

Going Beyond Relevance: Role of effort in Information Retrieval

Manisha Verma

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science
University College London

I, Manisha Verma, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

The primary focus of Information Retrieval (IR) systems has been to optimize for Relevance. Existing approaches to rank documents or evaluate IR systems does not account for “*user effort*”. Currently, judges only determine whether the information provided in a given document would satisfy the underlying information need in a query. The current mechanism of obtaining relevance judgments does not account for time and effort that an end user must put forth to consume its content. While a judge may spend a lot of time assessing a document, an impatient user may not devote the same amount of time and effort to consume its content. This problem is exacerbated on smaller devices like mobile. While on mobile or tablets, with limited interaction, users may not put in too much effort in finding information.

This thesis characterizes and incorporates *effort* in Information Retrieval. Comparison of explicit and implicit relevance judgments across several datasets reveals that certain documents are marked relevant by the judges but are of low utility to an end user. Experiments indicate that document-level effort features can reliably predict the mismatch between dwell time and judging time of documents. Explicit and preference-based judgments were collected to determine which factors associated with effort agreed the most with user satisfaction. The *ability to locate relevant information* or *findability* was found to be in highest agreement with preference judgments. Findability judgments were also gathered to study the association of different annotator, query or document related properties with effort judgments. We also investigate how can existing systems be optimized for relevance *and* effort. Finally, we investigate the role of effort on smaller devices with the help of cost-benefit models.

Acknowledgements

As I write this, I am filled with gratitude, excitement and some nostalgia. Last four years were not only intellectually but sometimes physically and emotionally challenging to say the least. I have learned to fail multiple times, change my course midway and yet persist to finish what I started. This thesis marks an end to an eventful journey. That said, I have to thank my advisor Emine Yilmaz, for her patience and encouragement when I had most certainly decided to quit (which happened a lot!) through the course of this PhD. We have discussed several ideas during my study at UCL, which has been rewarding in several ways. I have to thank her for all the opportunities and responsibilities she gave us as a group.

I must also thank Sebastian Riedal, for building a formidable and immensely talented group. His encouragement and advice were extremely useful and I promise to imbibe some of his qualities as a researcher in the future.

I have to thank so many PhD students at UCL, without whom it would have been impossible to navigate the perils *and* celebrate the merits of PhD. Marzieh, you are one of the first people I spoke to, when I came to 1ES. You have been a source of *inspiration* and *strength*! I have had countless discussions about research, academia and everything in between with the big gang of boys at UCL. Matko, Weinan, Shangsong and Rishabh, you have been a source of inspiration *and* competition whenever I needed to push myself. I have to thank Diego, Jiyin and Ke Zhou for collaborating with me on projects that I would not have attempted to do on my own. I also have to thank Jiyin for reviewing whatever needed reviewing so thoroughly and asking tough questions, most often forcing me to think harder and write better.

I have to thank Marc for being an excellent desk mate, senior and mentor. I learnt a lot about startups from Marc and Andy at ContextScout. As I write this, I have spent 2 eventful months in Yahoo! Research NYC, working on an exciting problem with an excellent team. I have to thank Yifan, Kevin and Meizu for an awesome summer. Summer in Google Zurich was equally eventful. I have to thank Aayush, Dan and Sophie for reminding me that having fun is sometimes as important as obtaining good results.

I have to thank Tuli and Abro, for you are my Monica and Chandler. That room will always be my room. Thank you for providing me a home whenever I chose to be homeless. I have to thank Mary Mitchell, for being there when no one was around. Thank you for being present during the hospital visits and constant food supply when I could barely sit up. You have taken care of me like a daughter when things got tough. Supi, Bhupi and Alexa, you made me do things I would have balked at even if shown in movies. Thank you for tolerating my tyranny yet providing the constant supply of refreshing energy, humour and idlis. I have to thank members of SGI for encouraging me to do better. Natasha and Michelle never ceased to push me to aim high. I have to thank friends at Python course, I enjoyed teaching it every single time! Nishank Mehra, you are not the best flatmate but I enjoyed your company until you ditched me to go back to India.

I have been blessed with a *huge* family that never tires of rooting for my success. I cannot summarize the support and strength you have provided in the 20000 words limit. I have to dedicate whatever I have done to eight people (in exactly this order): Baba, *Dad*², *Mum*², T, Shivu and Nisarg. I could not have asked for a better partner in crime and dear friend than Nisarg. This is for us and everything we spoke of while walking in IIIT at 4am every morning without fail.

Contents

1	Introduction	13
1.1	On document relevance	13
1.2	Evaluation of information retrieval systems	15
1.3	Search effort vs. user satisfaction in mobile	17
1.4	Problem statement	19
1.5	Contributions	19
1.6	Publications	20
2	Background	22
2.1	Relevance	23
2.1.1	Relevance criteria	24
2.2	Information retrieval evaluation	27
2.2.1	Offline evaluation	29
2.2.2	Online evaluation	30
2.2.3	Agreement between online/offline evaluation	32
2.3	Effort in information retrieval	34
2.3.1	Modeling effort in information seeking tasks	35
2.3.2	Characteristics of effort	38
2.3.3	Time based evaluation of effort	41
2.3.4	System based evaluation of effort	42
2.4	Mobile search	44
2.5	Conclusion	46

3	Relevance vs. document utility	48
3.1	User behavior and relevance	50
3.1.1	Document evaluation model	51
3.2	Experimental setup	53
3.2.1	Dataset collection	53
3.2.2	Labeling methodology	54
3.2.3	Time measurement	56
3.3	Experimental results	57
3.3.1	Utility versus relevance	58
3.3.2	Effect of effort on document utility	61
3.4	Conclusion	69
4	Effort based judgments in IR	70
4.1	Factors associated with effort	72
4.2	Effort based judging	74
4.3	Effort-Preference correlation	80
4.3.1	Preferences vs. effort characteristics	82
4.4	Predicting effort and relevance	83
4.4.1	Features	84
4.4.2	Predicting findability	86
4.4.3	Relevance prediction	88
4.5	Effect of effort on retrieval evaluation	89
4.6	Conclusion	91
5	Characteristics of effort judgments	93
5.1	Methodology	95
5.2	Results	98
5.2.1	Annotator specific analysis	99
5.2.2	Query-specific analysis	101
5.2.3	Document-specific analysis	103
5.2.4	Inferring implicit labels from judging time	108

5.3	Conclusion	112
6	Incorporating effort in ranking	113
6.1	Effort aware ranking	114
6.1.1	Effort aware SVMRank	115
6.1.2	Effort aware LambdaMart	116
6.2	Experimental setup	117
6.2.1	Features	117
6.2.2	Effort label generation	118
6.2.3	Datasets and evaluation metrics	119
6.2.4	Baselines and systems summary	121
6.3	Results and discussion	122
6.4	Conclusion	130
7	Search effort vs. satisfaction on mobiles devices	131
7.1	Overview of economic models in IIR	133
7.1.1	Query interaction	134
7.1.2	SERP interaction	134
7.2	User Study and data statistics	135
7.2.1	Search topics	136
7.2.2	SERP population and presentation	136
7.2.3	App interface	137
7.2.4	Participants	138
7.2.5	Observed variables	138
7.3	Cost/Benefit vs. satisfaction analysis	140
7.3.1	Query cost-benefit and user satisfaction	140
7.3.2	Search cost-benefit and user satisfaction	142
7.4	Conclusion	145
8	Conclusion	146
	Bibliography	151

List of Figures

2.1	An example of search query and ranked list of documents where each document is marked as relevance (green) or non-relevant (red).	27
2.2	Two paradigms of evaluation	30
3.1	(Left) Cumulative distribution of judgment time for crowd and expert judges versus dwell time, and (Right) judging time versus dwell time.	58
3.2	Percentage of low utility documents labeled as relevant versus difference between judging time and dwell time for the (left) CrowdJ-TrecQ , (middle) ExpertJ-TrecQ , and (right) CrowdJ-NaturalQ datasets.	60
4.1	Sample effort hit	75
4.2	An example hit for effort and preference judging	81
4.3	Comparison of systems based on #relevant documents vs #low effort relevant documents ($P@10$)	90
5.1	Variability in judging time with relevance labels for TREC Web topics	95
5.2	Judging interface	96
5.3	Attribute instruction	97
5.4	Highlight instructions	97
5.5	Instructions	97
5.6	% annotators not familiar with topics	98
5.7	Topic familiarity and findability labels	100
5.8	Topic type and findability labels	102

5.9	Answer words/sentences/location vs. effort	103
5.10	Judging time, answer sentences/location and effort	104
6.1	Judging time of low (left) and high (right) effort topics	119
6.2	Percentage improvement of $LMart_{rf}$ over $LMart_{rel}$ for all queries .	127
6.3	Percentage improvement of $SVMr_{rf}$ over $SVMr_{rel}$ for all queries . .	127
7.1	An example of different types of costs used in economic models of interaction.	133
7.2	Search result page samples	137
7.3	Topic layout and feedback screens	139
7.4	Query length and user satisfaction (left) and query interaction net profit (right)	141
7.5	Clicked documents, viewed snippets and user satisfaction	143
7.6	Cost of reading a snippet and clicked document	144
7.7	Net profit of reading a snippet and clicked document	144

List of Tables

2.1	Comparison of proposed relevance criteria	25
3.1	Sample query-document pairs from the datasets	54
3.2	Datasets used for analysis	56
3.3	Dwell time vs. judging time on various datasets	59
3.4	Document features associated with effort	62
3.5	Regression model for TRECQ and NaturalQ median dwell time prediction. * denotes predictors significant at the $p < 0.05$ level.	64
3.6	Significance of features for predicting the mismatch between utility and relevance for ExpertJ-TrecQ dataset.	65
3.7	Significance of features for predicting the difference between judging time and dwell time for CrowdJ-TrecQ dataset.	66
3.8	Significance of features for predicting the mismatch between utility and relevance for CrowdJ-NaturalQ dataset.	67
4.1	Effort label distribution	77
4.2	Inter-rater agreement	77
4.3	Factor importance for satisfaction	79
4.4	Preference and effort factors agreement	79
4.5	Text features used for predicting findability and relevance	84
4.6	Webpage structure features used for predicting findability and relevance	86
4.7	Findability features ¹	87
4.8	Relevance feature importance	88

4.9	Actual vs. predicted relevance labels	88
5.1	% annotators not familiar with query	98
5.2	P(find familiarity)	101
5.3	P(find topic type)	103
5.4	Feature distribution and correlation with judging time	105
5.5	Document features and judging time correlation for different labels .	107
5.6	Summary features and judging time correlation for different labels .	108
5.7	Pearson's Rho and Cohens Kappa	108
6.1	Relevance and Findability based features	116
6.2	Query and label distribution of 2011-2014	119
6.3	Low (high) effort queries	119
6.4	Relevance based evaluation of rankers for 2011-2014 Web Tracks .	123
6.5	Joint relevance and effort based evaluation of rankers for 2011-2014	125
6.6	Gain of rel+find joint models per query type	128
6.7	Feature weight determined using $NDCG_{rf}@20$	129
7.1	Topic descriptions	136
7.2	Data summary	140
7.3	Pearson's ρ between satisfaction & net query profit	141
7.4	Pearson's ρ between satisfaction & net search profit	141

Chapter 1

Introduction

1.1 On document relevance

Human beings frequently need to find and compare information, whether it is buying a new house or finding a new job. Over the years, the nature of information avenues has changed, from asking people or foraging print media to simply accessing information online. With rapid developments in Internet services and burgeoning online content, users can find required information within minutes. The Internet has evolved into a document¹ collection of 50 billion publicly accessible Web pages. With the active evolution of *search engines*, users can now concentrate on a tiny fraction of these documents to find desired information. Users express their need in form of a *query* (text, audio or image). Search engines measure the expected utility of documents, their so-called relevance with respect to a user's query and display highly scored documents (in turn highly relevant) for the user to browse.

The estimation of relevance is a very complex step for most collections and search settings. It goes beyond simple pattern matching between query terms and documents. In some cases, a document that does not contain all the search terms may still be relevant to the query if it addresses the underlying information need. For instance, a document about the flower '*Lilium longiflorum*' would be considered relevant to the query '*Easter lily*' even if query terms '*easter*' and '*lily*' do not appear in its text since '*Lilium longiflorum*' is the scientific term for '*Easter lily*'

¹In information retrieval, a document may refer to a very wide variety of things: books, websites, images, videos, or music tracks to name a few.

flower. Then, approaches that exploit semantic similarity [1, 2, 3] between query terms and document text are used to determine document relevance. Hardly any user can be expected to precisely know the exact documents they are searching for at the time of query input. According to Belkin [4], this so-called Anomalous State of Knowledge (ASK) requires users to have some notion of documents they need without knowing the full range of available information. In such scenarios, relevance models have to account for a considerable degree of uncertainty in the user-provided query and carefully interpret the available sources of evidence.

Researchers have tried to build consistent and universally applicable descriptions of relevance [5, 6, 7, 8, 9, 10] from two points of views: (1) system-based relevance and (2) subjective or user-based relevance. The system-driven approach treats relevance as a static and objective concept as opposed to the cognitive user-oriented approach that considers relevance to be a subjective and personalized measure of how a document addresses the underlying information need which may be task or situation dependent. Saracevic [11] distinguishes between five basic types of relevance. These are: (1) system or algorithmic relevance, which captures the relation between the query (terms) and the collection of retrieved document(s); (2) topical relevance which denotes the relation between subject or topic expressed in a query and subject or topic covered by documents; (3) cognitive relevance, related to the information need as perceived by the user; (4) situational relevance, depending on the task interpretation; and (5) motivational and affective, which describes the relation between intents, goals and motivations of the user and retrieved document(s). *Topicality* [12] has emerged as an important and frequently used factor to determine relevance. A document is considered relevant to the user's query if its content topically overlaps with the user's information need. It should be able to answer a user's query either partially or completely.

At present, Information Retrieval (IR) systems are designed to optimize for topical relevance. It is assumed that topically relevant documents shall answer user's information need, which in turn will yield higher user satisfaction. New algorithms [13, 14, 15] for ranking documents are designed such that most relevant

documents appear at the top of the Search Engine Result Page (SERP). Evaluation metrics [16, 17, 18, 19, 20] are also designed to compare systems on the basis of how many relevant documents are retrieved and how high are they shown on SERPs.

However, one may argue that relevance is a complex concept, which goes beyond *topicality*, that might vary with users and situations. Rightly so, researchers have shown that, besides topicality, several other factors constitute relevance: page authority, novelty or freshness, scope etc. We can find a wide range of empirical studies investigating the distribution, nature, and dynamics of relevance and how people assess it. Examples include: Barry *et al.* [21], Wang *et al.* [22], Tombros *et al.* [23] and Xu *et al.* [24]. These studies unanimously describe relevance as a composite notion, suggesting that topicality on its own is not sufficient to reliably judge document relevance.

To summarize, while simplified notion of *topicality* is often used to denote document relevance, a large number of studies have shown that *document relevance* is in fact a composite and multi-dimensional concept [21, 22, 23, 24] which should be incorporated in design and evaluation of information retrieval systems.

1.2 Evaluation of information retrieval systems

Search engines today return a ranked list of documents for a given user query. New algorithms or models can be evaluated on a set of queries and documents, where each query is associated with a list of documents. Here, document relevance with respect to the input query can be used to evaluate overall system effectiveness. At present, there are two popular paradigms of evaluation in Information Retrieval research. On one end we use small test collections to measure the performance of a system, while on the other end, we evaluate system effectiveness in the wild with search data obtained from live users at a large scale.

Evaluation of system effectiveness based on pre-designed small-scale test collections is known as *batch evaluation* or *explicit evaluation*. These *test collections* are a corpus of documents, where a subset of the corpus is *manually judged* for relevance with respect to a small set of queries. **Trained judges** are provided with

certain guidelines to determine document relevance and are asked to evaluate an individual or a pair of documents with respect to a query. The manual judgment of relevance is a time-consuming process which results in small but re-usable collections. While relevance is a composite notion, it is still currently captured by a *single grade* at judging time. Thus, if we were to evaluate IR systems more effectively, we would need test collections with labels other than relevance that capture more dimensions than just topicality.

The second approach to evaluate system effectiveness is to test it with live users. *User-based evaluation* relies on observing and measuring user's interaction with a document to determine its relevance. It is more time consuming and user data is not very reliable due to the presentation [25], click [26] and user bias [27]. This is also known as *implicit evaluation* where the system *implicitly* determines document relevance using either user behavior or time on a clicked document with respect to an input search query.

One would expect that batch evaluation would agree with the user-based evaluation of systems. But it has been shown in the past [28, 29, 30, 31, 32, 33, 34] that these two forms of evaluation do not agree with each other. These studies did not establish any direct correlation between user satisfaction reported by the users and the number of relevant documents in search results. They observed that improvements in test collection based evaluation do not always translate into a direct benefit for the end users (as measured by the number of relevant documents). Hersh *et al.* [28] observed that user satisfaction and implicit relevance judgments do not always correlate which suggests that relevance judgments do not completely capture all the aspects that might affect user satisfaction.

However, there are also studies [35, 36] that have found higher agreement between system effectiveness and user performance measures. But these studies have also shown that users actively adapt to the performance of a retrieval system. For instance, Smucker *et al.* [36] found that users change their behavior depending on the precision of the results list. Maskari *et al.* [35] found varied correlations between user effectiveness and evaluation metrics such as Precision and Average

Precision (AP). For instance, they found a weak correlation between AP and user effectiveness and high correlation between Precision and user effectiveness.

To summarize, existing evaluation mechanisms (batch and user-based) do not agree [28, 29, 30, 31, 32, 33, 34] with each other or agree with each other [36, 35] when systems under comparison have wide differences. We believe these differences arise because the judged relevance of a document does not comply with user's expectation from the document. On the basis of existing literature and its shortcomings, this thesis investigates the following hypothesis:

Explicit and implicit evaluation of systems do not align because existing relevance judgment paradigm does not account for '*Effort*' required to locate, read and digest *relevant* information from a given document.

Judges do not consider document utility with respect to an end user while judging documents. Trained judges are asked to identify document relevance regardless of how much *time and effort* it may take to consume it. While a judge can take several seconds, even minutes to evaluate a document, an end user may not be willing to spend as much time consuming it, even if it is relevant. Thus, despite being relevant, the document is of minimal utility to the user if they can not find the required information quickly or cannot properly understand it. We believe that the mismatch between explicit and implicit evaluation is an outcome of *effort* needed to locate, read and digest relevant information from a given document is different for judges and users.

1.3 Search effort vs. user satisfaction in mobile

Search is best summarized as an interactive process. Some search goals/tasks need users to issue some queries, read several snippets and click multiple results to find the required information. When an information need is underspecified or has several components, one query may not be sufficient to find a satisfactory answer. In such cases, a user may issue multiple queries or examine multiple documents to find relevant information. This series of queries and clicked documents constitute a search *session*. User satisfaction is one way of measuring search success which

is modeled as a function of several features derived from user behavior [37, 38] as number of queries issued, documents clicked or dwell time on clicked pages or collected explicitly from users [39].

However, recently a class of formal models have been proposed to model user interactions, more specifically in terms of cost (or effort) and benefit analysis using econometrics. Several models have been proposed [5, 40, 41, 42, 43] to compute the overall gain or net search success of a user. Some of these models have also been empirically evaluated [44, 45] with real user data. We believe that such models are an effective way to formally model effort from a user's perspective. Until now, we motivated how effort may be an important factor besides relevance but left out discussion on modeling effort spanning multiple queries and documents, i.e. a *search session*.

Our hypothesis is that existing work on cost-benefit models is useful in measuring both success *and* effort (in terms of cost) invested by the user in searching for relevant information on a search engine. However, current empirical evaluations are mostly restricted to a desktop setting. For instance, Azzopardi *et al.* [44] evaluate the cost of issuing search queries. In [46], authors consider the economics (in terms of gain and cost) to examine the interplay between querying and assessing. In this thesis, however, we evaluate the utility of existing models in mobile search. We particularly investigate the correlation between net benefit calculated using query and session cost-benefit models and explicit user satisfaction reported by the users. Query model explores the relationship between net gain and cost associated with issuing queries of different lengths on mobile. Session model evaluates the cost and gain from different actions such as querying, snippet and clicked result examination during session search. Our second hypothesis is as follows:

Existing cost-benefit models can be used to model *session effort* or success but desktop models are not directly useful on mobile.

