irrelevant document, which is not the focus of this work.

Factors characterizing effort

We investigate three factors that may be associated with effort: ease of finding information in a document, ease of reading and understanding its content. Existing work does not study the role of these factors in judging effort, however, one or more of these factors have been used to improve quality of search results for end users. For instance, readability has been used to filter [10, 25] or personalize [11, 24, 36] search results according to user's language proficiency. It has been shown that webpage readability levels impact users understanding of the document. Chandar *et al.* [8] have also shown that readability affects assessor disagreement.

It has also been shown that users actively *find* relevant or interesting information on a page [13, 27, 26] and may not sequentially read entire webpages. Guo *et al.* [15] studied cursor movements and found that users read relevant documents at length *after* scanning them. Scanning indicates that user is actively looking for required information on the page.

Understanding or consuming document is important in satisfying an information need. Information foraging theory [28, 29] has been used to show that users actively seek, filter, read, and extract information to satisfy information need. Thus, while previous research uses above parameters independently to tailor search results for end users, it is not known which parameter is more important for differentiating two *equally relevant* documents. We aim to obtain explicit judgements for all these parameters to identify which factor is highly associated with judging effort for a relevant document.

3. METHODOLOGY

By comparing relevance judgements with implicit signals of user satisfaction obtained via click logs, recent work by Yilmaz *et al.* [44] shows that 1) there can be significant mismatches between the utility of a document to an actual user and relevance of the document, and 2) some of these mismatches can be explained by factors related to the effort needed to find and process relevant information. These findings were based on relevance judgements and the behaviour of real users, but did not involve direct judgements of effort or analysis of how such judgements could be incorporated into the overall evaluation of information retrieval systems. Our primary purpose in this work is to show that it is possible to get reliable judgements of effort from relevance assessors and that incorporating these judgements into retrieval evaluation could lead to differences in system rankings. For this purpose, we first identify factors associated with effort. We design a judging interface and get judgements associated with these factors. We then analyse which of these effort related factors tend to be important for user satisfaction.

3.1 Factors Associated with Effort

We base our selection of effort related factors on previous work [15, 26, 27, 42] and the user model proposed in Yilmaz *et al.* [44] where a user does not read an entire web page sequentially. These studies suggest following model: when users first access a web page, they quickly scan it to determine portions of the document relevant to the query (*findability*). This is followed by reading these parts (*readability*) and finally understanding these nuggets of information (*understandability*). Based on this behaviour, given an information need, we hypothesize that the effort needed to satisfy the information need is affected by three primary factors:

- **Findability**: Effort needed to find the relevant information in a document.
- **Readability**: Effort required to read a document.
- **Understandability**: Effort required to understand a document to satisfy the information need.

Findability

Given an information need, the first step required to satisfy the need is to find relevant part(s) of the document. It has been shown [15, 27, 42] that users do not read entire webpages but first scan them for relevant parts. Effort needed to find the relevant portion of the document could have a significant effect on user satisfaction. Even if the document is highly relevant, the user may give up and end up being unsatisfied if it takes her too long to find required information in the document.

Readability

Once a part of the text that is relevant to the information need has been found, the user then has to read it to extract useful information. Reading a verbose document containing long sentences and difficult vocabulary may take a lot of effort for the user and may cause the user to be less satisfied, all other things being equal. Readability of a document can be quite subjective as it depends on the reading ability of the user: A fairly advanced reader will navigate difficult documents with relative ease as compared to a non-native speaker who struggles with the language. In this work we only focus on readability of document text, we shall study variance in effort due to user's readability level in future.

Understandability

Given that user may read only parts of the document, she has to process and understand the content in order to satisfy the desired information need. Even if document text is readable, if the information is not presented in a coherent manner, there are flaws in the description or the information is spread throughout the document, it can be difficult to understand. Such factors, which can lead to the user being unsatisfied, we denote as problems of understandability. Understandability can also be affected by the layout of the page. For example, pages with a lot of outlinks or advertisements distract users [13, 30] and make it difficult to extract the relevant information from the page.

It is worth noting that there may be other user specific factors that attribute to effort such as language proficiency or expertise in search topic that can be investigated in future. With this work, we aim to identify which of the three factors are important representatives of user effort in determining document relevance. We posit that given two documents of *same* relevance grade, users will prefer a low effort document over high effort document. We also determine how these factors correlate with user preferences. The user study investigating these questions and our findings are presented in the following sections.

4. EFFORT BASED JUDGING

We use crowdsourcing mechanisms in order to get explicit judgements of effort. We judge each of our hypothesized factors separately: the effort needed to find information (Findability), the readability of the document (Readability), and the understandability of the document (Understandability). Each of these effort-related aspects is measured on four point scale: *'very easy'*, *'easy'*, *'somewhat difficult'* and *'very difficult'*. We also ask judges to provide judgements about how satisfied they are with the document, and the relevance of the document. A sample HIT is shown in Figure 1.