More specifically, we investigate how existing cost-benefit models of economics correlate with user effort and satisfaction in mobile search which is absent from existing literature. Such analysis also provides an insight into which mobile specific

effort (or cost) parameters need to be incorporated into existing models.

1.4 Problem statement

With the above mentioned motivations we explore ways to define, characterize and incorporate *effort* in IR. We explore the characteristics of effort, gather effort based judgments and propose models that also incorporate effort in ranking. Finally, we investigate the role of relevance and effort on mobile to understand difference between desktops and mobile. We attempt to answer the following research questions:

RQ1. Can we empirically evaluate the role of effort in explaining the mismatch between batch and online evaluation?

RQ2. Determine which factors associated with *effort* can effectively distinguish between two *equally relevant* documents and correlate with satisfaction?

RQ3. Investigate whether annotator, query or document specific properties affect these effort judgments of relevant documents?

RQ4. How do rankings derived from explicit effort labels differ from those generated by effort labels derived from judging time of the document?

RQ5. How to account for effort in learning-to-rank [47] models and evaluate them with respect to relevance-based models?

RQ6. Do existing desktop based cost-benefit analysis models empirically correlate with user satisfaction in mobile search?

1.5 Contributions

The primary contributions of this thesis are as follows:

- In Chapter 3, we empirically evaluate the role of effort in addressing the gap between offline and online evaluation with three datasets. We also demonstrate the significance of effort specific features in predicting the mismatch between judged document relevance versus its utility to an end user.
- In Chapter 4, we identify which parameters constitute *effort*. We collected judgments for three parameters: easiness to read (*'readability'*), ease of finding (*'findability'*) and easiness to understand (*'understandability'*) the rele-

vant portions of the document. Explicit labels and preference labels suggest that *findability* is the most important factor associated with effort which highly correlates with satisfaction labels.

- In Chapter 5, we investigate which properties of a judge/query or document may affect effort judgments. Our analysis clearly indicates that judges take more time to judge *high effort* documents than *low effort* documents.
- Based on the findings in Chapter 5, we use judging time as effort labels. We also propose and evaluate two pairwise learning-to-rank models that optimize for both relevance and effort in Chapter 6. The proposed models perform well on two out of four datasets when evaluated on basis of rank biased and time biased metrics.
- In Chapter 7, we conduct a user study to collect search session logs to empirically evaluate the effectiveness of existing cost-benefit models in mobile search. We found that optimal parameters of these models differ from desktops and that satisfaction is better correlated with net query profit but weakly correlated with net search profit in mobile.

1.6 Publications

The following publications are based on the work presented in this thesis:

- Yilmaz, E., Verma, M., Craswell, N., Radlinski, F., & Bailey, P. Relevance and effort: An analysis of document utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, ACM.
- Verma, M., Yilmaz, E., & Craswell, N. On obtaining effort based judgments for information retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 2016, ACM.
- Verma, M., & Yilmaz, E. Search Costs vs. User Satisfaction on Mobile. In *European Conference on Information Retrieval*, 2017, Springer International

Publishing.

- Verma, M., & Yilmaz, E. On finding relevant information quickly. *Under Review*
- Verma, M., Yilmaz, E., & Craswell, N. Study of relevance and effort judgments across devices. In *Proceedings of the 2018 ACM on Conference on Human Information Interaction and Retrieval*, 2018, ACM.

Following publications have inspired this work but are not directly used in this thesis:

- Verma, M. Going Beyond Relevance: Incorporating Effort in Information Retrieval. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, ACM.
- Verma, M., & Yilmaz, E. Characterizing Relevance on Mobile and Desktop. In *European Conference on Information Retrieval*, 2016, Springer International Publishing.
- Verma, M., & Yilmaz, E. Entity oriented task extraction from query logs. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, ACM.
- Verma, M., & Yilmaz, E. Category Oriented Task Extraction. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, 2016, ACM.
- Verma, M., Yilmaz, E., Mehrotra, R., Kanoulas, E., Carterette, B., Craswell, N., & Bailey, P. *Overview of the TREC Tasks Track 2016*, In TREC, 2016, TREC.

Chapter 2

Background

Internet and search engines have together facilitated a world where users have copious amounts of digital information at their disposal. However, users can still consume only a small amount of information at any given (limited) time. Thus, to efficiently sift through billions of documents, users are provided with a search box and 10 blue links as results by search engines today. One can search entire web with text queries and is in turn provided with an ordered list of multiple links and snippets as results. It is important to provide users with documents that satisfy their information need to maintain user satisfaction and avoid user abandonment. Of course, users may be satisfied by search system only when displayed documents are relevant to the issued query.

Given an information need, *relevance* provides the basis for determining what information is likely to satisfy those needs, thus what information is worthy of retrieval. Researchers rely on **relevance** to design algorithms that retrieve documents that address the underlying information need of a query. Documents are graded or evaluated with respect to a query on basis of their relevance. Ideally, a system should rank highly relevant and irrelevant documents on the top and bottom of the list, respectively. Thus, one can distinguish between two ranking algorithms by analyzing the number of relevant documents retrieved and their positions in the ranked list. Relevance, thus, sits at the core of Information retrieval (IR) systems. Evaluation metrics explicitly use document relevance to compare two IR systems.

Given the importance of relevance, significant work exists on defining rele-

vance, identifying parameters that constitute relevance, i.e. factors important for a user to determine web page or document relevance with respect to a query. In the following subsection, we briefly cover the proposed definitions of relevance and factors important for users.

2.1 Relevance

Early work [10, 8, 6, 48, 49] focused on defining and scoping relevance in Information seeking and retrieval. Broadly, two definitions of relevance [6, 11] have been proposed: (1) objective or system-based relevance; and (2) subjective or human (user)-based relevance. The system-driven approach treats relevance as a static and objective concept based on objective measures of topical similarity, as opposed to the cognitive user-oriented approach that considers relevance to be a function of the applicability of the information to a user's need, problem, situation, and context, based on the user's subjective judgment.

Cooper *et al.* [49, 48] proposed a restricted definition of relevance, where the relevance of a sentence to an information need is dependent entirely on whether an answer to the need can be deduced from it or not. They also touch on conditional relevance, in which case one document is relevant for an information need only in the presence of another document. They also compare their logical relevance definition to *utility*, where they argue that it is not only the relevance of the document that determines its usefulness but also the *ease* with which this *relevance can be detected* by the system or the user. This a key point of discussion in upcoming chapters as we study effort and incorporate it in retrieval.

From user's perspective, it has been repeatedly proposed and validated [10, 6, 8] that relevance is a dynamic, situated and multi-dimensional concept. Schamber *et al.* [8] concluded that relevance is a multidimensional concept that is affected by both internal and external factors. They discuss how topicality is not enough to capture relevance of a document. They also discuss factors like utility and satisfaction in detail. Their primary suggestion was that relevance, while complex, can be measured if approached from user's perspective. Borlund [10] studies this

further. They summarize propositions on relevance criterion, degrees, and levels of relevance. They especially focus on situational relevance. They study the relationship between relevance and development of information need by studying user interaction with the search engine over time.

2.1.1 Relevance criteria

Some empirical work explores factors affecting user's relevance assessments of a documents with respect to a query. A large set of criteria have been identified, many of which relate to non-topical characteristics, aspects of the searcher and the situation. Table 2.1 summarizes relevance criteria studied in recent work [12, 8, 22, 23, 24, 50] in ascending order of publication year. Considerable overlaps in the criteria across studies suggest that strong general patterns exist across users to determine document relevance.

Tombros *et al.* [23] studied the impact of two situational variables on relevance criteria: task type and task stage. The study involved 24 participants, each searching for three controlled tasks on the Internet, using their preferred method of searching. The study points to variations between the features and criteria used by searchers to assess relevance based on task. A similar experimental web searching study by Kelly *et al.* [51] tested a pre-determined set of document features and found that different elements of web pages were used to assess relevance for different types of search tasks.

Xu *et al.* [24] conducted a study to investigate criterion that users employ to make relevance judgments. They proposed that topicality, novelty, reliability, understandability, and scope characterize relevance. They found that topicality and novelty were two most important dimensions for relevance judgments, followed by understandability and reliability.

Saracevic [11] in his synthesis of several decades of work proposes seven groups of relevance criteria: content, objects, validity, situational match, cognitive match, affective match, belief match. He mentions that effort should be considered when relevance with respect to a user is defined. He further provides a definition of the theory of relevance to an individual and incorporates effort into this definition.

Schamber <i>et al.</i> [8]	Wang <i>et al.</i> [22]	Tombros <i>et al.</i> [23]	Xu <i>et al.</i> [24]	Taylor <i>et al.</i> [50]	Zhang <i>et al.</i> [12]
Accuracy Currency Specificity Geographic area Reliability Accessibility Verifiability Clarity Dynamism Presentation Quality	Topicality Orientation Discipline Novelty Expected quality Recency Availability Special requisite Authority Relation/origin	Text content content numbers titles query terms amount of text Structure layout links Quality scope/depth authority recency general quality novelty pictures Physical properties file errors language, connection speed subscription	Scope Novelty Reliability Topicality Understandability	Accuracy Advertisement Affectiveness Authority Bias Breadth Definitions Depth Descriptions Guidelines History Novelty Recency Source Structure Time Tips Topic Understandability	Topicality. Novelty Understandability Scope Reliability

Table 2.1: Comparison of proposed relevance criteria

Taylor’s work [50] 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 judgments.

Zhang *et al.* [12] investigated five factors: a) novelty, (b) topicality, (c) understandability, (d) scope and (e) reliability. While *topicality* captures how related the document is to the topic of information need, *scope* characterizes how broad and specific document is to satisfy the given information need. Typically, information must be perceived as accurate to be considered relevant, this captures ‘*reliability*’. Their study showed that scope and novelty did not affect relevance. They also found understandability did not explain relevance judgments as completely as novelty and topicality did.

There also exists some work on aggregating relevance judgments obtained for several criteria. For instance, Costa *et al.* [52] propose a new model to aggregate multiple criteria evaluations for relevance judgments. They conclude that aboutness, coverage, appropriateness and reliability estimate document relevance.

Above list of relevance judgments criteria is rather long and has some limitations. Firstly, there are typically many factors that affect user’s notion of relevance. It is infeasible to ask users to assess each of these factors individually. Secondly,

different studies use synonyms to describe similar concepts (for example utility and usefulness [53]) and some factors almost entirely overlap in their meaning (for example new, novel and recent). Some may have subtle differences that may confuse the annotator or user which makes it difficult to determine their impact or effect across studies. Another limitation of these studies is that they seldom identify dependencies among different situations or conditions under which each factor may become more or less important to users. For instance, novelty and authority may be a primary factor in the beginning of a search session but time would be come important if user is under pressure to find information quickly.

With explosion of content on Internet, some other aspects have also emerged to be crucial for an end user to assess a document. Some researchers have worked on identifying factors besides relevance, that may influence user's interaction with the search engine and further help in improving user satisfaction. For example, *Recency* [54, 55, 56, 57] of the document may influence and dominate ranking algorithms in special verticals such as news. While these parameters can be encoded as features to predict system-based relevance, *manual evaluation* of each document for such an exhaustive list of parameters is impossible. Thus, it is useful to know which primary factors besides relevance are important and perhaps *only* rate each query-document pair for these factors.

Finally, while this exhaustive list is useful in differentiating between a relevant and non-relevant document, existing work does not determine which factors are of importance when *two equally relevant documents* are compared with each other. In this thesis, we argue that today, systems also need to differentiate between *two relevant documents*.

With burgeoning content on the Internet and rapid improvement in ranking algorithms, today there may be *multiple equally relevant* documents for a search query. Therefore, a user then would draw maximum value from a document that is not only relevant but also requires *less effort/time* to locate and consume the relevant information. This is even more important on small hand-held devices such as mobile or smart watches where both device accessibility and available time may

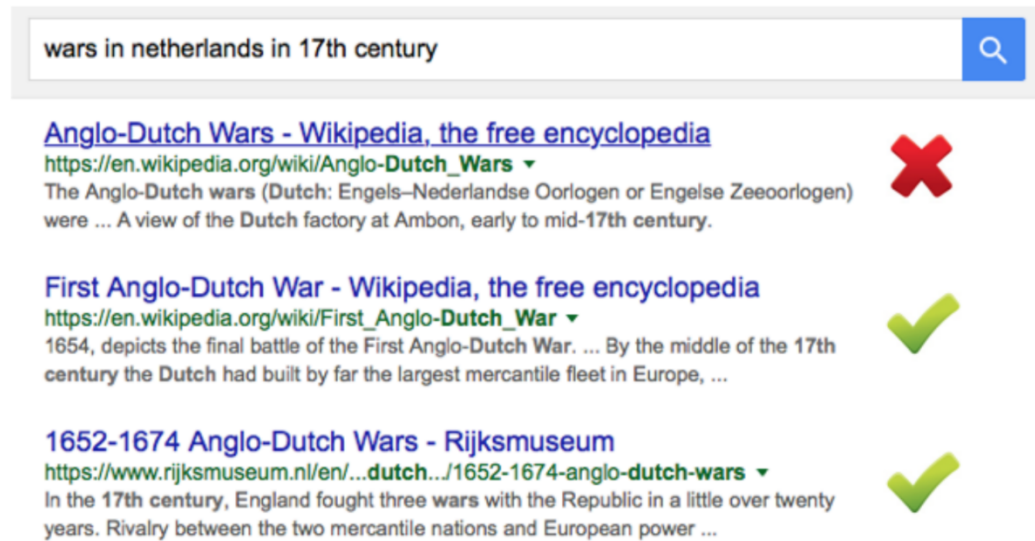


Figure 2.1: An example of search query and ranked list of documents where each document is marked as relevance (green) or non-relevant (red).

be severely limited. In this thesis, we focus on *understanding differences between two equally relevant documents on basis of effort* which has not been investigated in previous work. We also gather explicit judgments for effort which is in-line with existing work [24, 23, 12] on relevance characterization.

2.2 Information retrieval evaluation

The underlying objective of evaluation is to determine system effectiveness. A good system would rank documents such that user's information need is satisfied. Figure 7.1 shows an example of a search query and a ranked list of documents. The relevant documents are marked in green and non-relevant documents are marked in red. Before we review different types of evaluation metrics, we describe some popular metrics that are used throughout this thesis for evaluating retrieval performance.

Precision and recall

Precision (P) and Recall (R), defined in Equation 2.1 are popular measures used to evaluate ranked lists. While precision measures the fraction of retrieved documents that are relevant, recall measures the fraction of relevant documents that are

retrieved.

$$\text{Precision} = \frac{\text{\#relevant documents retrieved}}{\text{\#retrieved documents}} \quad \text{Recall} = \frac{\text{\#relevant documents retrieved}}{\text{\#relevant documents}} \quad (2.1)$$

Mean average precision

Most standard among the IR community is Mean Average Precision (MAP), which provides a measure of quality across recall levels. For a single information need, Average Precision (AP) is the average of the precision (P) obtained for the set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs or queries Q . Formally, if the set of relevant documents for a query $q_j \in Q$ is $\{d_1, \dots, d_{m_j}\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k , then

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk}) \quad (2.2)$$

DCG and NDCG

A final approach is the measures of cumulative gain [58], in particular discounted cumulative gain (DCG) and normalized discounted cumulative gain (NDCG). Like precision at k , it is evaluated over some number k of top search results. For a set of queries Q , let $R(j, d)$ be the relevance score assessors gave to document d for query j , DCG(k) and NDCG(Q, k) are defined as follows:

$$\text{DCG}(k) = \sum_{m=1}^k \frac{2^{R(m)} - 1}{\log_2(1 + m)} \quad \text{NDCG}(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \cdot \text{DCG}(k) \quad (2.3)$$

where Z_{kj} is a normalization factor calculated to make it so that a perfect ranking's NDCG at k for query j is 1.

At present, there are two ways of determining relevance of a document: a) explicit judgments and b) implicit judgments [59, 60]. These two paradigms of judgment gathering and system evaluation are briefly reviewed in the following subsections.

2.2.1 Offline evaluation

The primary way of obtaining relevance judgments is to get them assessed manually. In explicit evaluation, judges or experts are provided with concise definition of relevance and are asked to judge some documents in a corpus¹ with respect to a query and its description (if available). Each query-document pair is annotated independently where rating can be on binary or graded scale. An example of such evaluation is shown in Figure 2.2a. This set of queries, judged documents and remaining corpus forms a test collection that can be for repeated evaluation of different systems. This kind of evaluation paradigm is known as *Offline evaluation* or *Batch evaluation*.

Systems are evaluated using small test collections with limited set of information needs and static relevance judgments in platforms such as TREC², NTCIR³ or CLEF⁴. Batch evaluation is fast, repeatable, and relatively inexpensive (only initial cost of building test collection) and the data collected can be reused many times. However, test collection based evaluations make several simplistic assumptions about both real users and their information needs, what constitutes relevance, and many other aspects of retrieval (e.g. how summaries are presented to users, etc.). However, batch evaluation is commonly used in evaluating the quality of retrieval systems, especially when re-usability is a prime concern for enabling rapid experimental iteration among a number of alternatives.

Explicit judging limits the size and scope of test collection, as exhaustively evaluating each document with respect to a query is not only expensive but time consuming. Thus, while test collections with manual judgments are small and costly they make evaluation of multiple IR systems fast and repeatable. A comprehensive review of offline evaluation is given in [61, 62].

¹Corpus represents large collection of documents. For instance clueweb12 is a corpus of 733 million documents.

²<http://trec.nist.gov/>

³<http://research.nii.ac.jp/ntcir/index-en.html>

⁴<http://clef2018.clef-initiative.eu/>

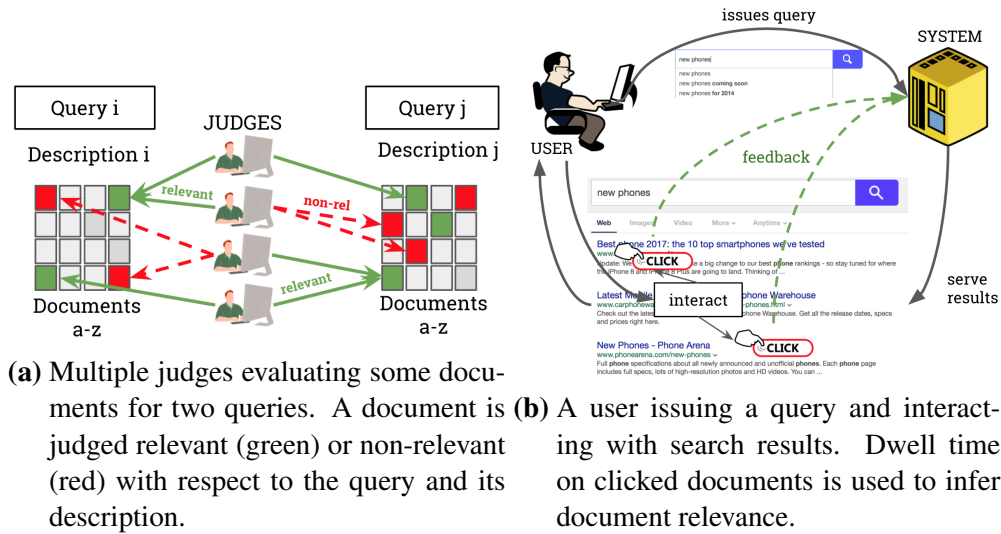


Figure 2.2: Two paradigms of evaluation

2.2.2 Online evaluation

Implicit judgments [26] are obtained by recording and analyzing user behavior on a document in the wild which yields larger test collections. Implicit judging also requires some heuristics (how many seconds should translate into what grade) and a lot of users to make statistically significant conclusions about page relevance.

The amount of time user spends on a clicked document, also known as its *dwell time* [25, 27], is used extensively in IR to determine webpage relevance. Implicit judgments can be useful in performing large scale evaluation of IR system with live users. This user-based approach of evaluation where actual users are observed and systems are evaluated on basis of these interactions is known as *Online Evaluation*. An example of how implicit judgments are collected from user interaction is shown in Figure 2.2b.

While user-based approaches can be used to evaluate end-to-end user satisfaction, they are expensive to perform since one has to collect data from enough users to evaluate and compare several systems. User-based data may also be difficult to analyze due to variance in tasks, populations and time. It has been found that click-through statistics are often highly affected by issues such as presentation bias [26] and perceived relevance of the documents. The perceived relevance of a document is mainly based on the summary (snippet) of the document and can be different

than the actual relevance of the document; hence, users may end up clicking on a document and find out that it is not relevant [63].

To overcome this problem, dwell time has been proposed as an implicit signal of relevance and dwell time is shown to be a good indicator of user satisfaction [60, 64, 65, 66]. There has been many different studies that compare dwell time with relevance. Kelly *et al.* [67] gives an overview of different research that has been done to analyse the correlations between dwell time and relevance. The Curious Browser [60] experiments showed that when users spend very little time on a page and go back to the results list, they are very likely to be dissatisfied by the results, and that a dwell time threshold of 20/30 seconds could be used to predict user satisfaction.

Morita and Shinoda [68] examined the relationship of three variables on reading time: the length of the document, the readability of the document and the number of news items waiting to be read in the user's news queue. Very low correlations (not significant) were found between the length of the article and reading time, the readability of an article and reading time and the size of the user's news queue and reading time. Based on these results, the authors examined several reading time thresholds for identifying interesting documents. When applied to their data set, they found that the most effective threshold was 20 seconds, resulting in 30% of interesting articles being identified at 70% precision.

Later, Buscher *et al.* [65] showed that using documents with dwell time less than 30 seconds as negative feedback resulted in better improvements in ranking performance than any other dwell time thresholds. They further showed that showing users only documents that have dwell time greater than 30 seconds have resulted in better ranker performance. Over many years, a dwell time value of 30 seconds has become the standard threshold used to predict user satisfaction [69, 70, 71].

In general, a very low dwell time can be reliably used to identify irrelevant documents. The converse of this is not necessarily true: a user may spend a long time searching for relevant information in a document and may fail to find the needed information. Hence, long dwell does not necessarily imply relevance [72]. Further-

more, the dwell time threshold that can be used to predict relevance is shown to vary depending on the task [71, 64]. Therefore, it is difficult to say that a document with dwell time above a certain threshold is relevant, whereas a document with a very low dwell time is likely nonrelevant [72].

2.2.3 Agreement between online/offline evaluation

In previous subsections we gave a brief overview about offline and online evaluation. One would expect that test (or batch) collection based evaluation would agree with user-based evaluation of systems. However it has been shown in the past [28, 29, 30, 31, 73] that these two forms of evaluation do not agree with each other, or agree with each other only when there is a significant gap in terms of the quality of the systems compared [30, 35].

These studies did not establish any direct correlation, stating that improvements in test collection based evaluation do not always translate into a direct benefit for end users (as measured by the number of relevant documents). However, it must be noted that the assessment procedures used in these studies did not take into consideration variation in personality and cognitive characteristics among users; instead, these studies used techniques which assumed that all users exhibited the same or similar characteristics.

Hersh *et al.* [28] observed that the explicit relevance labels reported by the participants during the study did not always correlate with the relevance judgments obtained from the assessors. This suggests that relevance judgments do not completely capture all aspects that might affect a user. Primary reason for this mismatch are the simplified assumptions around relevance, search topics and user behavior that are used to build test collections.

Disagreement between user ratings and system predictions is not limited to web search. It has also been found in tasks such as Query performance prediction [74] or Music retrieval [75]. Hauff *et al.* [74] found that correlation between the predictions derived from query performance prediction methods and the predictions obtained from human assessors was quite weak at both the topic and the query suggestions level. Similarly, Hu *et al.* [75] found weak correlation between system

effectiveness measure Average Precision (AP) (Section 2.2) and eight user centered measures which indicates that higher AP scores did not make the task of music similarity judgment easier

Recently, there is also some body of work [34, 76, 77] that found that user metrics such as precision and position biased metric DCG [58] (Equation 2.3) correlate well with user search effectiveness. For instance, Huffman *et al.* [77] found that search satisfaction (user reported) can be predicted reliably using the precision (manually judged) of first three search results. They also investigated prediction of satisfaction for different classes of queries such as informational, navigational or transactional. They found that rank of the key result was an important predictor for navigational queries, while for other kinds of queries, the information accumulated in later results was more important.

Kelly *et al.* [76] found statistically significant relationships between precision and user's evaluations of system performance. They observed that precision was a strong predictor of user ratings and explained more variance. Finally, the number of documents the users examined significantly influenced their evaluations, even when the difference was a single document.

Recent work [78] shows that existing online and offline metrics independently correlate with satisfaction but does not compare agreement between online and offline metrics. They also do not vary the quality of the systems, as done previously [79, 80, 76], to investigate the variance in correlation of both online and offline metrics with reported user satisfaction.

Scholer *et al.*, however, drew similar results as [35, 30], and found that the position biased gain metric DCG@1 [58] (Equation 2.3) on manual judgments agreed with user performance for systems with large gap in performance. This effect is statistically significant when relevance is treated as a binary criterion

The mismatch between online and offline evaluation could arise due to the disagreement between what judges and users consider relevant. One hypothesis is that judges are not trained to account for utility of a document with respect to an end user of an IR system. Trained judges are asked to identify document relevance

regardless of how much effort or time it may take to consume it. While a judge can take several seconds, even minutes to evaluate a document, an end user may not be willing to spend as much time consuming it, even if it is relevant. Thus, even if the document is relevant, it is of minimal utility to the user. We believe that the primary reason of mismatch between explicit and implicit document judgments is a result of effort needed to find and consume required information from a given document. While document relevance is important, so is the effort required to determine its utility. We begin with a brief overview of existing literature associated with effort, more specifically on how researchers define and characterize effort at present.

2.3 Effort in information retrieval

Effort is defined as ‘strenuous physical or mental exertion’ or ‘a force exerted by a machine or in a process’ in english dictionary. Similarly, ‘*effort heuristic*’ [81] is a mental rule of thumb in which the quality or worth of an object is determined from the perceived amount of effort that went into producing that object. Overall, it suggests investing energy in some action for a certain outcome. In this thesis our focus is to investigate the properties that constitute ‘*effort*’ and their integration in retrieval.

There are different forms of activity in web search that can be measured as some sort of effort. We can compute system level or user level effort in information retrieval. A user may invest effort at three different levels while addressing an information need. The first is document specific effort where the user investigates *a single document* with respect to the search query. The user must evaluate different nuggets of information *within* the document to make a decision about its utility. This thesis revolves around estimating what factors constitute *document level effort* and how they can be incorporated in design of retrieval models.

The second is the effort required to find and consume relevant information from *multiple documents* in a list of search results. Where a single document contains insufficient information to answer a search query, a user may have to visit multiple documents to address the underlying information need. Finally, if the information need is underspecified or has several components, one query may not be sufficient

to find a satisfactory answer. Hence, a user may have to issue multiple queries or examine multiple documents to find relevant information. This is known as *session search* where a complex information need is satisfied by issuing multiple queries and examining several documents. In this thesis, we investigate the correlation of session-level effort with explicit labels of user satisfaction in Chapter 7.

2.3.1 Modeling effort in information seeking tasks

Prior research has investigated the nature, importance and influence of effort on information seeking behavior. Not surprisingly, there is significant work on system/user effort definition and evaluation in IR research. We begin with an overview of models from information foraging theory [82, 83, 41] that measure user's effort in terms of some cost incurred in taking any action while answering a search query. The underlying hypothesis of these models is that a user would adapt a strategy that maximizes their gain of information per unit cost in addressing an information need.

Information foraging theory [84, 41] studies the dynamics of user choices between examining one document known as *within-patch exploitation* and exploring other documents i.e searching for next document that is also known as between-patch exploration. One of the assumptions of such models is the law of diminishing returns or Charnov's Marginal Value Theorem [85], i.e. the information gain of the users decreases as they spend more time consuming information related to a search topic.

Pirola *et al.* [84] model information seeking tasks with the help of optimal foraging theory. Optimal foraging theory outlines how animals such as predatory birds hunt their prey while minimizing the access or navigation costs. A optimal forager, for example a bird of prey hunts such that it gains maximum energy per effort invested in finding, pursuing and hunting its prey within the constraints of its environment. Similarly, an *optimal information forager* is one who best solves the problem of maximizing the rate of valuable information gained per unit cost, given the constraints of the search environment. These constraints include the validity of different sources and the costs of finding, accessing and consuming them. Their experiments with two information seeking tasks showed that participants used

an interleaving set of activities devoted to foraging, sense making, and knowledge construction and they relied on schematic representations to judge the utility or relevance of information sources. Each document is considered to be an information patch that forager consumes to satisfy an information need. While searching, the forager is expected to make a series of decisions between either consuming one patch or select a different document amongst the list of documents (or patches) in a search result list.

They proposed the ACT-IF cognitive model of foraging using production rules. They used word frequency and co-occurrence statistics in documents to estimate judgments of information. The model simulations were compared to a dataset collected from real user's search interactions in a browser. Each user was required to select and examine a cluster of documents with respect to a search task. The ACT-IF model yielded a good fit to users' ratings of the number of relevant documents in a given cluster. It could also explain the differences in the number of clusters selected for queries of different difficulties by the users.

While ACT-IF lays the ground work for approximating user costs and benefit, it is restricted to cost or gain obtained only from document assessment. It does not account for several other actions that users take while satisfying an information need in practise. For instance, it does not account for snippet examination which is very common in both desktop and mobile search [86, 87]. It also cannot model situations where there is a significant overlap between documents in search results [88].

SNIF-ACT 1.0 and 2.0 [83] uses information foraging theory to model how people exploit information cues, specifically text associated with hyperlinks in the webpage, to make navigation decisions such as judging which link to click next or when to abandon the search process. SNIF-ACT 1.0 and SNIF-ACT 2.0 showed that the measure of *information scent*, a word co-occurrence based measure of link utility, provides a good description of how people evaluate relevance of link texts and their search goals. On comparison with a simple Position model that scores links based on their positions on the web page, they found that both versions of

SNIF-ACT provide much better fit to human data than the Position model, showing that information scent reliably predicts user interaction with web documents.

SNIF-ACT 1.0 and 2.0 assume that users sequentially scan results which is not always true. It has been shown [89, 90, 86, 91] that users scan pages non-linearly and switch between scanning and reading a document. In such cases, static information scent derived solely on the basis of word co-occurrence would not be useful in determining which link the user shall follow next. SNIF-ACT would not reliably model navigation sequences for complex webpages since they do not account for different visual layouts. Another limitation of SNIF-ACT models is that they completely ignore the effect of user's prior knowledge about the search topic on their navigation. The models provide a rough estimation of navigation which may not apply to different users in real-time.

While models based on information foraging theory *et. al.* [84, 83] encode user's search cost and gain, their focus is to simulate or predict documents users may read and when the users should switch between documents. Information scent acts as a proxy for document relevance and time as a measurement of effort. We also explore simple cost-benefit models in context of mobile search in Chapter 7 and compare the correlation between explicit user satisfaction labels and net profit obtained from the models. However, our focus is to evaluate query and session based cost-benefit models. Existing foraging models only model *sequential ranked list examination for queries* but do not address aspects such as cost of querying, issuing multiple queries and examining search snippets, actions that are prevalent in session based searches. It is not clear how information needs that require users to issue multiple queries, non-linearly examine several snippets and click on one or more documents can be incorporated into existing information foraging models. Existing foraging models also do not shed any light on optimal number of queries that need to be issued in addressing an information need, which may be device dependent, something that we shall investigate in Chapter 7.

Another limitation of these models is that information scent calculation is limited to hyperlinks and word co-occurrence. Given that webpages today are fairly

complex, we need a better methodology to define information scent. In Chapter 3, 4 and 6 we shall use features such as document length, readability and frequency of tables, images or lists to capture this complexity. ACT-IF or SNIF-ACT models can be used to predict or simulate which results in the rank list will be clicked by a user and when they would abandon the website. However, these models would not differentiate between two documents that have the same information scent (or relevance). But in this thesis our objective is to determine what characteristics of effort can be used to differentiate between two equally relevant documents. Finally, it is not clear how information foraging theory models could be used to design and implement learning-to-rank models in practice that can learn from large collection of documents labeled for relevance, something we shall address in Chapter 6.

2.3.2 Characteristics of effort

Existing literature defines and studies effort differently during various stages of information searching and gathering. For instance, previous study [92] defines effort as number of documents clicked/viewed during a search session. The work presented in this thesis is different in that we study what factors best correlate with effort while consuming a *single document*.

Some work has also studied user effort in searching or judging a document. Villa *et al.* [93] 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 thesis, 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 is different in that it identifies effort required to consume a *relevant document*.

Sormunen in [94] studied factors that affected assessor effort in judging documents. The authors reported that the consistency of assessments is difficult to achieve if the assessors feel topic descriptions are ambiguous. They also found that a result page without any spam documents was preferred to one with spam;

and an irrelevant document high in a result list negatively impacts user satisfaction. Smucker *et al.* [95] measured assessor effort to identify errors in judgments.

Although Jones *et al.* [96] do not study effort directly, their finding is that given two documents of equal utility, users prefer the one with the lower spam score. Judging effort has also been studied for images, in [97] authors found that judging accuracy was not significantly affected by image size. However, it was found that size of an image significantly impacted the time required to judge it, with larger images taking longer to judge. Chandar *et al.* [98] have also shown that readability affects assessor disagreement.

Effort spent in scanning/reading search result lists has also attracted attention. Eye tracking [90, 86, 91] studies have shown high user-level variation in scanning strategies. Thomas *et al.* [90] found that for easy tasks, users limited scanning top few results but for complex tasks, they scanned deeper in the list. Lorigo *et al.* [86] examined the number of fixations, fixation duration, and time spent on tasks for two search engines. They also found that task type was shown to influence SERP viewing time and the number of fixations on selected web documents. In informational tasks, users spent less time on SERPs and had greater pupil dilation as compared to navigational tasks.

Guan and Cutrell [91] manipulated the positions of target results in navigational and informational tasks. Overall, participants devoted more time to tasks and were less successful in finding highly relevant results when they were displayed at lower positions on the search results list. This effect was especially pronounced for informational tasks (as opposed to navigational tasks). The eye tracking data showed that there was a decreased probability of looking at results at lower positions, explaining the poor performance of unsuccessful searches. Brumby *et al.* [99] performed similar eye tracking experiments in context of menu item selection in presence of several items where one was a relevant target item and all other ‘distractor’ items were of different relevance grades. They found that people rarely visited all items in the menu before selecting one option. They also visited fewer items in the menu when the distractor items were less relevant (poor) to the task.

Whereas, when the distractor options were more relevant to the task, they visited more items in the menu. Their findings suggest that menu item selection is sensitive to the context provided by expected relevance of all the examined items and not just to the most recent item.

Readability [100, 101, 102, 103] of the document may influence both the time and effort a user has to spend finding required query-specific information from a webpage. Collins *et al.* [104, 105] have shown that the webpage readability levels impact users understanding of the content. Maskari *et al.* [106] investigated relation of *Findability* with relevance. Their study found that users employ several cognitive processes while retrieving information, including learning, comprehension and speed in spotting information which contributed to users effectiveness of the search process. Reader *et al.* [151] conducted a controlled study in which the participants were asked to read four relevant documents of different difficulty grades. We also use a similar experimental setup in Chapter 4 where effort judgments are gathered for two *equally* relevant documents. Their objective was to examine how users allocate their time across texts and they found that users indeed adaptively use different strategies to read text. Inspired by information foraging theory [83, 41], they outlined two strategies of examining texts. The first strategy, sampling, involves inspection of all the documents and then reading selected nuggets. The second strategy is satisficing where texts are evaluated simultaneously as they are read by the users. Overall, readability has been found to be an important indicator of cognitive effort, which we shall also investigate in subsequent chapters but in a different capacity.

It has also been shown that users actively *find* relevant or interesting information on a page [107, 108, 68] and may not sequentially read entire web pages. Guo *et al.* [72] 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 [109, 41] 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 between two *equally relevant* documents. We aim to obtain explicit judgments for all these parameters to identify which factor is highly associated with judging effort for a relevant document.

To the best of our knowledge, there is no work on identifying effort related factors important to clicked relevant pages. Existing work only studies examination of search result pages, not how much effort was spent once the result was clicked. While, we do not conduct an eye tracking study, we investigate what captures effort for a *relevant document* in subsequent chapters.

2.3.3 Time based evaluation of effort

Smucker *et al.* [16] propose Time biased gain (TBG) to evaluate search effectiveness on the basis of time spent and information gained as user scans ranked document list. This is quite different from our work, as our focus is to determine effort to consume a document independent of the collection. But we use TBG for evaluation and show that our approach performs better than existing methods.

In work motivated by XML and multimedia retrieval systems, de Vries *et al.* [20] present a user model based around a '*tolerance to irrelevance*'. Their evaluation model assumes that users start reading some passage in a document and continue until either satisfied with relevant information and/or that relevant information '*is starting to appear*', or reach their time-based (or user effort-based) irrelevance threshold. Upon reaching an irrelevance threshold, the users proceed to the next system result. This work shares similarities with ours in terms of an abstract user model, but was motivated more by addressing the issues of not having a predefined retrieval unit within video and XML retrieval test collections.

Their measurement of effort is the time spent on inspecting irrelevant information which is different from what is used in this thesis. We posit that effort is a complex entity which may encapsulate several factors and this thesis attempts to characterize and evaluate such factors. Our focus is not a ranked list of results but a

single document.

2.3.4 System based evaluation of effort

Perhaps the first metric of effort was proposed by Cooper [110], where the proposed metric *Expected Search Length (ESL)* computes the expected user effort as the average number of documents the user has to browse in order to retrieve a given number of relevant documents. ESL is followed by several other metrics such as expected search duration [18]. Similarly, other works [111, 17] also propose measures of effort over ranked list of documents. The primary aim of these studies is to formulate and evaluate effort for user interaction with the system.

Carterette *et al.* [111] study the variation of utility based models within which they also investigate effort a user must put forth to achieve a particular amount of utility. They study characteristics of various measures incorporating these models.

Effort based evaluation of retrieval systems has also been proposed by Nicola *et al.* [17]. The proposed Twist measure evaluates the effectiveness of a system with respect to the effort required to *retrieve* desired information. They compare and contrast gain vs effort it requires to scan search rank lists. These definitions of effort differs from ours in that we aim to explicitly capture effort per *relevant document* independent of its position in the search results. The above metrics are concerned with user effort or tolerance to relevant or irrelevant documents in the search results. This body of prior work focuses on effort based evaluation of retrieval systems. However, this thesis attempts to characterize and evaluate effort per *relevant document*.

Jiang *et al.* [112] investigated how existing metrics can be adapted to account for effort. Similarly, Zuccon *et al.* [113] incorporate readability⁵ into evaluation. However, in this thesis, we focus on how to characterize effort required to consume relevant documents, collect judgments for the same and incorporate relevance and effort in learning-to-rank models.

There is some work on evaluating reading effort in XML documents. The authors [19] investigate the effectiveness of two metrics: a) effort required to assess

⁵they use readability measures such as LIX[114] to denote ‘understandability’

the document's relevance and b) character level effort required to read the relevant passages of the document. The proposed system presents a user with a ranked list of passages where passages are ranked *in decreasing order of relevance* to the user's search query.

Their study assumes that each document is a list of passages with decreasing order of relevance, which in reality may not be true and may, in-fact, hinder user's comprehension of the document. Their experiment is suitable for tasks such as XML retrieval where each document consists of independent but coherent chunks of information that when ordered arbitrarily do not hinder user's document comprehension. While, their work forms a good motivation, our objective is not to score passages with respect to user's query but to holistically evaluate a relevant document for user's effort with respect to her search query that is applicable in the real world.

Existing work does not quantify or analyze effort per document since it would be useful to know how much effort user must put forth to consume it. Document level effort information can be used to optimize retrieval. It can also be used to discount perhaps relevant but high effort documents to improve user satisfaction. Effort information can also be combined with other information (such as topic expertise, language proficiency) about the user to personalize search results. This will automatically reduce the time user spends on each document, in turn improving effort or utility based metrics proposed in existing work.

To summarize, while some related work investigates what factors impact assessment of relevance, our focus is not to measure relevance but to differentiate between two relevant documents on the basis of effort. Our work also differs from systems that measure effort required to find documents in a collection or effort required to scan search result lists.

Since we incorporate effort into ranking in this thesis, we also give a brief overview of learning-to-rank literature. Learning-to-rank models are traditionally designed to optimize for relevance. However, some work exists on incorporating factors besides relevance into training and evaluation of rankers. For instance, re-

searchers have explored how to balance relevance and freshness [115] of results. Researchers have proposed [115, 116] modification to existing rankers that account for both freshness and relevance. Our work is similar to this work, as we also propose modifications to pairwise approaches to account for effort and relevance. However, freshness and effort are different parameters, where effort may conflict with relevance (high relevance, high effort vs. low relevance, low effort), which does not apply in case of properties such as freshness.