For this study, we use data from Kazai et al. [22], which

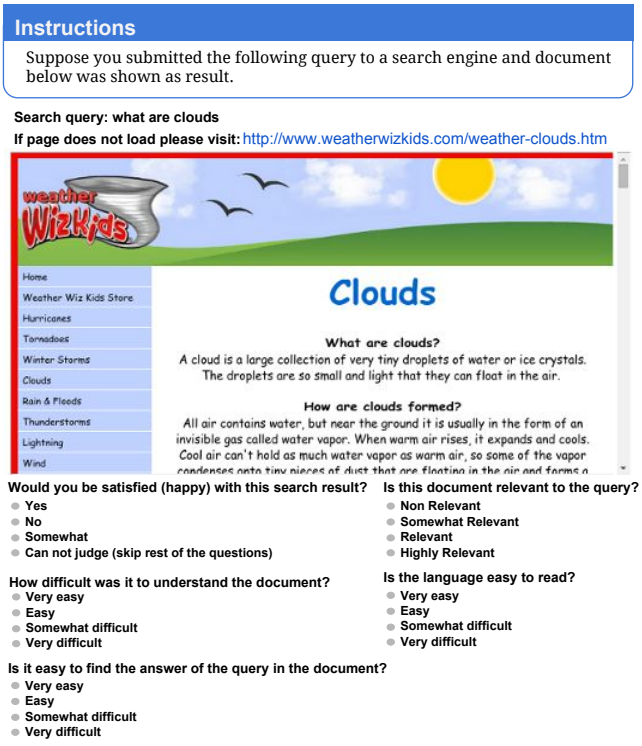


Figure 1: Sample Effort Judging HIT

consists of queries from TREC Web Track Ad Hoc task in 2009 and 2010. The full dataset contains 1603 URLs and 100 queries, where each query URL pair is judged on 5 grades of relevance by expert judges. Our main focus in this paper is to analyse how and whether effort affects user satisfaction for a relevant document. In this study, we control for relevance by excluding any non-relevant documents from our analysis, eliminating the lowest of the 5 grades. We also exclude inactive URLs from our analysis. Since effort is a factor that can affect user satisfaction only when a document is relevant,

Our goal with new crowd judgements is to eliminate differences due to relevance, and focus on effort-related differences identified in [44]. We therefore obtain our new judgements on pairs of documents with variation in effort but equal relevance. The signs of high and low effort are taken from [44]. Specifically, the number of words in a document (i.e., document length) and the readability level of the document measured by the readability measure LIX [5] are signals that can be associated with the effort needed to satisfy the information need in a document. We control for relevance by choosing documents with the same expert relevance grade (of the 4 remaining expert labels) and also eliminating one-word queries since such queries are ambiguous to untrained crowd judges. Hence, for each query in our dataset, from documents of same relevance grade, we include documents that have the maximum difference in terms of number of words and LIX. This way we ensure that both high and low effort documents are covered in our analysis, and can see which of our three hypothesized factors are seem to vary according to crowd judges.

Table 1: Effort Label Distribution

	very easy	easy	somewhat difficult	very difficult
Findability	89	41	17	19
Readability	96	46	13	11
Understandability	78	55	23	10

Table 2: Inter-annotator Agreement

Feature	Alpha (α)
Findability	0.35
Readability	0.22
Understandability	0.27
Satisfaction	0.38
Relevance	0.38

We used Amazon Mechanical Turk[1] (Mturk) to obtain preference labels where each tuple *(query, url)* in a HIT was judged for \$0.04 by three labellers. We use the majority vote to determine final grade of each document. After removing spurious labels (determined by time spent on task) and 'can't-judge' cases, our dataset consists of 80 queries and 166 documents. Ground truth relevance labels from TREC collection has following distribution: 114, 29, 11, 12 marked 'highly relevant', 'relevant', 'somewhat relevant' and 'non relevant' respectively. Table 1 summarizes data collected from this experiment. Relevance labels obtained from Mturk has following distribution: 76, 52, 22 and 16 documents have been marked 'highly relevant', 'relevant', 'somewhat relevant' and 'non relevant' respectively. We obtain following judgement for satisfaction: 143, 15 and 8 documents have been marked 'yes', 'somewhat', 'no' respectively. The median and standard deviation of time spent on task was 36 seconds and 141 seconds respectively.

In order to measure the reliability of the judgements obtained and the inter-annotator agreement, we use Krippendorff's alpha (α) [18]. As shown in Table 2, Alpha values of effort judgements lie in range of 0.22 and 0.38, which is comparable or even higher than the alpha values observed in previous work that measures the inter-annotator agreement between assessors that judge document relevance [3, 23, 32].

The inter-annotator agreement appears to be the highest for relevance and satisfaction (*0.38*). In terms of effort based judgements, findability has the highest inter-annotator agreement (*0.35*), which is comparable to that of relevance and satisfaction. On the other hand, inter-annotator agreement between understandability and readability seems to be lower (*0.27* and *0.22*, respectively). One explanation for this is that ease of finding information is an objective question, while understandability and readability are subjective. They depend on the judge's background knowledge, reading level, intellectual capacity, etc. Therefore, judgements associated with findability seem to be more reliable than the other two judgements. However, the inter-annotator agreement for understandability and readability is still comparable and even higher than the agreement values reported for relevance judgements in the previous work [3].

Given that retrieval evaluation aims at predicting user satisfaction, the primary focus in retrieval evaluation has been

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

on getting judgements of relevance and assuming that user satisfaction is a direct function of relevance. If we were to analyse the correlations of the judgements for each pair of factors using Spearman correlation coefficient, satisfaction seems to be significantly correlated with two factors: relevance and findability (with Spearman correlation coefficient of 0.375 and 0.24, respectively). Both of these correlations are statistically significant with $p \leq 0.05$. High correlation between relevance and satisfaction is inline with previous work [20] which shows that they are correlated for higher relevance grades. The other two factors associated with effort, understandability and readability do not seem to be significantly correlated with satisfaction based on Spearman correlation. Our results confirm that in contrast to the common assumption, user satisfaction is not just a function of relevance but effort to find relevant information is also a significant factor that affects user satisfaction. Even though readability and understandability do not seem as significant factors in effort, we believe that one reason for this is due to the highly subjective nature of these judgements. Hence, these aspects associated with effort should be investigated on a personal basis and they should be considered in context of personalized retrieval and personalized evaluation of retrieval systems, which is outside of the scope of this study.

In order to confirm our analysis, we further use ordinal logistic regression to predict satisfaction of a document given the three effort based factors and relevance. The analysis is shown in Table 3. Using the aforementioned four features, we were able to predict satisfaction with Adjusted R^2 coefficient of 0.32. The table shows feature used in the regression study, together with the associated p values. It can be seen that while relevance, findability and understandability tend to get positive weight when predicting satisfaction, readability coefficient is negative but not significant. Similar to the conclusions of the correlation analysis, our regression results confirm that relevance and findability are significant factors in predicting user satisfaction. Thus, effort based judgements associated with findability should be incorporated into retrieval evaluation if the goal is to evaluate user satisfaction.

Overall, our study supports following hypothesis:

- Effort based factors are significant for user satisfaction.
- Findability is an important factor to characterize effort.

The aforementioned study assumes that judgements based on user satisfaction should be the gold standard used in evaluation and aims at identifying factors that are significant factors for predicting satisfaction. However, these findings are still subject to the reliability and validity of the judgements associated with satisfaction. In order to further validate the findings based on satisfaction analysis, we conducted a follow up study on agreement of user preferences with effort factors. With preference based judging, judges tend to have freedom to decide between documents and are not restricted to evaluate them with respect to some predefined factors. Hence, preference based judgements are useful in getting unbiased decisions about what users prefer to see in a document without making the judges think about particular aspects associated with a document (such as relevance). Therefore, we collect preference based judgements between two documents and study the correlation between the three effort based factors with user preferences,

Table 3: Factor importance for Satisfaction

Feature	p-value
$Findablity^{+}$	0.003
$Readability^{-}$	0.364
$Understandability^{+}$	0.054
$Relevance^{+}$	0

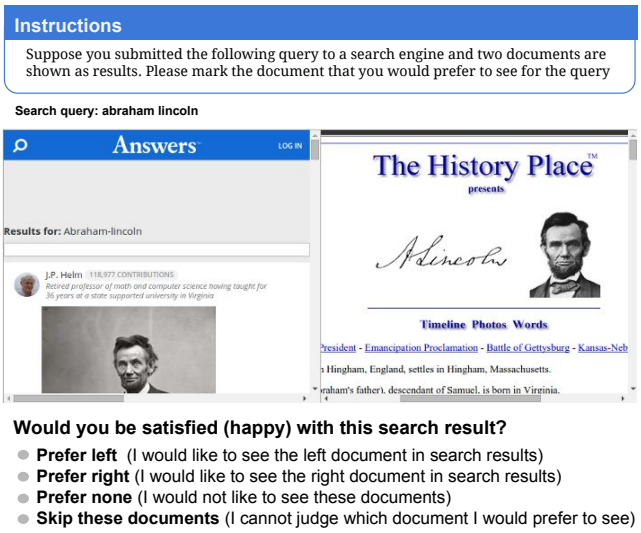


Figure 2: Sample Preference Judging HIT

analysing whether any of the effort related factors are significantly correlated with user preferences. The study and analysis are presented in the following section.

5. EFFORT-PREFERENCE CORRELATION

Primary aim of this experiment is to study and analyse preference correlation with effort factors defined in previous section. Preference judgements have been previously shown to be more reliable with better inter-annotator agreement than absolute judgements. For instance, Carterette *et al.* [7], studied assessor agreement and compared time spent on pairwise preference judgements and graded judgements. We use very similar guidelines and judging interface to the one used in that study. The judging interface used can be seen in Figure 2. Here two documents are shown side by side in separate frames to enable independent scrolling of either page. One important aspect associated with our judging interface is that we do not ask the judges to pick the document that is *more* relevant (which would bias them to think about relevance as opposed to what is really important for them). Instead, we provide judges with minimal instructions and just ask the judges to pick the document they would prefer.