Some researchers have also proposed [117] solutions for vertical ranking where objective is to balance multiple aspects of relevance. There is also work [118] on training rankers with multiple objectives. Essentially, they propose modifications to LambdaMart to first train for primary metric and then optimize wherever possible for secondary metric. Our work is similar to this work as we also propose extensions to existing learning-to-rank approaches to first optimize for relevance then followed by effort.

2.4 Mobile search

There is some work on mining large scale user search behavior in the wild. Several studies have looked into *when* does user search for information on mobile. Existing research [119, 120] has shown that today mobiles are used extensively to satisfy information needs. For instance, Church *et al.* [121, 122, 123] characterize *what* kind of mobile information needs arise and *how* users find information. They adopted method of diary entries to gather information about users search tasks. They observed that most mobile information needs arise due to social context (conversation with friends) or repeated daily tasks (finding directions home).

Researchers have found that user search logs on mobiles and desktop have many differences in terms of query length, click patterns, and search time [124, 125, 120]. Kamvar *et al.* [126, 127, 128] analyze large scale query logs to distinguish between queries issued from mobile. They also extensively analyze category of these queries and topics of clicked documents. On comparing searches across desktops and mobiles, they concluded that users treat smart phones as extensions of

their desktop computers. They suggested that mobiles would benefit from integration with computer based search interface. These studies found mobile queries to be short (2.3 - 2.5 terms) and high rate of query reformulation. Small scale studies like [129, 120] also report differences in search patterns across devices.

One key result of Church *et al.* [124] was that 90% search results did not get any clicks from users on mobile which they attributed to unsatisfactory search results. They suggested that different parameters such as user's location or time of the day should be taken into account while serving results on mobile. Song *et al.* [125] 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. They also proposed a framework to transfer knowledge from desktop search such that search relevance on mobile and tablet can be improved.

Recent work on mobile search has also explored use of user's touch interaction to improve retrieval for cross-device information needs [130]. Mouse movements have been shown to be effective in pre-fetching [131] and re-ranking [132] search results on desktops. For example, Huang *et al.* [133] found that mouse cursor activities on SERPs align with user's eye movements and those activities can be further used to infer document relevance.

More recently, Guo *et al.* [134] compare user behavior from two different laboratory studies on mobile and desktop designed for seven search tasks. They report higher dwell time⁶ on mobile (~44.3sec) than desktop (~31.2sec). They use several user interaction specific features to determine page relevance. They found that the user's inactive time (on clicked document) is positively correlated with its relevance, whereas gesture speed has a negative correlation with document relevance.

The underlying assumption of their work is that user interaction on a webpage can be used to predict its relevance. In their study, they assume that when users struggle to find the information in a webpage, they will mark it as non-relevant.

⁶Nicholas *et al* [135] also reported slightly higher dwell time on mobile vs. desktop.

However, topical relevance is *independent* [12] of user interaction. A user may struggle to find required information for several reasons such as small screen size, low readability or inconvenient location of the answer in the webpage.

We believe that such features in [134] capture *user effort* and do not reflect on topical relevance of the webpage. Topical relevance objectively evaluates only the *presence* of information required to answer the query in a webpage and is not associated with user's action of finding it. However, their work sheds light on the merit of such features which in-turn reinforces the importance of considering *effort* in retrieval.

Other work includes abandonment prediction [136] on mobile and desktop. Recently, Williams *et al.* [137] investigated role of good abandonment in context of mobile search. They proposed and analyzed role of gesture based features to predict good abandonment. Researchers have also investigated query reformulation on mobile [138] and understanding mobile search intents [124]. Buchanan *et al.* [139] proposed some ground rules to design web interfaces for mobile.

Existing work does not compare mobile or desktop on basis of cost and benefit framework which may be useful in determining user success in search. While existing work models search success via manually designed features, in this thesis we take an alternate approach, in that, we compare user success (reported by a real user) with user effort and gain using models based on econometrics. To the best of our knowledge, existing cost-benefit models have only been empirically evaluated on desktops but not on mobiles. Given that both modalities are different, we believe that model parameters derived from desktop based search studies cannot be directly used on mobile. In this thesis, we perform a user study to evaluate cost-benefit models on mobile and show that indeed desktop models would need to be modified to incorporate mobile specific user behavior.

2.5 Conclusion

In this chapter we gave a brief overview about relevance, its definition and role in Information retrieval. We also discussed parameters that influence users while determining the relevance of a document with respect to their information need. Re-

searchers have proposed several criteria, such as novelty, authority or topicality, that affect user's decision about document relevance. However, existing studies neither study dependencies between these parameters nor investigate how their influence on document relevance evolves with user's information need. It is also extremely difficult to get large scale judgments for these parameters to evaluate their importance for relevance at scale. While these factors are important and impact user's notion of relevance, it is difficult to use them for information retrieval effectiveness evaluation.

We also discussed types of IR effectiveness evaluation. It has been repeatedly shown that batch evaluation does not agree with user based evaluation of IR systems. We posit that *effort* is a source of this disagreement and that current IR judgments reflect document relevance but do not encapsulate effort required to read, locate and understand relevant text in a document. We empirically evaluate this hypothesis in Chapter 3.

In Chapter 4 we study potential parameters, such as the ability to read, locate and understand relevant document text, that encapsulate user effort and identify the most important parameter. We also train a model to provide document level judgments for this parameter (thus enabling large scale evaluation and comparison of IR systems). We also evaluate the effect of different parameters on these judgments in Chapter 5 and finally show how effort judgments can be incorporated in ranking in Chapter 6 .

We also reviewed existing work in characterizing differences between mobile and desktop search. At present, there is no work that explains how existing cost-benefit models will differ across devices. Therefore, we study the utility of existing models of evaluating interactive information retrieval on mobile in Chapter 7. We attempt to understand how existing cost-benefit models capture user effort within a session and whether existing desktop based model parameters can be directly used for mobile search.

Chapter 3

Relevance vs. document utility

As noted in the previous chapters, information retrieval effectiveness evaluation typically takes one of the two forms: batch evaluation based on static test collections and online experiments that use implicit signals from users (such as time spent on the page) as indicators of relevance. Test collections consist of a small number of information needs and static relevance judgments obtained manually either from the experts [140] or from crowdsourcing [141] experiments. On the other hand, user-based evaluation involves observing a real user and inferring the relevance of the webpage by either using the time spent on the page i.e dwell time or by mining other actions [72] of the user.

Ideally, the outcome of batch evaluation should be predictive of the satisfaction of real users of a search system. Yet research has shown that these two forms of evaluation often do not completely agree with each other [28, 29, 31, 33, 73], or agree with each other only when there is a significant gap in terms of the quality of the systems compared [30, 34, 35]. One of the main reasons behind this mismatch is the simplifying assumptions made in batch evaluation about relevance and how users behave when they use a search system. One such assumption is that the users independently evaluate each document with respect to a query. Relevance assessors are not the owners of the query or the search task but are assigned query-document pairs for evaluation. They are only shown the query-document pair but are not provided any context about the user's topical knowledge, search history or goals at the time of judgment. This lack of context can result in an incomplete or

incorrect judgment [142] about document relevance. Another assumption is that the annotators would agree on the scope of the information need. Some annotators may interpret query terms differently from other annotators which may yield different assessments. Therefore, there is an increasing interest in modeling the user needs and interaction with a search engine in collection-based effectiveness evaluations [16, 79, 80, 143].

We claim that a key source for disagreement between batch evaluation and user-based online experiments is due to the disagreements between what judges consider as relevant versus the *utility* of a document to an actual user. The main goal of this chapter is to investigate the following hypothesis.

Hypothesis 1: Existing relevance judgment paradigm does not account for effort, which, in turn causes the mismatch between explicit and implicit evaluation of systems.

To address this goal, we rely on the implicit feedback (such as dwell time information) gathered from real users of a retrieval system. We compare the indicators of *utility* inferred from implicit feedback to judgments obtained from relevance judges and identify sources of disagreement.

We further focus on the reasons of mismatch between relevance judgments and implicit signals obtained from the clicked documents (e.g. dwell time). Our hypothesis is that existing relevance judgments do not account for ‘*effort*’ needed to find and consume relevant information, which in turn causes the mismatch between online and offline evaluation. There is a difference between what judges *think is relevant* and what users *find relevant* in real time, i.e. the estimation of ‘*effort*’ required to extract utility from a relevant webpage is not the same for both populations.

Relevance judges are typically explicitly asked to identify the relevance of documents they assess. Therefore, they must evaluate each document thoroughly before marking it as relevant or non-relevant. In performing these judgments, judges often spend a significant amount of effort on documents that may not have significant relevant content or that may be hard to read. On the other hand, users usually simply wish to fulfill an information need and are often much less patient when determining

if a particular document is relevant. If they do not see evidence of sufficient relevance quickly or if they think that relevant information is difficult to consume, they tend to give up and move on to another document. Therefore, even if a document is relevant to a query, it provides only minimal utility to an actual user if finding and understanding the relevant portions of the document is difficult.

Overall, our findings suggest that *effort* plays an important role in user satisfaction. Our results also show that features related to the effort to find and consume relevant information (for example, readability level of the document) could be used as ranking features when retrieval systems are designed as they have a significant impact on the utility of a document to an actual user.

We begin with a user model that considers different stages of user interaction with a search engine and compare this with the behavior of judges in Section 3.1. Through the user model, we show how the *utility* of a document with respect to an actual user could be different from the relevance of the document in Section 3.3.1. We also show through a regression model that features related to *effort* can predict if a *relevant* document is of low utility to users in Section 3.3.2. Finally, in Section 3.4 we conclude with findings and limitations of this work.

3.1 User behavior and relevance

Given a set of search queries and clicked documents, the objective of this chapter is to understand the difference between the relevance judgments obtained from the judges and the utility of the documents to a real user, the first step is to consider how users assess clicked documents. We discuss stages, derived from previous eye tracking studies [65, 144, 145, 146], a user may traverse to examine a clicked document with respect to a query.

Existing research [65, 144, 146] that evaluates information seeking behavior on a webpage with eye tracking studies has shown that users *switch between scanning or skimming parts* of the document which is followed by concentrated reading of relevant sections when searching for relevant information. For instance, Buscher *et al.* [144] found that while scanning users mostly looked at the beginning of the lines

and quickly scanned downwards. In contrast, during reading behavior, they looked over the full length of each line indicating concentrated reading. We incorporate the two stages of information consumption with the following user model.

3.1.1 Document evaluation model

The two-stage model of how users examine a clicked webpage is as follows:

- **Stage 1: *Initial Assessment***

Upon clicking on a document (with an expectation of finding relevant content), users make a rapid adjustment to their expectation. A user may quickly scan the webpage to locate information nuggets, i.e. sentences, tables or paragraphs, that could potentially address the underlying information need. A user may examine these information nuggets in depth in the next stage.

- **Stage 2: *Extract Utility***

Assuming the user expects that they can extract value by identifying an answer to their question or information need, the user is now willing to commit time to read the content, view multimedia, or complete a transaction.

At the time of searching, a user needs to go through both stages to extract value from the content of a clicked document. However, if a document does not seem promising during the initial stage-one assessment, i.e. scanning, the user may choose to go back to the search result page, issue another query or abandon the search task altogether. Examining another document may be a particularly good strategy if the user believes that other documents exist that may be useful to consume. This is similar to the principle encoded in the information foraging theory [84].

At the time of judging, a query-document pair is assigned to an assessor who evaluates the document's relevance for the search query. Usually a judge is provided with a search query, description of the information need if available and a snapshot of the webpage for evaluation. An assessor's goal is only to identify whether the given document is relevant. As accuracy in judgments is important, judges are willing to invest more time to ensure that their answer to the document relevance

assessment is correct. Judges also sometimes spend substantial time deciding the degree of relevance, for instance considering guidelines to determine if a document should be marked as ‘*relevant*’ or ‘*highly relevant*’.

We observe that for documents where judges take a long time to make a relevance assessment, the assessment must itself be difficult. Therefore, in subsequent sections, we take long judging time to indicate a *high-effort* document. We validate this hypothesis in Chapter 5 by collecting explicit judgments for effort and find that judges consistently spend more time in judging *difficult* or *high effort* documents. In contrast, we hypothesize that when users spend a long time on a document, they are either spending time consuming the content (Stage 2) or are willing to spend a long time on Stage 1.

Our objective is to study the potential mismatch between what is considered relevant by the users and what is identified as relevant by the judges. Given a query-document pair we now compare the dwell time, i.e. the time spent on a clicked document by an *end user* with its judging time, i.e. the time an assessor (expert or crowd-worker) spends on evaluating the *same* document with respect to the query. This comparison would help to identify and closely examine documents where implicit relevance determined from the user dwell time does not align with the judge’s explicit evaluation of document relevance for a search query. To summarize, when the dwell time (time spent on a document by actual users) and the judgment time spent on a document are considered with respect to the above model, one of the following four cases must hold:

- 1. Low dwell time, low judgment time :** One possibility in this case is that the document is obviously non-relevant, and both users and judges reach this assessment quickly. Alternatively, the document may be highly relevant for the information need and users require very little time to extract relevant information from the document. For instance, a question-answering information need may require users to simply read a single sentence or a subset of words or numbers to extract utility.
- 2. High dwell time, low judgment time :** This scenario would occur when a document is clearly relevant, and real users spend substantial time on the second stage

extracting utility from the document.

3. Low dwell time, high judgment time : As judges take a long time, it is unlikely that the document is obviously relevant or obviously non-relevant. However, users abandon the document quickly. In terms of time, this bucket consists of documents on which users and judges do not agree which is of interest to our study in subsequent experiments.

4. High dwell time, high judgment time : A document that perhaps requires some in-depth consideration from both judges and users. The document could be relevant or non-relevant. If the document is non-relevant, it would require high effort from both judges and users before they reach a decision about its relevance. On the other hand, if the document is relevant, both users and judges perhaps engage with its content and consume it in-depth to extract utility.

Next, we describe how we use this model to infer document *utility* with respect to actual users, showing that of the four cases most mismatches between relevance and utility tend to occur on documents under case 3, documents that the users do not spend much time on but are labeled as relevant. We further show that these mismatches are mainly caused by effort to find relevant information.

3.2 Experimental setup

We use three datasets to compare relevance judgments obtained from judges with document utility for actual users. Each dataset is parameterized by three aspects: (a) the source of queries that are judged, (b) the types of judges performing the judgments, and (c) the way in which dwell time data was collected.

3.2.1 Dataset collection

The first two datasets CrowdJ-TrecQ and ExpertJ-TrecQ are derived from Kazai *et al.* [147], which was used to analyze systematic judging errors in IR. This data consists of queries from TREC Web Track Ad Hoc task in 2009 and 2010. It was constructed by scraping the top 10 search results from Google and Bing for the 100 queries from Web Track 2009 and 2010, resulting in a total of 1603 unique query-URL pairs over 100 topics. This method of re-sampling documents for the TREC

Queries from ExpertJ-TrecQ and CrowdJ-TrecQ	
Query	URL
science fair project ideas	https://www.sciencebuddies.org/science-fair-projects/project-ideas
multiple sclerosis	http://www.webmd.com/multiple-sclerosis/default.htm
carmax san antonio	http://www.carmax.com/enus/locations/texasusedsanantonio7152.html
dow jones industrial average	http://topics.bloomberg.com/dow-jones-industrial-average/
ways to make extra money	http://christianpf.com/ways-to-earn-extra-money-from-home/
Queries from CrowdJ-NaturalQ	
Query	URL
irs free tax preparation online	http://www.freefile.irs.gov/
american cocker spaniel	http://en.wikipedia.org/wiki/American_Cocker_Spaniel_Puppies
symbolism of water	http://www.whats-your-sign.com/symbolism-of-water.html
do you get paid for jury duty	http://www.jud.ct.gov/jury/faq.htm
melissa rycroft biography	http://www.biography.com/people/melissa-rycroft-20980961

Table 3.1: Sample query-document pairs from the datasets

topics was preferred in order to ensure up to date coverage of the topics and high overlap with the query-URL pairs that appear in the logs of the commercial search engine, which we aim to use in our analysis.

Our third dataset (CrowdJ-NaturalQ) consists of queries sampled from the actual traffic of a commercial search engine. We mined the anonymized query logs from the commercial search engine for a seven-week period starting in late 2012 and extracted queries and clicked results for further analysis. To reduce variability from cultural and linguistic variations in the search behavior, we only consider entries from searchers in the English-speaking United States locale. We sample pairs of URLs shown in the top two positions of the organic Web results.

We further restrict CrowdJ-NaturalQ to those pairs of URLs for which each URL has at least 30 *impressions*, i.e. is shown at least 30 times in the search result page for the query. We use URL pairs whose dwell time distribution is significantly different from each other with $p\text{-val} < 0.05$ computed using two-tailed t-test. This produces a set of about 5,000 query-URL1-URL2 triples. As an example, five queries and documents from each dataset are shown in Table 3.1.

3.2.2 Labeling methodology

While 1603 query-URL pairs in ExpertJ-TrecQ were judged by highly trained judges (experts) that are employed by a commercial search engine, for CrowdJ-TrecQ , they were judged by crowdworkers recruited via Crowdflower¹. The judg-

¹<http://www.crowdflower.com>

ment interface was setup such that the collected judgments are comparable to those provided in TREC Web Track where given a query-document pair, a judge is asked to annotate the document’s relevance with respect to the query. The judging interface showed the judges a query and a web search result (in an iframe) and asked them to rate the search result’s topical relevance to the query using a five-point Likert scale from *Bad* to *Ideal*. Following description was used for individual grades:

- **Ideal (4):** User seeing this result would think it is a key result.
- **Highly Relevant (3):** User seeing this result would find it highly relevant.
- **Relevant (2):**User seeing this result would find it relevant and will be happy with the result.
- **Somewhat Relevant (1):** User would find it somewhat relevant since it contains only partially relevant or incomplete information.
- **Bad (0):** User seeing this result would find it non-relevant and will be unhappy as this result provides no useful information for the query.

Each query-URL pair was annotated by 3 judges and the majority vote was used to identify the final label for a document. Judges could skip a query-document pair (by selecting the ‘I cant tell’ option). They could also research a query by viewing the top 10 results from Google² and Bing³. The query-url pairs were judged by 20 professional judges to construct ExpertJ-TrecQ dataset. CrowdJ-TrecQ dataset consists of the same query-url pairs but judged by 45 US crowd workers recruited via crowdsourcing platform Crowdfunder. To participate, workers are first qualified via a basic Web judging task, but receive no training.

We use the same judging interface to gather labels for CrowdJ-NaturalQ , where each document is labeled on the same grades listed above by five judges on Crowdfunder. Each judge in CrowdJ-TrecQ and CrowdJ-NaturalQ is payed 0.03 cents per HIT which totals the cost of judgment to \$144 and \$750, respectively.

The properties of all three datasets are summarized in Table 3.2. Overall, we

²<http://www.google.com>

³<http://www.bing.com>

Dataset	Queries	Judges	Clicks collected on
CrowdJ-TrecQ	Manually constructed for TREC	Crowdsourced	Natural search rankings
ExpertJ-TrecQ	Manually constructed for TREC	Trained experts	Natural search rankings
CrowdJ-NaturalQ	Sampled from actual query distribution	Crowdsourced	Randomized rankings

Table 3.2: Datasets used for analysis

observed that the agreement rate between professional Web judges was higher than the crowd workers where Fleiss’ Kappa (κ) was 0.59 for ExpertJ-TrecQ and 0.33 for CrowdJ-TrecQ . We obtained a similar agreement rate using Krippendorff’s alpha (α) of 0.35 for CrowdJ-NaturalQ dataset.

3.2.3 Time measurement

Since our goal is to study the utility of a document with respect to an actual user versus a relevance assessor, we simplify the analysis by only considering relevance on a binary scale, converting the graded relevance judgments into binary. We follow the same approach adopted previously [148, 149] to convert graded judgments to a binary scale. In our analysis, all documents labeled as bad are considered to be *non-relevant*, and all others are labeled as *relevant* to the query. We shall now elaborate how we determined dwell time and judging time for each document.

Dwell Time Measurement: To get the dwell time information for the documents in our datasets, we use click logs of a commercial search engine over a 3 month period starting in September 2013. The dwell time information was collected by observing all clicks on the search engine results during this period. Note that dwell time information was not available for all the documents judged in the datasets as users tend to click only on documents that they assume will be relevant based on the document snippet. To make sure that we have a reliable estimate for dwell time, we focus on the documents that have been clicked at least 30 times during the 3 month period. When the documents for which we have dwell time information available are considered, we end up with 4399 documents for the CrowdJ-NaturalQ dataset, and 1538 documents for the ExpertJ-TrecQ and CrowdJ-TrecQ datasets. For each document, we have dwell time information from many different users and we use the median dwell time on a document as the dwell time for that document as it has been shown to be a more reliable indicator of relevance than the mean [150].

Since our datasets are constructed by using frequently-clicked documents for which reliable dwell time information is available, a significant proportion of documents are labeled as relevant by our judges. This does not constitute a problem for our analysis as we are mainly interested in studying the documents that are labeled as relevant but are of low utility to the users.

Judgment Time Measurement: Each query-URL pair in all three datasets is labeled by multiple judges. The time it takes a judge to assess a document with respect to a search query is measured using the platform Crowdfunder (for CrowdJ-NaturalQ and CrowdJ-TrecQ) or using javascript (for ExpertJ-TrecQ). For each query-document pair Crowdfunder provides the response question associated with the relevance and the time a judge spent on the HIT. This is often used as a proxy for the judgment time when relevance judgments are collected via crowd-sourcing experiments. Given that each query-document is judged by at least three judges, we use the median judging time across three or more judges as the judgment time for each query-url pair.

3.3 Experimental results

The plot on the left in Figure 3.1 shows the cumulative distribution for the judging time for ExpertJ-TrecQ and CrowdJ-TrecQ datasets versus the dwell time on these datasets. The plot shows that expert judges usually require more time to label a document than crowd judges: 95% of the documents were labeled within 140 seconds by the expert judges as opposed to approximately 90 seconds for the crowd judges. The plot also shows that on certain documents users spend a substantially longer time than the judges. This is expected according to our user model as users tend to go through both Stage 1 and Stage 2 of the user model if they decide that the document is worth examining in Stage 1 whereas the judges mainly go through Stage 1. The plot on the right shows how dwell time on a document compares with judging time on that document for the CrowdJ-TrecQ dataset. It can be seen that there is no linear correlation between dwell time and judge time – judges may spend long time judging documents that have a low dwell time and vice versa⁴.

⁴Remaining datasets have very similar behavior to the CrowdJ-TrecQ dataset.

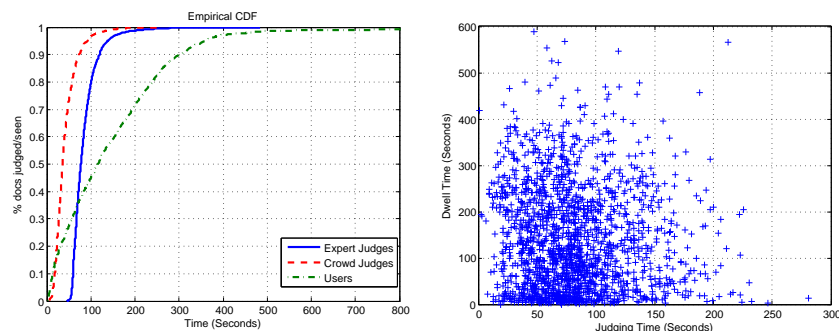


Figure 3.1: (Left) Cumulative distribution of judgment time for crowd and expert judges versus dwell time, and (Right) judging time versus dwell time.

3.3.1 Utility versus relevance

The underlying objective is to study the *utility* of a document for an actual user versus the relevance of the document. For this purpose, we divide the documents of three datasets into four scenarios according to our user model as described in Section 3.1: 1) low dwell time and low judgment time, 2) high dwell time and low judgment time, 3) low dwell time and high judgment time, and 4) high dwell time and high judgment time. In order to label each document as having low or high dwell time, and low or high judgment time, we use the following thresholding strategies:

Low vs. High Dwell Time: In Section 2.2.2, we provided an overview of how dwell time has been used in the literature to infer the relevance of a document. Previous work [72, 60, 64, 65, 150] showed that a short dwell time (typically less than 20 or 30 seconds) reliably indicates that the document was not found to be relevant by the user, as he or she decided to stop considering the document quickly. More recently a dwell time threshold of 30 seconds has become the standard threshold used to predict user satisfaction [69, 72, 65, 70, 71]. Therefore, we use a dwell time threshold of 30 seconds to identify documents with low dwell time. Hence, a document d_k is considered relevant if its median dwell time t_{kd} is equal to or greater than 30 seconds, i.e. $t_{kd} \geq 30$ and non-relevant if its median dwell time is less than 30 seconds, i.e. $t_{kd} < 30$ seconds.

Low vs. High Judgment Time: For each dataset \mathcal{D}_i , we compute its median

Judg. \ Dwell	High	Low	High	Low	High	Low
	High	593/625	112/134	588/644	88/114	1903/1957
Low	650/654	116/125	595/635	121/145	1974/1987	236/237
Datasets	CrowdJ-TrecQ		ExpertJ-TrecQ		CrowdJ-NaturalQ	

Table 3.3: Dwell time vs. judging time on various datasets

judgment time $Md(\mathcal{D}_i)$. We use it as a threshold to divide each dataset into two parts. A document d_k whose judging time t_{kj} is less than the dataset’s median judging time, i.e. $t_{kj} \leq Md(\mathcal{D}_i)$ is considered to have *low judgment time*. Similarly, a document whose judging time is greater than the dataset’s median judging time, i.e. $t_{kj} > Md(\mathcal{D}_i)$ falls into the category of *high judging time*.

Given these definitions of low vs. high dwell time and low vs. high judgment time, Table 3.3 shows the total number of relevant documents versus the total number of documents for each of the four cases of the user model. Considering each of the four cases separately with respect to our user model enables us to infer the utility of a document with respect to an actual user as follows:

1. Low dwell time, low judgment time: In our datasets, most documents that fall under this category tend to be labeled as relevant by the judges (116 out of 125 documents for the CrowdJ-TrecQ dataset, 121 out of 145 documents for the ExpertJ-TrecQ dataset and 1974 out of 1987 documents for the CrowdJ-NaturalQ dataset). This is perhaps a result of how these datasets were constructed i.e these documents were clicked by multiple users. Given that judges labeled most of these documents as relevant, it seems that such documents were highly relevant to the query and the user was able to quickly locate the required information to satisfy the information need.

2. High dwell time, low judgment time: Of the documents that fall into this category, we found that 99%, 93% and 99% in CrowdJ-TrecQ , ExpertJ-TrecQ and CrowdJ-NaturalQ are marked relevant by the judges. Given that judges have marked most of these documents relevant, the high dwell time indicates that these pages are also engaging for the users.

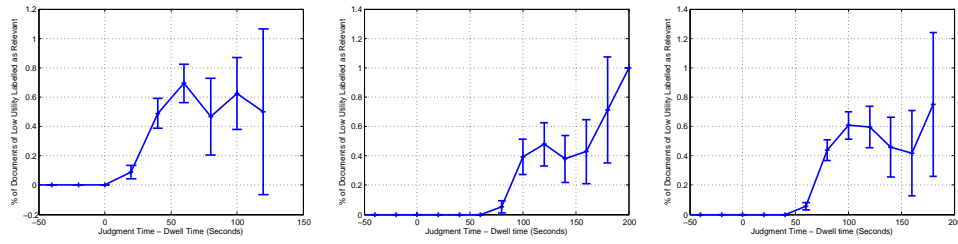


Figure 3.2: Percentage of low utility documents labeled as relevant versus difference between judging time and dwell time for the (left) CrowdJ-TrecQ , (middle) ExpertJ-TrecQ , and (right) CrowdJ-NaturalQ datasets.

3. Low dwell time, high judgment time: A lower fraction of documents is labeled relevant by the judges (112 out of 134 (84%) for the CrowdJ-TrecQ dataset, 88 out of 114 (77%) for the ExpertJ-TrecQ dataset and 213 out of 218 (97%) for the CrowdJ-NaturalQ dataset). Since the judges take a long time, it is unlikely that these documents are obviously relevant or non-relevant. However, low dwell time indicates that the users are not willing to spend same effort as judges to extract useful information from the document. Therefore, the documents that fall under this category tend to be of *low utility* for the user.

4. High dwell time, high judgment time: Of the documents that fall into this category, we found that 94%, 91% and 97% in CrowdJ-TrecQ , ExpertJ-TrecQ and CrowdJ-NaturalQ are marked relevant by the judges. Higher judging time and dwell time indicates that both judges and users spend significant amount of time in extracting utility from these documents.

Out of these 4 cases, we are mainly interested in case (3) as in that case the utility of a document to an actual user seems to be different than the relevance of the document labeled by the judges. We perform the analysis (presented in the subsequent sections) on these documents. Our hypothesis in this chapter is that this difference between utility and relevance occurs when a judge spends far too much time on a document compared to a real user (case 3 in Section 3.1.1).

Figure 3.2 shows how the percentage of documents that have high judging time and low dwell time, i.e are of low utility for the users but are labeled as relevant, changes as the difference between judging time and dwell time increases

for CrowdJ-TrecQ (left plot), ExpertJ-TrecQ (middle plot), and CrowdJ-NaturalQ (right plot) datasets. The x-axis depicts the difference between the dwell time and the judging time. For each document d_k , we compute difference between its judging time (t_{kj}) and median dwell time (t_{kd}), i.e. $t_m = t_{kj} - t_{kd}$. We further group these differences into buckets of length 20 seconds. On the y-axis, we plot the mean percentage of low utility documents for each bucket, i.e. $U_l(t)$ where $t \in (t_i, t_i + 20)$, along with the error bars. The error bars are computed using the percentage of low utility documents for different values of judging and dwell time difference (t_m) in a bucket. For example, the error bar for the bucket of 60-80 seconds is calculated using the percentage of low utility documents for different values of t_m in the range of (60,80) seconds, i.e. $U_l(t)$ where $t \in (60, 80)$. The error bars basically represent the uncertainty in the average percentage of low utility documents for a given range of judging and dwell time difference. It can be seen that as the difference between judge time and dwell time increases, the percentage of documents that are labeled as relevant but are of low utility to users tends to also increase, showing that such cases are more likely to happen when the judges spend significantly more time on a document than the users.

Judges are likely to spend more time on documents that require high effort to find and extract relevant information and the users, quite often, may not be willing to put in this effort. Therefore, our hypothesis is that the mismatch between utility and relevance is likely to be caused by factors related to effort. In the next section, we show that effort is indeed a significant factor that causes the disagreements between relevance and utility of a document.

3.3.2 Effect of effort on document utility

Given that under case 3 of our user model, most documents that are of low utility to users are labeled as relevant in our dataset, we analyze the factors that might cause these disagreements between utility and relevance. As shown in the previous section, most of these mismatches tend to occur when judges spend a long time judging documents that users quickly decide to be of low utility. Our further hypothesis is that these may be *high-effort* documents, where people need to work relatively hard

Name	Description	Name	Description
$ARI(d_i)$	Automated Readability Index of d_i	$ARI(sentquery_i)$	Automated Readability Index of sentences with query terms in d_i
$LIX(d_i)$	LIX Index of d_i	$LIX(sentquery_i)$	LIX Index of sentences with query terms in d_i
$numsent(d_i)$	Number of sentences in d_i	$numsent(sentquery_i)$	Number of sentences with query terms in d_i
$numwords(d_i)$	Number of words in d_i	$numwords(sentquery_i)$	Number of words in sentences with query terms in d_i
$numQ(sentquery_i)$	Number of query terms in sentences with query terms in d_i		

Table 3.4: Document features associated with effort

to extract relevant information, and users decide it is not worth the effort. This might be due to several factors such as document length, difficult to read or that it contains incomplete answer to the search query.

In order to test this hypothesis, we extracted several features that are associated with effort required to locate, extract and consume relevant information in the document. Several studies have shown that readability of the document text affects both user’s reading behavior [151, 152] and understanding [102, 103, 104] of its content. This makes it imperative to include features that measure how difficult it may be for a user to consume the content of the document with respect to a given search query. In particular, we extracted three readability indices and several features associated with the document length or the location of the query terms in the document. Table 3.4 provides the list of features used in this experiment.

Prior work has shown that users tend to scan documents and only read parts of the documents that they find relevant [153]. Hence, we do not assume that the users read the entire document, but instead they may search for query terms and read sentences [144] that contain some or all the query terms. Therefore, we also extract features related to the readability of the sentences in which query terms occur. We assume that users who find query terms in a sentence may also read the neighboring sentences, more specifically the previous and the next sentence. Thus, we also include these sentences to compute query-based summary features from the

document.

To measure the readability of the documents and sentences with query terms, we use Automated Readability Index (ARI) [154] and LIX. Chandar *et al.* [98] used these statistics to analyze the effect of the readability level of a document on assessor disagreement. They found that document length was not a significant factor in explaining disagreement and documents with lower readability levels provoked higher disagreement amongst judges. Since these findings are interesting and counter-intuitive, we wanted to test their importance in explaining disagreements between users and judges. ARI produces an approximate representation of the US grade level needed to understand the text and is defined as follows:

$$ARI = 4.7 \frac{chars(d_i)}{words(d_i)} + 0.5 \frac{words(d_i)}{sent(d_i)} - 21.43 \quad (3.1)$$

where $char(d_i)$ is the number of letters, numbers, and punctuation marks, $words(d_i)$ is the number of words, and $sent(d_i)$ is the number of sentences in a document d_i .

LIX [114] is another index used to represent readability level of a document. It is computed as

$$LIX = \frac{words(d_i)}{period(d_i)} + \frac{longWords(d_i) * 100}{words(d_i)} \quad (3.2)$$

where $words(d_i)$ is the number of words, $period(d_i)$ is the number of periods (defined by period, colon or capital first letter), and $longWords(d_i)$ is the number of long words (more than 6 letters) in a document d_i .

These readability estimates output a grade level, where higher level indicates a requirement of higher proficiency in English language and vocabulary. Readability level of sentences with query terms are calculated by treating them as independent documents. Primary features for a document (d_i) and sentences in d_i that contain the query terms ($sentquery_i$) are summarized in the Table 3.4.

We first analyze factors that might cause users to spend different amount of time on clicked documents. We use linear regression to predict dwell time from the features described above, and identify the factors that have a significant contribution

Feature	TRECQ				NaturalQ			
	B	SE	t-val	p-val	B	SE	t-val	p-val
$ARI(d_i)$	0.49	0.15	3.10	0.001*	0.04	0.08	0.54	0.587
$LIX(d_i)$	0.45	0.54	0.82	0.40	0.38	0.24	1.63	0.103
$numsent(d_i)$	-0.01	0.09	-0.08	0.93	-0.05	0.08	-0.61	0.539
$numwords(d_i)$	-0.01	0.01	-0.36	0.71	0.00	0.00	1.36	0.175
$ARI(sentquery_i)$	-0.67	0.23	-2.85	0.004*	0.05	0.08	0.65	0.517
$LIX(sentquery_i)$	-0.25	0.43	-0.57	0.56	-0.60	0.19	-3.24	0.001*
$numsent(sentquery_i)$	0.16	0.29	0.54	0.58	0.16	0.47	0.35	0.729
$numwords(sentquery_i)$	3.37	6.44	0.52	0.60	-5.93	2.67	-2.22	0.026*
$numQ(sentquery_i)$	-0.36	0.55	-0.66	0.50	-0.91	0.74	-1.23	0.219

Table 3.5: Regression model for TRECQ and NaturalQ median dwell time prediction. * denotes predictors significant at the $p < 0.05$ level.

for predicting dwell time.

Table 3.5 shows the breakdown of regression model for predicting median dwell time of query-url pairs in ExpertJ-TrecQ , CrowdJ-TrecQ and CrowdJ-NaturalQ datasets, respectively. Note that since the ExpertJ-TrecQ and CrowdJ-TrecQ datasets contain the same query-URL pairs, there is a single column in the table labeled *TRECQ* for these two datasets, and another column *NaturalQ* for CrowdJ-NaturalQ dataset. **Bs** are unstandardized regression coefficients in seconds, and **SEs** are standard errors of those coefficients. **ts** are t-statistics, and features that have a significant contribution to the model, i.e. have $p \leq 0.05$, are highlighted in bold. Both linear regression models explained some amount of variance in dwell time. The adjusted R-squared was $R^2_{adj} = 0.641$, and F-statistic was $F(9,4581) = 5.82$, $p < 0.0001$, for *NaturalQ* model. Similarly, $R^2_{adj} = 0.651$, and $F(9,1563) = 4.9$, $p < 0.001$, for *TRECQ* model. The residual standard error was 14.8 and 19.2 seconds for *TRECQ* and *NaturalQ*, respectively.

It can be seen that features related to the readability of the document and the sentences with query terms ($ARI(d_i)$ and $ARI(sentquery_i)$) seem to have a significant effect on the amount of time users spend on the documents on the *TRECQ* dataset. For the CrowdJ-NaturalQ dataset, readability of the entire document does not seem as important as the readability of the sentences with query terms, as $LIX(sentquery_i)$ and $numwords(sentquery_i)$ seem to be significant factors in the model whereas $LIX(d_i)$, $ARI(d_i)$ or $numwords(d_i)$ do not seem to be significant.

Feature	B	SE	t-val	p-val
$ARI(d_i)$	6.28E-02	3.43E-02	1.832	0.067
$LIX(d_i)$	2.76E-01	1.41E-01	-1.961	0.050*
$numsent(d_i)$	3.71E-02	2.51E-02	1.481	0.139
$numwords(d_i)$	3.05E-03	1.52E-03	-2.003	0.046*
$ARI(sentquery_i)$	-2.39E-02	6.13E-02	-0.39	0.697
$LIX(sentquery_i)$	2.00E-01	1.14E-01	1.763	0.078
$numsent(sentquery_i)$	-6.94E-02	8.01E-02	-0.867	0.386
$numwords(sentquery_i)$	-1.34	1.75	-0.765	0.444
$numQ(sentquery_i)$	1.65E-01	1.50E-01	1.1	0.272

Table 3.6: Significance of features for predicting the mismatch between utility and relevance for ExpertJ-TrecQ dataset.

We believe that this is due to the properties of the dataset: queries in the CrowdJ-NaturalQ dataset are sampled from the logs of a real search engine where most queries tend to be navigational and it is easy to spot the parts of the documents that are relevant to the information needs for these types of queries.

Of the significant regression factors, some have positive regression weights and some have negative weights. The weight associated with readability of the entire document, $ARI(d_i)$, is positive for the *TRECQ* data. This means that users tend to spend longer time on documents that are more difficult to consume. From our current data we can not tell whether they are spending the additional time on Stage 1 or Stage 2 of our user model (Section 3.1.1). Meanwhile, weights associated with sentence-level readability, such as $ARI(sentquery_i)$, are negative. Similarly, in the case of *NaturalQ* dataset, factors that are related to the readability level of the sentences with query terms, such as $LIX(sentquery_i)$ and the length of the sentences with query terms ($numwords(sentquery_i)$) tend to get negative weights, suggesting that they have a negative effect on the total amount of time users spend on these documents. This reduction in dwell time when relevant sections are difficult to consume seems to be due to users searching for the query terms in the document and giving up quickly when they realize that these sections are too difficult to read.

We now focus on the reasons as to why the utility of a document with respect to actual users sometimes differs from the relevance of a document (bold cells in Table 3.3). We use linear regression to predict the difference between judging time

Feature	B	SE	t-val	p-val
ARI(d_i)	2.00E-02	4.23E-02	0.473	0.636
LIX(d_i)	-3.82E-02	1.45E-01	-0.264	0.792
numsent(d_i)	5.11E-02	2.39E-02	2.142	0.033*
numwords(d_i)	4.01E-03	1.54E-03	-2.607	0.009*
ARI($sentquery_i$)	2.46E-02	6.27E-02	0.393	0.694
LIX($sentquery_i$)	1.41E-01	1.15E-01	1.224	0.221
numsent($sentquery_i$)	-8.39E-02	7.77E-02	-1.08	0.280
numwords($sentquery_i$)	-1.21	1.70	-0.71	0.478
numQ($sentquery_i$)	1.71E-01	1.45E-01	1.184	0.237

Table 3.7: Significance of features for predicting the difference between judging time and dwell time for CrowdJ-TrecQ dataset.

and dwell time of a document and determine which properties may cause a mismatch between both user’s and judge’s time spent on examining the page. We fit the model to predict the difference between median dwell time and median judging time of a document.