Our main focus in this study is to analyse whether any of the effort related factors are important for user preferences. For this purpose, we use the same dataset as the one used in the effort based judging study (Section 4). We obtain preference based judgements on the documents of the same relevance grade. Thus, for each query, we show a pair of documents that are of the same relevance grade side by side to each user. We then analyse whether the users tend to prefer one document over the other, and whether any of the effort related factors are correlated with user preferences.

Table 4: Preference and Effort Factors Agreement

Factor	Percentage
Findability	0.607*
Readability	0.512
Understandability	0.511
Satisfaction	0.727*

We used Mturk to obtain preference labels where each triplet *(query, url1, url2)* in a HIT was judged for $0.04 by three workers. The total cost of the experiment was $23.5. We use the majority vote between three judges as the final judgement. After removing pairs with no clear preference (i.e., pairs which had 3 judges label 'Prefer left', 'Prefer right' and 'Both irrelevant') and the hits that were skipped by judges, we obtain a total of 81 triplets for our analysis. The median and standard deviation of time spent on preference interface are 17 seconds and 47 seconds respectively.

The average pairwise percentage agreement for the set of judgements obtained from this study is *0.60*, which is higher than random agreement of 0.5 ($p \leq 0.05$). Given that the judges are only shown same relevance grade documents, the fact that there is significantly higher agreement than random between the judges indicates that there are some additional factors that affect user satisfaction. When compared to the inter-annotator agreement values reported by Carterette *et al.* [7] (which focus on getting judgements associated with relevance, and asking judges to rate documents that could be of different relevance grades), the inter-annotator agreement in our study is comparable but slightly lower (approximately 0.7 versus 0.6). This is because our judging task is much harder since we focus on getting preference judgements on documents that are of *same* relevance grade.

Given the preference based judgements from the judges, we further analyse whether any of the effort related factors are significantly correlated with user preferences, i.e., whether the users tend to prefer low versus high effort documents and whether these preferences are statistically significant. Table 4 shows percentage agreement between preference and an effort factor. Basically it captures the percentage of pairs where if judges prefer one document over the other, the effort statistic also prefers that document, i.e. effort value of preferred document is lower than the other document. We also analyse the agreement of satisfaction based judgements (obtained via the effort based judging interface in the previous section) with user preferences.

It can be seen that satisfaction and findability are highly correlated with user preferences and these correlations are statistically significant ($p \leq 0.05$). The high correlation of satisfaction based judgements with user preferences further confirm the reliability of the satisfaction based judgements from the effort based judging study. Furthermore, our results here confirm that out of the three effort based factors, findability is the primary factor that can significantly affect user preferences and satisfaction. Given that the inter-annotator agreement between the preference judgements is 0.60, the agreement of 0.607 between findability and user preferences (obtained via the majority vote) is comparable with agreement between two random judges in terms of the documents they prefer.

Overall, our analysis shows that findability is an important factor that can affect user satisfaction and preferences, suggesting that retrieval systems should be built to optimize for findability together with relevance if the goal is to optimize user satisfaction. Our results further show that findability is another factor that should be considered together with relevance for evaluating user satisfaction. In the following sections, we focus on analysing how incorporating findability in building and evaluating retrieval systems could change the design of retrieval systems.

6. PREDICTING EFFORT VERSUS RELEVANCE

In this section, we focus on predicting the most important effort related factor: findability and analyse how building systems that optimize for this factor would require different types of features. In particular, we propose and investigate some features and their accuracy in predicting findability. We also use these features to predict relevance and compare and contrast features that are useful in predicting relevance and findability. Our hypothesis is that the features that are important for predicting findability are not necessarily correlated with features that are important for predicting relevance. This would suggest that in order to optimize for effort (or findability) together with relevance, search systems should include additional features (such as the ones proposed in this paper) that are designed to capture findability or effort to find relevant information.

First, we propose and describe several features that can capture the easiness of finding information in a document, then show the importance of these features for both predicting findability and relevance of a document.

6.1 Features

We propose several features that incorporate different dimensions of effort. The first set is text based features that are related to the content of the document and second is html oriented features that are related to the layout of the page.

6.1.1 Text features

We construct features from entire document text and from summary (part of the document that contains the query terms). Since a user may not always read the entire document if she has little time, often, the quickest way to judge a document is to search for the query terms and read the neighbouring paragraphs (i.e., the summary). We create document summary using sentences that contain query terms along with one sentence that appears before and after them. Similar features have been used previously in [44]. The features are summarized in Table 5.

- Typically, lengthier documents may require more effort than shorter documents. Hence, we use features that capture the length of document. They mainly cover number of words and sentences in a document. Similar values are also calculated for summary.