Table 3.6, 3.7 and 3.8 contain the breakdown of regression models for query-url pairs in ExpertJ-TrecQ , CrowdJ-TrecQ and CrowdJ-NaturalQ datasets, respectively. **Bs** are unstandardized regression coefficients, and **SEs** are standard errors of those coefficients. **ts** are t-statistics, and features that have a significant contribution to the model, i.e. have $p \leq 0.05$, are highlighted in bold. The adjusted R-squared was $R^2_{adj} = 0.44$, and F-statistic was $F(9,886)=10.83$, $p < 2.2e - 16$, for ExpertJ-TrecQ dataset. Similarly, the adjusted R-squared was $R^2_{adj} = 0.38$, and F-statistic was $F(9,759)=10.56$, $p < 2.2e - 16$, for CrowdJ-TrecQ model. Finally, $R^2_{adj} = 0.37$, and $F(9,4577)=40.66$, $p < 2.2e - 16$, for CrowdJ-NaturalQ model. The residual standard error was 4.2, 3.9 and 3.5 seconds for ExpertJ-TrecQ , CrowdJ-TrecQ and CrowdJ-NaturalQ , respectively.

Features that are related to the readability of the entire document, such as $LIX(d_i)$, $numsent(d_i)$ and $numwords(d_i)$ seem to have an important contribution as to why users find a relevant document of low utility. All three features have a positive coefficient across all datasets. It suggests that the gap between dwell time and judging time increases with increase in document length and its readability. This indicates that users optimize for properties such as readability or length

Feature	B	SE	t-val	p-val
ARI(d_i)	-5.20E-03	1.40E-02	-0.371	0.71
LIX(d_i)	1.19E-01	4.29E-02	2.763	0.005*
numsent(d_i)	3.14E-02	1.52E-02	2.071	0.03*
numwords(d_i)	8.34E-04	4.37E-04	-1.91	0.05*
ARI($sentquery_i$)	1.42E-02	1.47E-02	0.967	0.33
LIX($sentquery_i$)	-1.38E-01	3.39E-02	-4.077	4.65E-05*
numsent($sentquery_i$)	-1.57E-01	8.48E-02	-1.856	0.06
numwords($sentquery_i$)	-1.81E-01	4.86E-01	-0.373	0.70
numQ($sentquery_i$)	1.48E-01	1.35E-01	1.094	0.27

Table 3.8: Significance of features for predicting the mismatch between utility and relevance for CrowdJ-NaturalQ dataset.

besides document relevance which is not considered by relevance assessors. This further suggests that we should measure the utility of a document for an actual use by incorporating effort as well as relevance into our judging procedure.

Discussion and limitations

In this chapter, we wanted to investigate whether effort required to consume a document can effectively explain the difference between relevance determined by implicit user feedback (e.g. dwell time) and manually judged relevance by expert or crowdsourced judges. With the help of three datasets gathered using different methods, we first showed that indeed a fraction of documents exists for which the relevance judgments obtained by the assessors is different from the implicit judgments obtained using dwell time. We saw that 84%, 77 % and 97% percent of documents in our datasets fall in the category where users spend very little time on the page but judges spend significantly more time judging them as relevant.

Our hypothesis was that existing judgments do not capture how much effort it takes an end user to extract and consume relevant information from a webpage which leads to the mismatch between explicit and implicit relevance judgments of clicked documents. To test this hypothesis, we first predicted dwell time with a set of effort and relevance based features. We used readability of the page and query-based snippets as features to capture effort. We found that different features were important in reliably predicting the dwell time across datasets. Positive coefficients

of document-based readability features shows that users take longer to read difficult documents which is in line with the previous work [71]. We also found that different readability features calculated from query biased snippets have negative coefficients, which indicates that users may spend less time on documents where query focused nuggets were difficult to read.

We further tested whether relevance and effort based features could reliably predict the difference in dwell time and judging time across all datasets. Document length and readability features were found to be significant in predicting this mismatch. It is interesting to note that query-based snippet features were not significant in predicting the difference across all datasets but were significant in predicting dwell time. However, the negative coefficients indicate that difference between dwell time and judging time decreases as the length of query-based snippets increases. This suggests that document with query specific nuggets would elicit similar time of examination from both judges *and* users.

Positive coefficients of readability features, however, indicate that the gap between judging time and dwell time would widen as query-based nuggets in the document become difficult to read, i.e. users will spend far less time examining such documents than judges. Overall, our findings show that effort related features can explain the mismatch between judging time and dwell time.

It is worth noting that our work has several limitations, some of which we shall address in subsequent chapters. First and foremost, we do not explicitly ask users to label documents for relevance, instead we derive it from dwell time, which would yield a cleaner dataset for analysis. Unfortunately, gathering explicit labels from users while they visit pages in real-time is challenging and impractical as it deteriorates overall user experience. Thus, while it is a standard practice to use dwell time as an indicator of relevance, it may have induced some noise in our datasets. We also did not gather explicit labels for the amount of effort they invested in extracting useful information from a clicked document. This also suffers from the same challenges as explicit labeling of relevance. However, we do gather these judgments in subsequent chapters from judges, to determine which factors repre-

sent effort (Chapter 4) and how do different judge, query and document specific attributes affect these judgments (Chapter 5), respectively.

Finally, we also use a very limited set of features for our experiments, which we shall expand on in the next chapter. It is important to note that regression analysis just confirms that there is a mismatch between judges and user, we could explore more sophisticated non-linear models to better predict both dwell time and the difference between dwell time and judging time in the future.

3.4 Conclusion

In this chapter, we empirically evaluated whether implicit document relevance differs from manual explicit judgments obtained from assessors. We used three datasets and showed that for a fraction of documents, dwell time based relevance disagreed with manual judgments of relevance.

Our hypothesis was that the existing procedure of obtaining relevance judgments does not capture how much effort a user must invest into extracting utility from a document. We also discussed a user-model based on eye-tracking studies and related research on web-page examination [65, 144, 145, 146] which shows that user behavior and expectation could be different from that of the judges when reading a document. To test our hypothesis, we used a set of handcrafted features to a) predict dwell time, and b) predict the difference between dwell time and judging time. Regression analysis showed that features related to readability of a document play an important role in the possible mismatch between relevance of a document versus its utility to an end user.

Given these findings, in subsequent chapters we investigate what factors may explicitly capture effort. We gather explicit judgments to understand what factors are closely associated with effort. We ask judges to provide information regarding the effort required to find relevant information in a document as well as document readability and understandability. We investigate which parameter agrees the most with satisfaction which is subsequently used to evaluate the performance of several systems for relevance and *effort*.

Chapter 4

Effort based judgments in IR

In the previous chapter, we compared explicit judgments gathered from assessors and implicit relevance judgments derived from user dwell time. For some documents, we found that there was a difference in these two types of judgments. We used several features associated with document length, readability, and query-based snippets to predict the difference in document dwell time and judging time. We found that the disagreement of judges and the users occurred for documents that were long or had a relatively high reading level. Though we detected and analyzed the abandonment of documents marked relevant by the judges, we did not elaborate on what constitutes *effort*. With the help of some features, we showed that factors associated with ‘*effort*’ could effectively explain the mismatch between document’s utility for an end user and its relevance assigned by a judge but did not gather explicit labels for effort. In particular, we did not gather human judgments of factors besides relevance, that may contribute to effort or overall user satisfaction.

In this chapter, we build on the findings of the previous chapter by first identifying factors that characterize effort, where effort can be defined as the amount of time or actions required to *find, read and understand relevant information in a document*. We conduct further analysis of factors that distinguish *two relevant documents* on the basis of effort. Our aim is to determine whether effort impacts user’s preference of documents when relevance is controlled (i.e. kept constant). This chapter investigates the following hypothesis :

Hypothesis 2: Factors such as readability, the ability to find or understand relevant information are useful in estimating effort required to consume a document.

Mainly, we focus on three factors that might affect the effort required to find and consume relevant information 1) *Findability* or easiness to find relevant information in a document, 2) *Readability* or readability level of the document, and 3) *Understandability* or easiness to understand the content of the document. We investigate whether these factors: *understandability*, *readability* and *findability* are useful in distinguishing two documents of *equal* relevance grade.

We conduct experiments to obtain explicit judgments for these factors and analyze which of these factors are significant for user satisfaction. We show that 1) it is possible to obtain judgments 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 show that easiness to find 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.

Since document relevance is of prime importance in evaluation, significant work exists that studies factors that are important in determining relevance and whether they remain constant or evolve with user's interaction with search results. Taylor's work [50] with two longitudinal studies investigated the association between the search process and 15 different relevance criteria. They found both '*Structure*' and '*Understandability*' became more important to subjects during later search stages and are pre-requisite to positive relevance judgments. We use web page oriented features to capture '*Structure*', and language specific readability measures to capture '*Understandability*' in our experiments.

Analysis in Chapter 3 [155] 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 analyze the effect of incorporating effort into retrieval evaluation. We analyze systems submitted to TREC Web Track¹ Ad-hoc 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 elaborate on effort parameters and related user studies in Sections 4.1, 4.2 and 4.3, respectively. Our experiments and findings are reported in Section 4.4. Evaluation of Trec web track submissions is described in Section 4.5. We conclude our findings and discuss the future work in Section 4.6.

4.1 Factors associated with effort

By comparing relevance judgments with implicit signals of user satisfaction obtained via click logs, the previous chapter shows that 1) for certain documents there is a mismatch between the utility of a document to an actual user and its relevance, 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 judgments and the behavior of real users but did not involve direct judgments of effort or analysis of how such judgments 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 judgments of effort from relevance assessors and that incorporating these judgments 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 judgments associated with these factors. We then analyze which of these effort related factors tend to be important for user satisfaction.

We base our selection of effort related factors on previous work [72, 68, 108,

¹<http://trec.nist.gov/data/webmain.html>

156] and the user model proposed in previous chapter [155] 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 behavior, 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 [72, 108, 156] that users do not read entire web pages but first scan them for relevant parts. The 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. We shall discuss the relationship between user's expertise in topic and corresponding effort labels in next chapter.

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 the 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. We attribute these factors as problems of understandability which result in an unsatisfied user. Understandability can also be affected by the layout of the page. For example, pages with a lot of outlinks or advertisements distract users [107, 157] 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 above-stated factors are important representatives of user effort in determining document relevance. We posit that given two documents of the *same* relevance grade, users will prefer a *low effort* document over a *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.

4.2 Effort based judging

We conduct a crowdsourcing study to gather labels for different characteristics associated with effort. For each query-url pair, we ask the annotators to provide explicit labels for different characteristics along with relevance and satisfaction grades. Each query-url pair is annotated by multiple workers and we use the majority label for analysis in the subsequent sections. We describe the dataset, interface and judging criteria in detail in the subsequent subsections.

Dataset

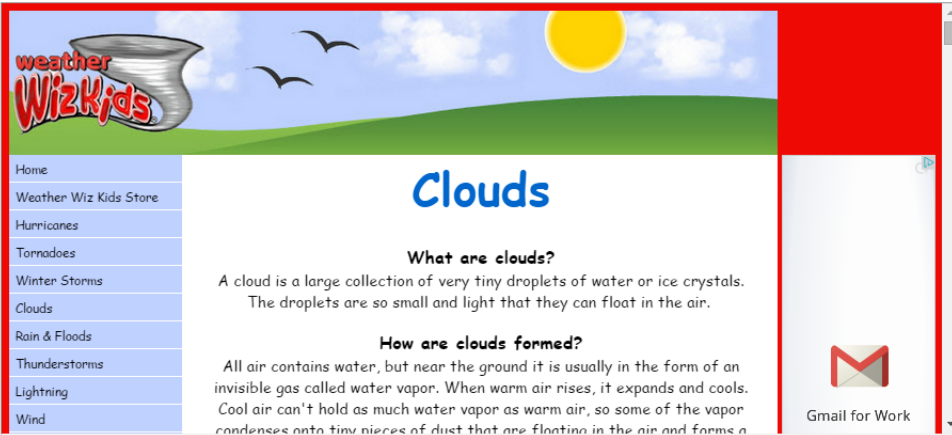
For this study, we use data from Kazai et al. [147], which consists of queries from TREC Web Track Ad Hoc task in 2009 and 2010. This dataset was also used in Chapter 3. The procedure of constructing this dataset is covered in detail in Sec-

Instructions

Suppose you submitted the following query to a search engine and the document below was shown as a result.

Search Query : what are clouds

If page does not load please visit <http://www.weatherwizkids.com/weather-clouds.htm>



1. Would you be satisfied (happy) with this search result?
 - Yes
 - No
 - Somewhat
 - Cant Judge (skip the rest of the questions)
2. Is this document relevant to the query?
 - Nonrelevant
 - Somewhat Relevant
 - Relevant
 - Highly Relevant
3. How difficult was it to understand the document?
 - Very easy
 - Easy
 - Somewhat difficult
 - Very Difficult
4. Is the language easy to read?
 - Very easy
 - Easy
 - Somewhat difficult
 - Very Difficult
5. Is it easy to find answer of the query in the document?
 - Very easy
 - Easy
 - Somewhat difficult
 - Very Difficult

Figure 4.1: Sample effort hit

tion 3.2.1. The full dataset contains 1603 URLs and 100 queries, where each query URL pair is judged on 5 grades of relevance by expert judges. Since the dataset comes from an old TREC collection, inactive URLs were excluded from the analysis. Since effort is a factor that can affect user satisfaction only when a document is relevant, in this study we exclude any non-relevant documents from our analysis, eliminating the lowest of the 5 grades.

Our goal is to control for relevance and focus on effort-related differences in this study. Therefore, in this chapter, we collect judgments on pairs of documents that are of the same relevance grade but may differ on the basis of effort. We use the significant features from Table 3.6 in Chapter 3 to sample query-document pairs for labeling. 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 [158] given in Equation 3.2 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 expert labels) and also eliminating one-word queries since such queries would be ambiguous and difficult to judge for crowd judges.

Hence, for each query in our dataset, we sample documents that have the same relevance grade but maximum difference between the number of words in the document and the readability index LIX of the document. This ensures that both high and low effort but relevant documents would be covered in our analysis. Post-sampling, our dataset consists of 80 queries and 166 documents, where for each query there are at least two relevant documents with a wide gap in the values of two features.

Labeling methodology

We gather judgments for each effort characteristic separately: the effort needed to find the 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 judgments about how satisfied they are with the document, and the relevance of the document. We use four grades to gather labels for relevance use : ‘*highly relevant*’, ‘*relevant*’, ‘*somewhat relevant*’ and ‘*not relevant*’. It is worth noting that all documents used in the study are relevant, hence we use ‘*not relevant*’ as a trap option to detect, block and remove judges that cheat or incorrectly label documents. The use of trap questions has been recommended previously [159] to improve the quality of labels gathered from crowdsourcing experiments. Satisfaction is measured on the following scales: ‘*yes*’, ‘*somewhat*’, ‘*no*’ and ‘*can’t judge*’. A sample hit with all the questions is shown in Figure 4.1.

We used Amazon Mechanical Turk² to obtain preference labels where each tuple (*query*, *url*) in a HIT was judged by three labelers where \$0.04 was payed for each HIT. We recruited judges from English speaking US region only. Each

²<https://www.mturk.com/>

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

Table 4.1: Effort label distribution

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

Table 4.2: Inter-rater agreement

judge was required to have at least 95% acceptance rate or had 5000 HITs approved by requesters on the platform. This ensured that we recruited judges who were familiar with crowdsourcing experiments and were efficient in completing tasks on Mechanical Turk. Each judge was allowed 30 minutes to complete each HIT. In total, 45 judges completed 580 HITs and the total cost of the experiment was \$20.4.

Judgment characteristics

We use the majority vote of the labelers in order to get the final judgment on a document. After removing spurious labels (determined by time spent on the task) and ‘*can’t judge*’ cases, ground truth relevance labels from TREC collection has following distribution: 114, 29, 11, 12 documents are marked ‘*highly relevant*’, ‘*relevant*’, ‘*somewhat relevant*’ and ‘*non-relevant*’, respectively.

Figure 4.1 summarizes data collected from this experiment. Relevance labels obtained from MTurk has following distribution: 76, 52 and 22 documents have been marked ‘*highly relevant*’, ‘*relevant*’, ‘*somewhat relevant*’, respectively. We obtain following judgment for satisfaction: 143, 15 and 8 documents have been marked ‘*yes*’, ‘*somewhat*’ and ‘*no*’, respectively. The mean and standard deviation of time spent on task was 63 seconds and 141 seconds, respectively.

In order to measure the reliability of the judgments obtained and the inter-annotator agreement, we use Krippendorff’s alpha (α) [160]. As shown in Figure 4.2, Krippendorff’s alpha (α) of effort judgments lie in the range of 0.22 and 0.38 which is comparable or even higher than the alpha values observed in the previous work that measures the inter-annotator agreement between assessors that judge relevance of the documents [161, 162, 163].

The inter-annotator agreement appears to be the highest for relevance and sat-

isfaction (0.38). In terms of effort based judgments, 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 a more objective question, while understandability and readability are more subjective. They depend on the judge's background knowledge, reading level, intellectual capacity, etc. Therefore, judgments associated with findability seem to be more reliable than the other two judgments. However, the inter-annotator agreement for understandability and readability is still comparable and even higher than the agreement values reported for relevance judgments in the previous work [161].

Relationship between satisfaction and effort characteristics

Given that retrieval evaluation aims at predicting user satisfaction, the primary focus has been on getting judgments of relevance and assuming that user satisfaction is a direct function of relevance. The Spearman's rank correlation (r_s) between satisfaction labels and relevance judgments ($r_s=0.375$, $p\text{-val}=0.03$) is the highest. The correlation between findability and satisfaction ($r_s=0.24$, $p\text{-val}=0.01$) follows next. High correlation between relevance and satisfaction is in-line with the previous work [77] which shows that they are correlated for higher relevance grades. The other two factors associated with effort, understandability ($r_s=0.15$, $p\text{-val}=0.23$) and readability ($r_s=0.09$, $p\text{-val}=0.10$) do not seem to be significantly correlated with the satisfaction.

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 significant factors in effort, we believe that one reason for this is due to the highly subjective nature of these judgments. Hence, these aspects associated with effort should be investigated on a personal basis and they should be considered in the context of personalized retrieval and personalized evaluation of retrieval systems, which is outside the scope of this study.

Feature	B	SE	t-val	p-val
Findability	0.13	0.042	3.02	0.003
Understandability	0.11	0.057	1.94	0.054
Relevance	0.18	0.041	4.37	0
Readability	-0.05	0.056	-0.91	0.364

Table 4.3: Factor importance for satisfaction

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

Table 4.4: Preference and effort factors agreement

To further validate this claim, we use a linear regression model to predict satisfaction of a document given these three factors associated with effort and relevance. The breakdown of regression model is given in Table 4.3. **Bs** are unstandardized regression coefficients in seconds, and **SEs** are standard errors of those coefficients. **ts** are t-statistics, and factors that have a significant contribution to the model, i.e. have $p \leq 0.05$, are highlighted in bold. The adjusted R-squared was $R^2_{adj} = 0.32$, and F-statistic was $F(4,162) = 17.43$, $p < 6.42E - 10$, for this dataset.

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 judgments 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.

While explicit judgments give an indication of how documents can be evaluated on the basis of effort, it remains to be seen whether these judgments also reflect in user preferences. Ideally, documents where the information is ‘easy to find’ or the information is ‘easy to read’ should be preferred over those that have been labeled difficult to find, read or understand. Therefore, we conduct a follow up study which determines whether user preference of a document agrees with the explicit judgments gathered above. 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 judgments are useful in getting unbiased decisions about what judges prefer to see in a document without making them explicitly think about particular aspects associated with a document (such as relevance). Therefore, we collect preference based judgments between two *equally relevant* documents and study the correlation between the three effort based factors and judge preferences, analyzing whether any of the effort related factors are significantly correlated with the user preferences. The study and analysis are presented in the the following section.

4.3 Effort-Preference correlation

Primary aim of this experiment was to study and analyze preference correlation with effort factors defined in Section 4.1. Preference judgments have previously [164] shown to be more reliable with better inter-annotator agreement than absolute judgments. We design a similar experiment where we ask multiple judges to indicate which of the two relevant documents would they prefer more for a given query. We compare these preference judgments with explicit labels gathered in the previous section to determine which factors associated with effort are important to reliably distinguish between two equally relevant documents.

Dataset

Our main focus in this study is to analyze 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 4.2). We use the same dataset of 80 queries and 166 documents where each query has at least two documents of the same relevance grade. The sampling and nature of URL's is explained in Section 4.2. We control for relevance such that we can reliably measure whether any effort related factors can significantly affect document preference.