- Secondly, to assess the difficulty of the documents and corresponding summaries, we use three readability indices, namely Coleman Liau index (CLI) [9], Automated Readability Index (ARI) [34] and LIX [5]. These metrics are calculated by counting number of words, sentences and are used as a rough estimate for a document's difficulty. These features are calculated both

Table 5: Text features used for predicting findability and relevance

Summary and Document Specific Features			
avgSumChar	Avg #characters in summary	avgDocChar	Avg #characters in document
sumWords	#words in summary	docWords	#words in document
sumPunct	#punctuations in summary	docPunct	#punctuations in document
sumSent	#sentences in summary	docSent	#sentences in document
sumSentQT	#summary sentences with query terms	docSentQT	#document sentences with query terms
Readability Features			
sumARI	ARI Index of summary	docARI	ARI Index of document
sumCLI	CLI Index of summary	docCLI	CLI Index of document
sumLIX	LIX Index of summary	docLIX	LIX Index of document
Other Features			
queryFreq	#query appears in page	minQPos	Min pos of query term in document
qTermstInTitle	#query terms in title	maxQPos	Max pos of query term in document
qWinB	Fraction of bold text with all query terms	tRatio	Fraction of #words and #tags in html

for entire document and summary containing query terms.

- Finally, query term specific features are used to capture relevance of document with respect to input query. These features include number of query terms in text and title, as well as their min, max and median frequencies in both document and summary. We also use min, max and median positions of query terms in both document and summary.

6.1.2 Html Features

Users interact more with rich and responsive web pages and their layout or structure can be instrumental in finding useful information. We use some features that leverage underlying information in html markup. These structure oriented features are given in Table 6.

- The first set of features are associated with tag distribution in a document. We posit that tag distribution captures how well information is organized in a web page. For instance, users may find it difficult to navigate pages with a lot of outlinks or images. Thus, we extract percentage of tables, images, headings, paragraphs, lists and outlinks from the web page.
- Outlink distribution of a page is useful because too many outlinks can be distracting and hinder readability of the document. We consider fraction of words in hyperlinks and words in text as feature. We also use fraction of links within a page, to same domain and other sites as features.
- Some parts of webpage tend to be more important and attract more attention [30] than others. For instance, headings are useful for skimming content. We use number of headings that contain query terms, their min, max and average position as features. Similar features are extracted from outlinks.
- A user may look for information by searching for query terms in a webpage. Hence, we also extract features from span (window) of text that contains all query terms. We use number of such spans in a document, their min, max and average positions, their average length and spans that cover headings as features.

Table 7: Findability Features

Feature	**p-val**	**Feature**	**p-val**
$fTable^-$	0.00	$minWinPos^+$	0.00
$avgSumChar^-$	0.01	$meanPosOut^+$	0.01
$docCLI^-$	0.02	$sumWords^+$	0.02
$maxWinPos^-$	0.04	$fImg^+$	0.02

We would like to emphasize that the features proposed above are by no means exhaustive. This is a first step in the direction of identifying features that could be significant for effort and more features can be used to capture different aspects of effort.

6.2 Predicting Findability

Given the aforementioned features, we focus on predicting effort through these features and analyse which features are significant for predicting effort. Since findability seems to be the most important factor for user satisfaction, we focus on predicting findability and compare and contrast the features that are important for predicting findability with features that are important for predicting relevance.

We use Ordinal Logistic Regression with normalized feature values ($\mu = 0$, $\sigma^2 = 1$) to predict Findability labels obtained from effort judging (described in Section 4), and report Root Mean Squared Error (RMSE) to measure the quality of predictions. For validation of regression analysis in predicting labels, along the same lines of the analysis done in Table 4, we compare the agreement of predicted Findability labels with preference judgements obtained from assessors by computing the fraction of documents preferred by the users that are predicted to be of high findability according to the regression model. To summarize, preference agreement is calculated as follows: given predicted grade of two documents, what is direction of preference (document with higher findability grade will be preferred over a document with a lower findability grade).

RMSE for the predictions is 0.37. We also compute preference agreement between predicted Findability grades. Preference agreement for predicted Findability grades is 0.587, which is comparable to agreement of 0.6 between Findability and preference judgements if actual judgements of Findabil-

Table 6: Html features used for predicting findability and relevance

Dom Oriented Features			
fHead	Fraction of headings (h1,h2..h6)	fBoldItalics	Fraction of bold, italics and strong
fTable	Fraction of tables	fOutlinks	Fraction of outlinks
fDiv	Fraction of Divs	fImg	Fraction of images
fPara	Fraction of paragraphs	fList	Fraction of Lists
Outlink Oriented Features			
fSameDomain	Fraction of hrefs to same domain	aRatio	Normalized #words in hyperlinks
fDiffDomain	Fraction of hrefs to different domain	aTxtRatio	Fraction of words in hyperlinks
fOutPage	Fraction of hrefs to same page		and text tags
Query Term Window Specific Features			
qWinH	Fraction of headings with all query terms	minWinPos	Min window pos with all query terms
qWinO	Fraction of outlinks with all query terms	maxWinPos	Max window pos with all query terms
qWinB	Fraction of bold text with all query terms	meanWinPos	Mean window pos with all query terms
Query Specific Features			
minPosH	Min pos of heading with query terms	minPosOut	Min pos of outlink with query terms
maxPosH	Max pos of heading with query terms	maxPosOut	Max pos of outlink with query terms
meanPosH	Mean pos of heading with query terms	meanPosOut	Mean pos of outlink with query terms
countH	#Headings with query terms	countOut	#Outlinks with query terms

ity were used in the analysis (in Table 4). These results suggest that the model can reliably predict Findability.

Table 7 contains statistically significant features (p-val $<$ 0.05) and direction of correlation coefficient. We posit that features that can help users find information quickly would be more important for Findability. We observe that both html and text features proposed above are important. Given that images and tables are useful in spotting information on a webpage, they (fImg and fTable) are significant in predicting Findability. While too many tables may make it difficult to find information, documents with more images have higher Findability grade. As expected, features such as minimum position of query terms in summary (minWinPos) and number of words (sumWords) in summary are also significant since these are directly correlated with the amount of effort needed to find the relevant information in the page. These results emphasize that position of query specific information on the page is important for reading/skimming entire document.

Our hypothesis is that if one solely focuses on predicting relevance, the features that are important for that purpose are likely to be different than the features that are important for findability, suggesting that retrieval systems need to use additional features such as the ones proposed in this paper in order to optimize for effort together with relevance. In order to validate our hypothesis, in next section, we use the aforementioned features to predict relevance and analyse the importance of features and how they differ from Findability features.

6.3 Relevance Prediction

Similar to the model for predicting Findability, we use our proposed features for predicting judgements of relevance obtained via the effort based judging interface. We use Ordinal Logistic Regression with normalized feature values ($\mu = 0$, $\sigma^2 = 1$) to predict relevance which results in 0.41 RMSE. Confusion matrix for the model is given in Table 8, where each cell is fraction of documents with actual label x_i and predicted label y_i. Table 9 shows features that are statistically significant (p-val $<$ 0.05) for predicting relevance, together with the direction of correlations.

Table 8: Actual vs. Predicted Relevance Labels

		Predicted		
		1	**2**	**3**
Actual	**1**	0	0.8	0.2
	2	0.05	0.59	0.34
	3	0	0.13	0.86

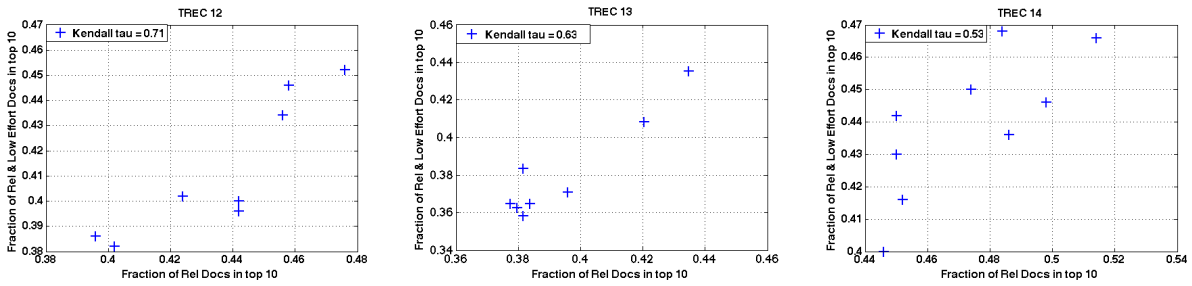
Table 9: Relevance Feature Importance

Feature	**p-val**	**Feature**	**p-val**
$tRatio^-$	0.01	$termsInTitle^+$	0.01
$countH^-$	0.02	$qWinO^+$	0.01
$maxWinPos^-$	0.04	$fBoldItalics^+$	0.02
$docWords^-$	0.04	$fImg^+$	0.04

As shown in previous work [38], document content features impact relevance most followed by query and structure specific features. While features related to document length (docWords and maxWinPos) have negative coefficients (suggesting user preference for documents with fewer terms and sentences), query and summary specific features (qWinO, sumWords and termsInTitle) have positive coefficients. Documents with higher text to tag ratio (tRatio) have lower relevance suggesting that verbose documents which lack structure are not preferred by users. At the same time, it is interesting to note that users do not prefer documents that contain a lot of headings as documents with more headings (countH) have lower relevance grade.

Table 9 also suggests that important features for predicting relevance are different than features that are significant for predicting Findability.

Overall, our results confirm our hypothesis that 1) retrieval systems that are optimized for relevance are not necessarily optimizing for effort, 2) in order to build retrieval systems that optimize for user satisfaction, systems should



(a) Top 10 systems of 2012 (b) Top 10 systems of 2013 (c) Top 10 systems of 2014

Figure 3: Comparison of systems based on #relevant documents vs #low effort relevant documents (p@10)

be optimized for Findability together with relevance, and 3) additional features that capture ease of finding information in the page (such as the ones proposed in this paper) should be used in building and optimizing retrieval systems.