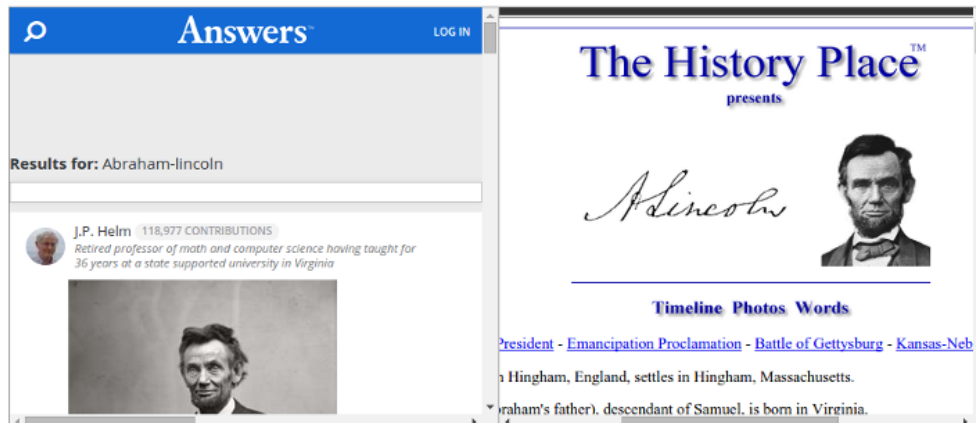
Labeling methodology

We follow guidelines suggested in previous work [164] to collect preferences over document pairs. Carterette et al. [164] study assessor agreement and compare time spent on pairwise preference judgments and graded judgments. We also show a pair

Suppose you submitted the following query to a search engine and two documents are shown as a result. Please mark the document that you would prefer to see for the query.

If page does not load please visit [S\(link1\)](#)

If page does not load please visit [S\(link2\)](#)



- Prefer Left. (I would like to see left document in the search results.)
- Prefer Right. (I would like to see right document in the search results.)
- Prefer None (I would not like to see these documents.)
- Skip these documents. (I cannot judge which document I would prefer to see, I would like to skip these documents)

Figure 4.2: An example hit for effort and preference judging

of documents for each query side by side to each user to further analyze whether the judges tend to prefer one document over the other, and whether any effort related factors are correlated with their preferences. The judging interface is shown in Figure 4.2. 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.

We used Mechanical Turk to obtain preference labels where each triplet (*query*, *url1*, *url2*) in a HIT was judged by three workers. We randomized the order of the pairs in each HIT to prevent cheating. We use the majority vote between three judges as the final judgment. We recruited judges from English speaking US region only. Each judge was required to have at least 5000 HITs approved by requesters on the platform. This ensured that we recruited judges who were familiar with and efficient in completing tasks on Mechanical Turk. Each judge was allowed 15 minutes to complete each HIT. In total, 53 judges completed 560 HITs of which 62 HITs were discarded due to cheating. We rejected HITs on which judges spent

less than 5 seconds since the median examination time of a single document in the previous experiment was 60 seconds. We paid \$0.04 for completion of each HIT and the total cost of the experiment was \$23.5.

Judgment characteristics

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 mean and standard deviation of time spent on preference interface are 33 seconds and 47 seconds, respectively.

The average pairwise percentage agreement for the set of judgments obtained though this study is 0.60 , which is higher than random agreement of 0.5 ($t(80)=2.05$, $p \leq 0.05$). Given that the judges are only shown documents of the same relevance grade, the fact that there is a significantly higher agreement than random between the judges shows that there are some additional factors that affect user satisfaction and there is an agreement amongst different judges. When compared to the inter-annotator agreement values reported by Carterette et al. [164] (which focus on getting judgments 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 judgments on documents that are of the same relevance grade.

4.3.1 Preferences vs. effort characteristics

Given the preference based judgments from the judges, we further analyze 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 4.4 shows percentage agreement between a 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. For comparison purposes, we also analyze the agreement of satisfaction based judgments (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 when tested using two tailed t-test. The high correlation of satisfaction based judgments with user preferences further confirm the reliability of the satisfaction based judgments 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 judgments is 0.60, the agreement of 0.607 between findability and user preferences (obtained via the majority vote) is comparable with the 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 along 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 analyzing how incorporating findability in building and evaluating retrieval systems could change the design of the retrieval systems.

4.4 Predicting effort and relevance

In this section, we focus on predicting the most important effort related factor among those described in Section 4.1, findability and analyze 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 the features that are important for predicting

Summary and document specific features			
sumChar	#characters in summary	docChar	#characters in document
sumWords	#words in summary	docPunct	#punctuations in document
sumPunct	#punctuations in summary	docSentQT	#document sent with query terms
sumSent	#sentences in summary	docWords	#words in document
sumSentQT	#summary sent with query terms	docSent	#sentences in document
Readability features			
docARI	ARI Index of document	sumARI	ARI Index of summary
docCLI	CLI Index of document	sumCLI	CLI Index of summary
docLIX	LIX Index of document	sumLIX	LIX Index of summary
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 q-terms	tRatio	Fraction of #words and #tags in html

Table 4.5: Text features used for predicting findability and relevance

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 chapter) 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.

4.4.1 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.

4.4.1.1 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 neighboring paragraphs (i.e. the summary). To create a summary of a document containing the query terms, we simply use a sentence that contains any query term and its immediate neighbors in the document. Similar features have been used previously in [155]. The features are summarized in Table 4.5.

- Typically, lengthier documents may require more effort than shorter documents. Hence, these features capture the length of the 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) [165], Automated Readability Index (ARI) [166] and LIX [158]. These metrics are calculated by counting number of words and sentences, and are used as a rough estimate for a document's difficulty. These features are calculated both for the entire document and the 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 the text and the 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.

4.4.1.2 Features associated with webpage structure

Users interact more with complex webpages today than with plain text documents and the layout of the document can be instrumental in finding information in a document. Thus, it is important to leverage underlying information in an html page to build stronger features. We propose the following set of features capturing different aspects of effort. Webpage structure oriented features are given in Table 4.6.

- The first set of features are associated with the tag distribution in a document. We consider tag distribution to be a signal of how well information is organized in a web page. Pages with a lot of text or images may not be useful as navigation would become difficult. Thus, percentage of tables, images, headings, paragraphs, lists and outlinks are extracted as features.
- Outlink distribution of a page is useful because too many outlinks can be distracting and hinder readability of the document. We consider fraction of

Structure 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 and text tags
fOutPage	Fraction of hrefs to same page		
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 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

Table 4.6: Webpage structure features used for predicting findability and relevance

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 the webpage tend to be more important and attract more attention from the users than others. Especially those that have headings (useful for skimming) or contain query terms as they help find information faster. We use number of headings that contain query terms, their min, max and average position as features. Similar features are extracted from the outlinks.
- Summary specific features are important for finding information quickly. We use number of such spans in a document, min, max and average position of such spans, their average length, and spans that cover headings as features.

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 but one could possibly add more features to capture different aspects of effort.

4.4.2 Predicting findability

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

Feature	Coeff	Feature	Coeff
<i>sumChar</i> *	-9.5	<i>sumWords</i> *	8.44
<i>maxWinPos</i> *	-2.4	<i>docARI</i> *	2.24
<i>docCLI</i> *	-1.6	<i>minWinPos</i> *	1.64
<i>docSentQT</i>	-1.39	<i>queryFreq</i>	1.09
<i>fTable</i> *	-0.68	<i>meanPosOut</i> *	0.88
<i>sumLIX</i>	-0.67	<i>fImg</i> *	0.63

Table 4.7: Findability features ³

predicting findability and compare and contrast the features that are important for predicting findability with features that are important for predicting relevance.

To avoid over-fitting on a small dataset, we convert different grades of findability to a binary scale. Our task now reduces to binary classification where the relative ordering between different classes (easy vs. difficult) is significant. Therefore, 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 4.2), 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 Figure 4.4, we compare the agreement of predicted Findability labels with the preference judgments obtained from relevance assessors by computing the fraction of documents preferred by the users that are predicted to be of high findability according to the regression model.

RMSE for the predictions is 0.37. These results suggest that the regression model is quite good at predicting the labels correctly. Preference agreement with predicted Findability grades is 0.587, which is comparable to agreement of 0.6 between Findability and preference judgments if actual judgments of Findability were used in the analysis (in Figure 4.4).

Among all the features used in the analysis, Table 4.7 shows the features that have high coefficients. Since features that can help users find information more quickly are more important for Findability, it is expected that both html and text features mentioned above will be important. Thus, it is not surprising that number

³* indicates statistical significance with ($p \leq 0.05$)

Feature	Coeff	Feature	Coeff
<i>docWords*</i>	-9.76	<i>sumWords</i>	6.3
<i>docPunct</i>	-9.6	<i>fBoldItalics*</i>	0.57
<i>maxWinPos*</i>	-2.07	<i>termsInTitle*</i>	0.65
<i>countH*</i>	-1.3	<i>qWinO*</i>	1.26
<i>tRatio*</i>	-0.64	<i>maxQPos</i>	1.37
<i>docSent</i>	-5.30	<i>fImg*</i>	0.41

Table 4.8: Relevance feature importance

		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 4.9: Actual vs. predicted relevance labels

of images, lists and tables are useful in predicting Findability. As expected, features such as minimum position of query terms in summary and number of words 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 the importance of answer location in the webpage in reducing the time user spends on reading/skimming the 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 chapter 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 analyze the importance of features and how they differ from Findability features.

4.4.3 Relevance prediction

Similar to the model for predicting Findability, we use our proposed features for predicting judgments 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 4.9, where each cell is fraction of documents with actual label x_i and predicted label y_i . Table 4.8 shows features that are statistically significant for predicting relevance, together with the direction and strength of correlations.

As shown in the previous work [23], document content features impact rele-

vance most followed by query and structure specific features. While features related to document length (docWords, docPunct and maxWinPos) have negative coefficients (suggesting user preference for documents with fewer terms and sentences), query and summary specific features (qWinO, termsInTitle, maxQPos and sumWords) have positive coefficients. Above table 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 be optimized for Findability together with relevance, and 3) additional features that capture the easiness to find information in the page (such as the ones proposed above) should be used in building and optimizing the 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 the retrieval systems.

4.5 Effect of effort on retrieval evaluation

Until now, we have focused on getting relevance judgments associated with effort and have shown that user satisfaction and preferences can be affected by effort related factors, in particular, by ability of find the relevant 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 [167] has shown that variations in relevance assessments does not necessarily lead to significant differences in retrieval evaluation. Given this finding, we further analyze 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 judgments for these years would be very costly and time consuming. Since our results suggest that findability can capture effort, and that findability labels correlate

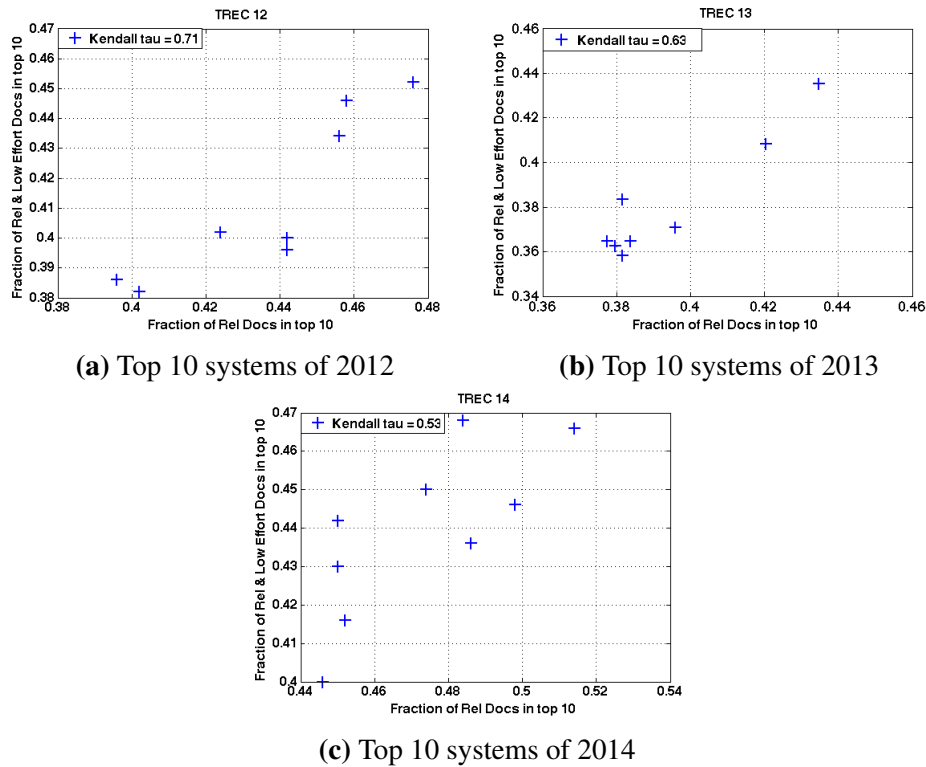


Figure 4.3: Comparison of systems based on #relevant documents vs #low effort relevant documents ($P@10$)

with user satisfaction, and that it is possible to predict findability with a good accuracy, we used the regressor in the previous section to predict findability label of a document.

Focusing on the top systems submitted in 2012-2014, we analyze 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). We then compared this value with the fraction of relevant documents retrieved in top 10 that are also low effort (i.e., findability). The results of this experiment for TREC 2012, TREC 2013 and TREC 2014 are in Fig 4.3a, Fig 4.3b and Fig 4.3c, respectively. The plots also show the Kendall's tau correlation between the ranking of systems obtained when the systems are ranked based on the number of relevant documents versus the 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 than each other when easiness to find 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 seem to be different than each other.

The same behavior 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 easiness to find information in a document for user satisfaction, the satisfaction of the users of these two search systems would be very different 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.

4.6 Conclusion

It has been shown that relevance and user satisfaction do not always agree [28, 29], and users may still be dissatisfied with their search despite being served relevant documents. Previous chapter [155] 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. However, we did not investigate what constitutes effort or how can effort judgments be obtained and incorporated in evaluation. We attempted to answer all these questions in this chapter.

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 the user preference. Our analysis indicates findability correlates well with the user satisfaction among all the above 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 analyze 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 measured as findability) is taken into account.

Our analysis suggests that effort based judgments 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 analyze different label aggregation strategies to incorporate all the effort parameters. We could also look into incorporating effort into existing evaluation metrics or proposing new effort based metrics for retrieval evaluation.

Chapter 5

Characteristics of effort judgments

While topical relevance labels have been extensively gathered and studied in the literature [7, 12, 24, 104], research on effort labels is relatively scarce. Topical relevance judgments have a limitation in that they do not represent *how much time* it would take an end user to *locate* the required information in a relevant document. As noted in Chapter 3, users may not invest sufficient time looking for relevant information. If a user finds it difficult to locate required information in a relevant document, she may abandon the document or the query altogether. It is worth noting that existing implicit signals such as dwell time [69, 150] will incorrectly imply that such documents are useful for the user.

In Chapter 4 [168] we designed experiments to identify factors that represent effort and correlate the most with satisfaction labels. We found that *Findability* i.e. the *ability to find* relevant information in a webpage is useful in distinguishing between *high* and *low* effort documents. On the basis of explicit and pairwise preference labels, we concluded that users prefer documents where information can be located quickly. We also showed that the existing systems that optimize for relevance do not perform well for judgments that represent relevance *and* effort.

In this chapter, we further investigate the nature of findability judgments. One may argue that *findability* (or effort) judgments may be subjective and can be affected by several factors. In the previous chapter, we gathered judgments with two different experiments to understand which factors characterize effort. We did not study different annotator, query or document properties that might influence these

judgments. Hence, in this chapter we investigate the following hypothesis:

Hypothesis 3: Factors associated with annotators, query, and document will affect effort labels.

We posit that the ability to find the required information from a webpage may be affected by several factors. For instance, annotator specific factors such as prior knowledge of a topic or language proficiency may greatly increase (or decrease) the amount of time they spend on locating relevant information in the webpage. Query specific factors such as query type and difficulty may influence annotator's ability to find and understand information in the webpage. Similarly, document specific factors such as length or its structure may equally aid or hinder a judge in finding required information. Therefore, in this chapter, we aim to answer two research questions. Firstly, whether annotator, query or document specific properties are correlated with findability (or effort) judgments. Secondly, for each property whether *high effort* documents can be reliably distinguished from *low effort* documents.

We posit that the *auxiliary information* gathered *during* the annotation of relevance judgments can be used to estimate factors such as *findability*. One such auxiliary information collected at no extra cost is *judging time*. During relevance judging, systems usually record the time of judgment, indicating how much time it took a judge to annotate the document for topical relevance. Even though judging time contains information about how easy it is for an assessor to locate the relevant information in a document, how this information can be used in retrieval and training has never been explored before.

Figure 5.1 shows the variance in judging time of relevant¹ documents for randomly sampled queries from TREC Web Track. The figure shows that relevance assessors can take varied amount of time to grade *two equally relevant* documents for a single topic. For example, for topic 182, judging time varies widely for 64 highly relevant documents. Since gathering effort labels manually at scale may be infeasible, we investigate whether judging time can be used as an estimate of effort. Hence, we also investigate the following hypothesis in this chapter:

¹1=Relevant document - 4=Key document

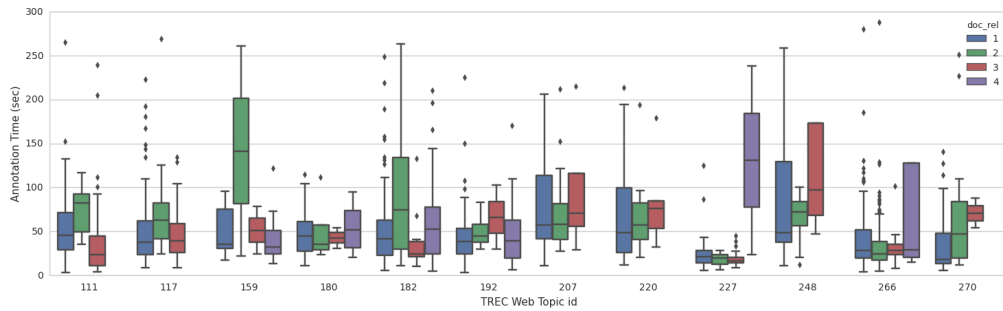


Figure 5.1: Variability in judging time with relevance labels for TREC Web topics

Hypothesis 4: Judging time can be used to estimate document effort.

Existing collections are built for topical relevance and do not contain effort-based labels. Therefore, we gathered explicit judgments from assessors using a crowdsourcing platform for this study. We sampled queries, documents and corresponding relevance judgments from a publicly available dataset. TREC Web collections [140]² contain manually assigned relevance labels with respect to a given search query. We use a subset of this collection to crowdsource effort specific judgments. We built an interface to ask annotators questions about their prior knowledge of query topic, findability and their satisfaction with the document for a particular query.

Our analysis uncovers two key trends. First, findability judgments vary with annotators prior knowledge of the query topic and the nature of the query. Second, judges spend more time locating information in *high effort or difficult* documents when compared to *low effort or easy* documents. We describe the collected dataset and its results in Section 5.1 and Section 5.2 respectively. Section 5.3 summarizes our findings and conclusions.

5.1 Methodology

The aim of this study is to investigate the effect of annotator, query and document specific parameters on effort labels. We crowd-sourced findability labels for a subset of TREC Web track query-document pairs with annotators recruited via Mechanical

²http://ir.dcs.gla.ac.uk/test_collections/

Suppose you submitted the following query to a search engine and the document below was shown as a result.

Query: symptoms of heart attack
Description: What are the symptoms of a heart attack in both men and women?

If the webpage does not load, please visit the following link.

Webpage Link: http://heart.org/HEARTORG/Conditions/HeartAttack/WarningSignsofHeartAttack/Heart-Attack-Symptoms-in-Women_UCM_436448_Article.jsp

Instructions View Remove

Heart Attack Symptoms in Women

Updated:Thu, 23 Feb 2012 8:00:00 AM

We've all seen the movie scenes where a man gasps, clutches his chest and falls to the ground. In reality, a heart attack victim could easily be a woman, and the scene may not be that dramatic. Senior Woman Sitting

"Although men and women can experience chest pressure that feels like an elephant sitting across the chest, women can experience a heart attack without chest pressure," said Nieca Goldberg, M.D., medical director for the Joan H. Tisch Center for Women's Health at NYU's Langone Medical Center and an American Heart Association volunteer. "Instead they may experience shortness of breath, pressure or pain in the lower chest or upper abdomen, dizziness, lightheadedness or fainting, upper back pressure or extreme fatigue."