In the following section, we focus on evaluating the quality of retrieval systems and show how incorporating effort into retrieval evaluation could lead to very different conclusions in terms of the quality of retrieval systems.

7. EFFECT OF EFFORT ON RETRIEVAL EVALUATION

Until now, we have focused on getting relevance judgements associated with effort and have shown that user satisfaction and preferences can be significantly affected by effort related factors, in particular, by ease of finding information in a document. Since the primary goal in retrieval evaluation is to measure user satisfaction, our results suggest that effort should be incorporated into retrieval evaluation.

Previous work [41] has shown that variations in relevance assessments does not necessarily lead to significant differences in retrieval evaluation. Given this finding, we further analyse whether incorporating effort as a factor in retrieval evaluation could lead to significant differences in the evaluation of systems. For this purpose, we use data from TREC Adhoc task 2012 to 2014. Getting effort based judgements for these years would be very costly and time consuming. Since our results suggest that findability information is a primary factor that affects user satisfaction and that it is possible to predict findability with a good accuracy, we used the regressor designed in the previous section in order to predict the easiness of finding information in a document. Focusing on the top systems submitted to these years, we then analyse how their performance would change if easiness of finding information in a document was incorporated into retrieval evaluation. For this purpose, we first evaluated the fraction of relevant documents retrieved by these top performing systems in top 10 (i.e., precision at 10 value). We then compared this value with the fraction of relevant documents retrieved in top 10 that are also low effort (i.e., findability). Figure 3 shows the result of this experiment for TREC 2012 (left plot), 2013 (middle plot) and 2014 (right plot). The plots also show the Kendall's tau correlation between the ranking of systems obtained when the systems are ranked based on number of relevant documents versus number of low effort relevant documents retrieved in top 10.

It can be seen that top performing retrieval systems tend to vary significantly in terms of effort needed to find relevant information and that even if two systems may retrieve identical number of relevant documents in top 10, their performance may be very different from each other when ease of finding information in the document is considered. For example, for TREC 2012, the fourth and fifth best performing systems in terms of fraction of relevant documents retrieved in top 10 seem to have retrieved almost identical number of relevant documents in top 10, whereas when the effort to find relevant information is also considered as a factor, their performance seems to be different from each other. The same behaviour can be seen in TREC 2014 for the third and fourth best systems according to the number of relevant documents retrieved in top 10. In this case, there is a big gap in the performance of these systems when effort to find relevant information is considered. Given the importance of findability in a document for user satisfaction, the satisfaction of the users of these two search systems would be very different from each other even though they retrieved similar number of relevant documents in top 10.

Overall, our results suggest that when effort to find relevant information is considered, performance of retrieval systems could be quite different as opposed to just focusing on relevance. Therefore, new evaluation metrics that incorporate effort together with relevance are needed for building retrieval methodologies that are better aligned with user satisfaction.

8. CONCLUSION

Extensive research has been carried out in the past on characterizing relevance. If the underlying assumption is that user satisfaction is impacted by relevance then outcome of highly ranked systems on basis of relevance judgements, should align with satisfaction of real users. However, it has been shown that relevance and user satisfaction do not always agree, and users may still be dissatisfied with their search despite being served relevant documents. Recent work [44] showed that the utility of a document with respect to an actual user can be different than its relevance, which in turn impacts user satisfaction. Their work leaves room for further research as they do not explain what constitutes effort or how can effort judgements be obtained and incorporated in evaluation. With this work, we attempted to answer all these questions.

We proposed three characteristics that could be useful in measuring effort, mainly – Findability, Readability and Understandability. To evaluate these factors we conducted two user studies – an effort based study where we asked for explicit grades for these parameters and a follow-up preference

study to validate whether effort parameters align with user preference. Our analysis indicates findability correlates well with user satisfaction among three parameters.

Having shown that findability is a reasonable predictor of user satisfaction, we compare important features for predicting findability with those useful for predicting relevance. Again, we observe useful predictors for findability and relevance capture different aspects. Towards the end, we analyse whether incorporating effort as a factor in retrieval evaluation could lead to significant differences in the evaluation of systems. Comparison of top performing runs on TREC Web track datasets of 2012-2014 suggests that performance of retrieval systems could be quite different when effort (in our experiments findability) is taken into account.

Our analysis suggests that effort based judgements can be explicitly collected from end users and can also be used to evaluate retrieval systems. There are several directions in which this work could progress. It would be interesting to analyse different label aggregation strategies to incorporate all the effort parameters. We would also look into incorporating effort into existing evaluation metrics or proposing new effort based metrics for retrieval evaluation.

9. REFERENCES

[1] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. The good and the bad system: Does the test collection predict users' effectiveness? In *Proc. SIGIR*, 2008.

[2] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proc. SIGIR*, 2005.

[3] O. Alonso and S. Mizzaro. Using crowdsourcing for {TREC} relevance assessment. *Inf. Process. Manage.*, 2012.

[4] P. Borlund. The concept of relevance in ir. *J. Assoc. Inf. Sci. Technol.*, 54(10):913–925, 2003.

[5] J. Brown and M. Eskenazi. Student, text and curriculum modeling for reader-specific document retrieval. In *Proceedings of the IASTED International Conference on Human-Computer Interaction. Phoenix, AZ*, 2005.

[6] B. Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proc. SIGIR*. ACM, 2011.

[7] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: Preference judgments for relevance. In *Proc. ECIR*, 2008.

[8] P. Chandar, W. Webber, and B. Carterette. Document features predicting assessor disagreement. In *Proc. SIGIR*, 2013.

[9] M. Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *J. Appl. Psychol.*, 1975.

[10] K. Collins-Thompson. Enriching the web by modeling reading difficulty. In *Proc. ESAIR*, ESAIR '13.

[11] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *Proc. CIKM*, 2011.

[12] C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting. *Inf. Process. Manage.*, 2012.

[13] D. DeStefano and J.-A. LeFevre. Cognitive load in hypertext reading: A review. *Computers in Human Behavior*, 2007.

[14] N. Ferro, G. Silvello, H. Keskustalo, A. Pirkola, and K. JÃd'rvelin. The twist measure for ir evaluation: Taking user's effort into account. *J. Assoc. Inf. Sci. Technol.*, 2015.

[15] Q. Guo and E. Agichtein. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. In *Proc. WWW*, 2012.

[16] J. Gwizdka and I. Lopatovska. The role of subjective factors in the information search process. *Journal of the American Society for Information Science and Technology*, 2009.

[17] M. Halvey and R. Villa. Evaluating the effort involved in relevance assessments for images. In *Proc. SIGIR*, 2014.

[18] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 2007.

[19] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *Proc. SIGIR*, 2000.

[20] S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *Proc. SIGIR*, 2007.

[21] F. C. Johnson, J. R. Griffiths, and R. J. Hartley. Task dimensions of user evaluation of information retrieval systems. *Information research*, 2003.

[22] G. Kazai, N. Craswell, E. Yilmaz, and S. Tahaghoghi. An analysis of systematic judging errors in information retrieval. In *Proc. CIKM*, 2012.

[23] G. Kazai, J. Kamps, and N. Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Inf. Retr.*, 2013.

[24] P. Kidwell, G. Lebanon, and K. Collins-Thompson. Statistical estimation of word acquisition with application to readability prediction. *JASA*, 2011.

[25] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proc. WSDM*, 2012.

[26] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proc. SIGIR*, 1994.

[27] M. Mosconi, M. Porta, and A. Ravarelli. On-line newspapers and multimedia content: An eye tracking study. In *Proc. SIGDOC*, 2008.

[28] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, 2005.

[29] P. L. T. Pirolli. *Information Foraging Theory: Adaptive Interaction with Information*. 2007.

[30] O. Rokhlenko, N. Golbandi, R. Lempel, and L. Leibovich. Engagement-based user attention distribution on web article pages. In *Proc. HT*, 2013.

[31] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestations of relevance. *J. Assoc. Inf. Sci. Technol.*, 58(13), 2007.

[32] P. Schaer. Better than their reputation? on the reliability of relevance assessments with students. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, Lecture Notes in Computer Science. 2012.

[33] L. Schamber and M. Eisenberg. Relevance: The search for a definition. 1988.

[34] R. Senter and E. Smith. Automated readability index. Technical report, DTIC Document, 1967.

[35] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, 2012.

[36] C. Tan, E. Gabrilovich, and B. Pang. To each his own: Personalized content selection based on text comprehensibility. In *Proceedings of WSDM*, 2012.

[37] A. Taylor. User relevance criteria choices and the information search process. *Inf. Process. Manage.*, 48, 2012.

[38] A. Tombros, I. Ruthven, and J. M. Jose. How users assess web pages for information seeking. *J. Assoc. Inf. Sci. Technol.*, 2005.

[39] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results? In *Proc. SIGIR*, 2001.

[40] R. Villa and M. Halvey. Is relevance hard work?: evaluating the effort of making relevant assessments. In *Proc. SIGIR*, 2013.

[41] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Process. Manage.*, 36(5), 2000.

[42] H. Weinreich, H. Obendorf, E. Herder, and M. Mayer. Not quite the average: An empirical study of web use. *ACM Trans. Web*, 2(1):5:1–5:31, Mar. 2008.

[43] Y. C. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *J. Assoc. Inf. Sci. Technol.*, 57(7), 2006.

[44] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: An analysis of document utility. In *Proc. CIKM*, 2014.

[45] Y. Zhang, J. Zhang, M. Lease, and J. Gwizdka. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proc. SIGIR*, 2014.