Even when the signs are subtle, the consequences can be deadly, especially if the victim doesn't get help right away.

'I thought I had the flu'
 Even though heart disease is the No. 1 killer of women, women often chalk up the symptoms to less life-threatening conditions like acid reflux, the flu or normal aging.

"They do this because they are scared and because they put their families first," Goldberg said. "There are still many women who are shocked that they could be having a heart attack."

Figure 5.2: Judging interface

Turk³ and used the interface in Figure 5.2 to gather judgments.

We randomly sampled 58 topics with two types of information needs: 32 queries with *faceted* i.e. underspecified information needs which may have several subtopics and 26 queries with *single* i.e. clearly defined topics with a single information need provided in the 2013 and 2014 TREC Web track collection. We used the query along with its description to provide more specific information to the annotators. It has been shown [169] that the quality of judgments improve when the evaluators are given intent statements.

TREC Web 2009-2014 collections consist of 113263 documents labeled for relevance. Since it is infeasible to collect findability labels for such a large set of documents, we created a representative dataset by sampling some relevant documents for each query. First, we used LambdaRank[170] to train a ranker with features given in Section 4.4.1. Then, we ranked all documents for each query and selected all relevant documents up-to rank 20 for this study. We ignored non-relevant and spam documents which is in line with the experiment in Chapter 4 [168]. We obtained 356 relevant documents for judging with above method to control for relevance and explicitly focus on *findability* labels. The relevance distribution of sampled documents is as follows: 1) Key Result⁴: 8, 2) Highly relevant: 28,

³<http://www.mturk.com>

⁴A webpage or site that is dedicated to the topic; authoritative and comprehensive, it is worthy

- **Findability:** We ask you to mention whether or not you were able to find the required information in page.
- **Satisfaction:** We ask you to mention how satisfied or happy you were with the page given the query.
- **Highlights:** You should mark all the relevant parts of webpage. Highlighting is only necessary when page is relevant to the query. The [image](#) shows how to highlight parts of page. You shall only be **paid if you mark text that contains the answer** to the query.

Figure 5.3: Attribute instruction



- **Instructions:** If you forget instructions, please click on '**Instructions**' button  to view them.
- **View:** To view the text you highlighted, you can click '**View**' button. 
- **Remove:** To remove the text you highlighted, you can click '**Remove**' button. Note that it *removes all highlights*.

Figure 5.4: Highlight instructions

3) Relevant: 118 4) Somewhat relevant: 215 documents respectively.

For an in-depth analysis of effort (or findability) labels, we asked annotators to provide the following information: a) their language proficiency (*proficiency*), b) background knowledge of query topic (*familiarity*), c) satisfaction with document content (*sat*), d) how much information (*info-found*) they found in the page and e) whether it was easy to locate (*find*) the required information. The definition of findability and satisfaction used in the study are shown in Figure 5.3. We used the following scales for each label:

- **Language Proficiency** (*proficiency*): Not proficient(1) - Highly proficient(4).
- **Topic Knowledge** (*familiarity*): No Knowledge(1) - Expert(3).
- **Amount of information** (*info-found*): None, Partial and all.
- **Findability** (*find*): Very difficult(1) - Very Easy(4).
- **Satisfaction** (*sat*): Not Satisfied(1) - Highly satisfied(4).

We also asked annotators to highlight whatever information was relevant to the query in the document as shown in Figure 5.4. The highlighted portion of the text has yellow background as show in Figure 5.2. This ensured that the annotators read the document and marked important information before answering all the questions. The annotation interface with instructions is available online⁵.

We binarize above labels for our analysis to account for sparsity in the dataset. We payed MTurk annotators 0.05 cents for annotating a single document. Each

of being the top result in a web search engine

⁵<http://128.16.12.66:4730/index>, batch:xaa, workerid:user1d

Id	Query	%
274	golf instruction	0
216	nicolas cage movies	0
280	view my internet history	5
266	symptoms of heart attack	9
264	tribe formerly living in alabama	65
199	fibromyalgia	75
189	gs pay rate	86
273	wilson's disease	100

Table 5.1: % annotators not familiar with query

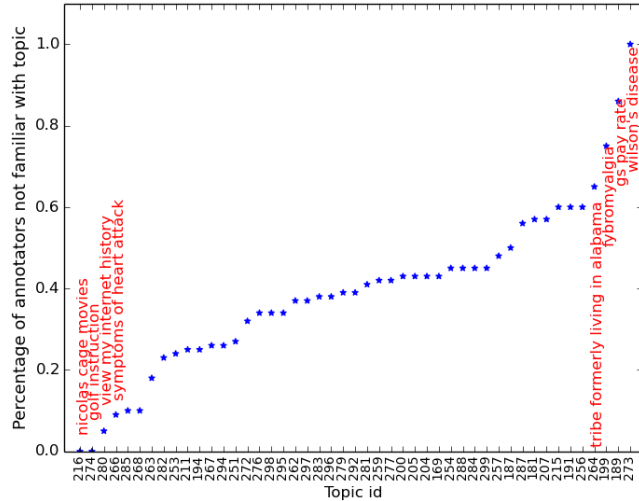


Figure 5.6: % annotators not familiar with topics

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 removed annotators that did not highlight any information in the document, did not scroll or move mouse on the document⁶, or their highlighted answers did not agree with other annotators. We computed cosine similarity between all the pairs of word vectors derived from the highlights and kept those judgments whose highlights had a similarity of ≥ 0.80 to control for variance in answers.

5.2 Results

Overall, we gathered labels for 58 queries, 356 documents from 70 judges. We analyze effect of annotator-specific variables, query and document dependent variables on effort labels. For each property (query, annotator or document specific), we analyze three things: 1) distribution of effort labels given the property, 2) distribution of judging time per property, and 3) distribution of judging time *and* effort labels for different values of effort labels.

⁶We tracked scroll and mouse movements via Javascript.

5.2.1 Annotator specific analysis

Judges play an important role in labeling documents. Annotator specific factors may affect how a judge evaluates a document with respect to a search query. We posit that language proficiency and topic knowledge are important factors that may affect a judge’s effort labeling decision.

5.2.1.1 Language proficiency

A document may get different labels when judged by annotators with different language skills. If an annotator lacks proficiency in a language, she may find it difficult to read or locate required information. However, in our dataset we did not observe a large variation in language proficiency of annotators. Of 70 annotators, only 8 annotators reported *slight proficiency* and 1 annotator reported *no proficiency* and rest reported *high proficiency* in English. This is perhaps an outcome of Mechanical Turk platform as workers on such platforms are expected/required to have sufficient knowledge of English to register for hits. We also did not observe any statistical difference in findability (*find*) labels of documents across different labels of language proficiency when tested with a chi-square test ($\chi^2(4, N = 732) = 11.63, p < 0.20$).

5.2.1.2 Query topic knowledge

Prior knowledge of the query may affect both the time and effort it takes a judge to evaluate a document. We also collect and analyze the effect of prior knowledge of query topic (*familiarity*) on annotation time and findability (*find*) labels. Some topics are relatively harder than others. Table 5.1 and Figure 5.6 depict the percentage of annotators reporting *no knowledge* for some of the queries. It can be seen that a large number of annotators are less familiar with *scientific topics* or *unique people/place/thing*.

Figure 5.7a shows the mean judging time with 95% confidence interval error bars across topic familiarity labels. Number of samples used to estimate mean and error bar size are indicated on top of each error bar respectively. Given that the judging time is a continuous (dependent) variable and topic familiarity is a categorical (independent) variable, we use Kruskal-Wallis H [171] with Dunn’s mul-

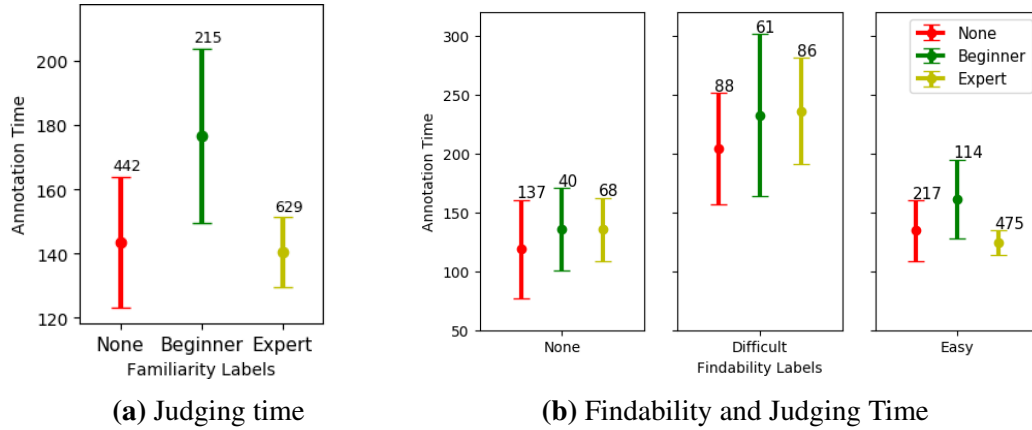


Figure 5.7: Topic familiarity and findability labels

multiple comparison test to determine whether or not population median of all the groups are equal. There was a statistical difference between the judging time of annotators with different levels of topical knowledge ($H(3)=74.36$, $p\text{-val}=7.10e-17$) with mean rank of 559.03 for judges with *no knowledge*, 579.03 for *beginner* and 715.2 for *expert* respectively. The corrected p-values for each pairwise comparison (*no knowledge, expert*), (*no knowledge, beginner*) and (*beginner, expert*) are 0.004, 0.36 and 0.03 respectively. This suggests that median judging time of annotators with *no knowledge* of topic is statistically different from *beginner* and *beginner* is statistically different from *expert*. We observed that the average time spent by the judges with *beginner* skills is the highest amongst all knowledge levels. This is expected as the judges may read the text more carefully to find required information which is in-line with the previous work [151] that studies the reading behavior of users on texts with different levels of difficulty.

When judging time is further divided by findability labels, we get the distribution shown in Figure 5.7b. It seems that the judges across all levels (none, beginner or expert) take longer to find required information in *Difficult* documents. However, it is worth noting that an annotator's knowledge about the query topic is inversely proportional to the number of documents marked *Difficult*.

Table 5.2 shows the probability of effort labels conditioned on query familiarity i.e. $P(\text{find}|\text{familiarity})$. Annotators with *no knowledge*, *beginner* and *expert* knowledge of query topic could not either find the required information (*None*) or

	None	Difficult	Easy
No Knowledge	0.31	0.20	0.49
Beginner	0.19	0.28	0.53
Expert	0.11	0.14	0.75

Table 5.2: P(find|familiarity)

found it difficult to find information (*Difficult*) in 0.51% (234/452), 47% (103/217), 25% (157/614) documents respectively. This indicates that the judges with prior knowledge about the query find larger fraction of documents easy as compared to judges with little or no expertise in the query topic. A chi-square test of independence was performed to examine the relationship between familiarity and findability labels. Chi-square test $\chi^2(4, N=1282)=83.3$, $p < 0.05$ suggests that an annotator’s prior knowledge about the query may influence her ability to locate the relevant information in a webpage.

Overall, the above analysis shows that query familiarity affects findability labels, where experts label a smaller fraction of documents as *Difficult* as compared to beginners. However, we also found that the judging time of *Difficult* documents is consistently higher than *Easy* documents across all levels of topic familiarity. We observed a fair agreement between workers while aggregating labels per query than previously reported. The inter-rater agreement (Krippendorff’s Alpha α) for *find*, *info-found* and *satisfaction* was 0.30, 0.24, 0.29 respectively. We posit that annotator-dependent variables cause variation in labels which in turn affects inter-annotator agreement.

5.2.2 Query-specific analysis

Previous work [172] has shown that the nature of a search query also affects the amount of time it takes to complete a search task. We sampled queries with two types of information need: *faceted* and *single* for our work. We analyze the effect of topic type on both judging time and findability labels.

Figure 5.8a shows the mean with 95% confidence interval error bars judging time for query with *faceted* and *single* information need. The judging time distribution of documents for *faceted* query does not statistically differ from *single*

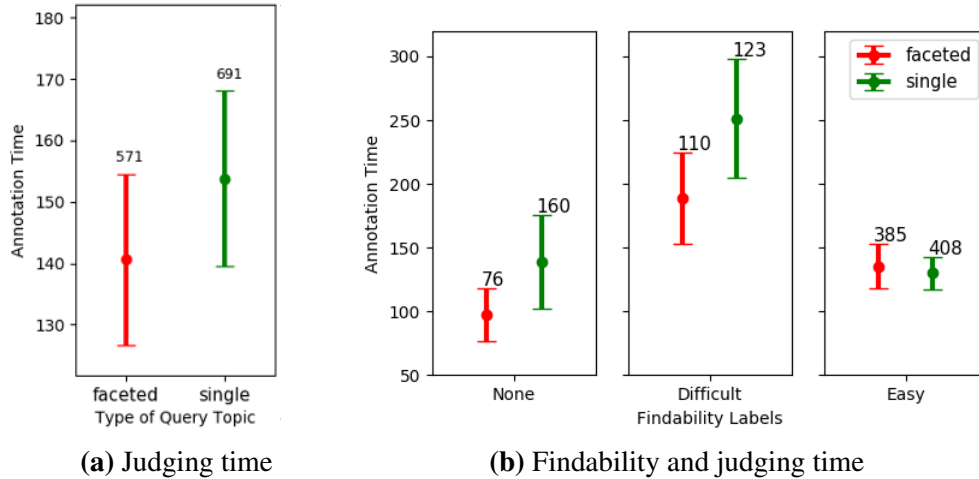


Figure 5.8: Topic type and findability labels

information need query. Figure 5.8b shows further division of judging time on the basis of findability labels. We use Mann-Whitney U test to compare categorical information needs and continuous variable judging time. Mann-Whitney U test indicates no statistical difference between judging time of different types of queries ($U=186035.5$, $p\text{-val}=0.08$) with median of 86.6 seconds for *faceted queries* and 92.5 seconds for *single queries* respectively.

We observed no statistical difference between judging time distribution of *None* documents ($U=3535.0$, $p\text{-val}=0.25$) for *single* and *faceted* information needs. There was also no significant difference between the judging time distribution of *Easy* documents ($U=83690.5$, $p\text{-val}=0.52$) for *single* and *faceted* queries. However, judging time distributions of faceted queries is statistically different from single information needs for *Difficult* ($U=6161.0$, $p\text{-val}=0.03$) documents.

Within faceted queries, the judging time distribution of *Difficult* documents is significantly higher than documents judged *None* ($U=2278.0$, $p\text{-val}=0.0003$) and *Easy* ($U=16207.0$, $p\text{-val}=1.64e-06$) respectively. The median judging time of *Difficult*, *None* and *Easy* documents for faceted queries was 132.5, 81.38 and 81.01 seconds respectively. Similarly, within the singular information needs group, the judging time distribution of *Difficult* documents is significantly higher than documents judged *None* ($U=5108.0$, $p\text{-val}=1.28e-07$) and *Easy* ($U=16248.0$, $p\text{-val}=1.94e-11$) respectively. The median judging time of *Difficult*, *None* and *Easy* documents for

	None	Difficult	Easy
faceted	0.13	0.19	0.68
single	0.18	0.23	0.59

Table 5.3: P(find|topic type)

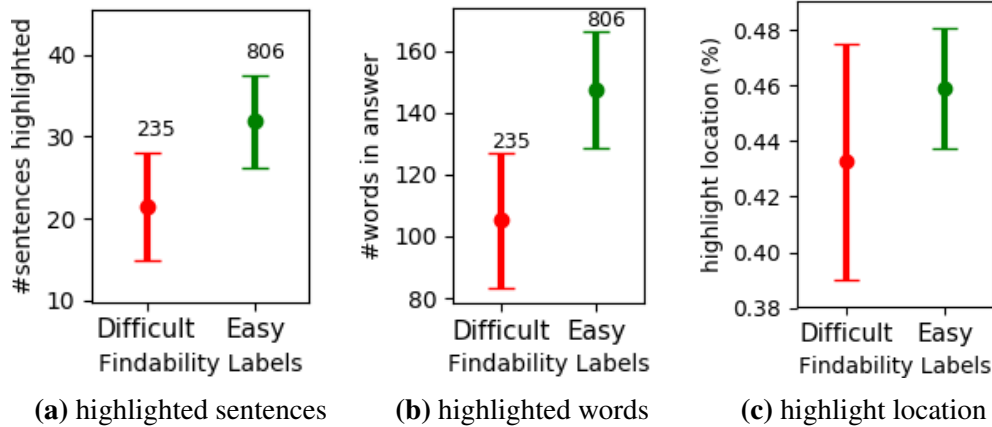


Figure 5.9: Answer words/sentences/location vs. effort

faceted queries was 162.98, 88.1 and 82.9 seconds respectively.

The probability of findability labels conditioned on query type is given in Table 5.3. Annotators have marked 32% documents of *faceted* queries as none or *Difficult* compared to 41% of the documents with *single* information need queries. This suggests that the annotators either do not find or find it more difficult to locate relevant information for queries with *single* information needs. This is expected as such queries require specific information nuggets which requires a judge to read the document more thoroughly. A chi-square test ($\chi^2(2, N=1282)=19.59, p<0.05$) indicates that the nature of information need (single or faceted) may influence a judge’s ability to locate relevant information in a webpage.

5.2.3 Document-specific analysis

In this section we analyze the association of document specific properties with effort labels. We asked the annotators to highlight all the information in the webpage that is relevant to the search query. We consider the highlighted portion of the webpage to be the *answer* for the search query. We examine two properties: number of sentences/words in the highlighted text and their location ⁷ in the webpage.

⁷It is the average depth of text relative to the page content

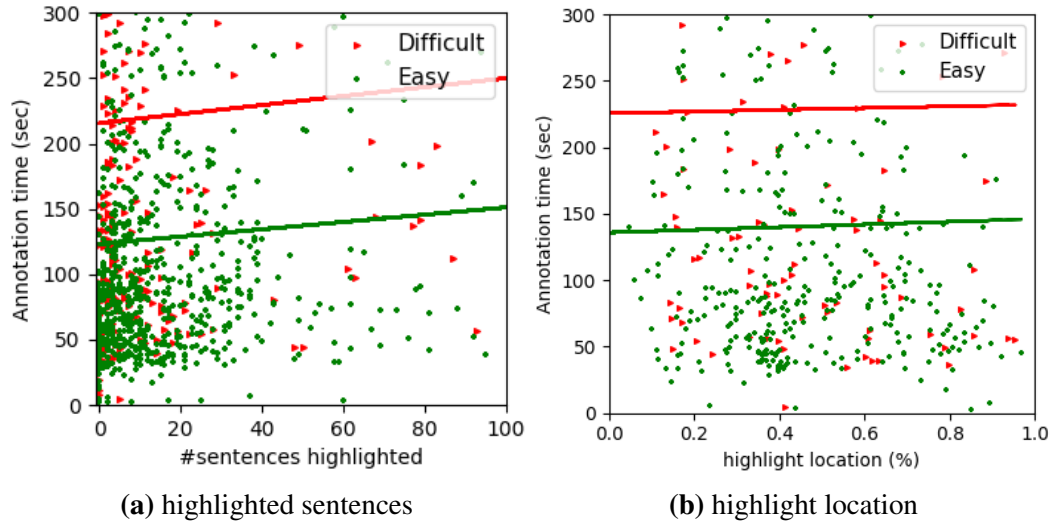


Figure 5.10: Judging time, answer sentences/location and effort

Figure 5.9a and 5.9b shows the average number of highlighted sentences and words with 95% confidence interval error bars respectively. The number of highlighted sentences was significantly higher in documents marked *Easy* (median=7) than the documents labeled as *Difficult* (median=4) when tested using Mann-Whitney U test ($U=4225.0$, $p\text{-val}=0.04$). The number of words was also significantly higher ($U=4044.5$, $p\text{-val}=0.01$) in *Easy* (median=53) documents when compared to *Difficult* (median=34) documents. It is interesting to note that the annotators highlighted fewer sentences and words in *Difficult* than *Easy* documents. The time and effort required to locate relevant information and effectively consume it in difficult pages may cause judges to focus their attention and energy on smaller nuggets of information in the webpage which perhaps lead to shorter highlights.

Figure 5.9c shows the distribution of highlighted text location across *find* labels. Location of highlights is not statistically different ($U=704.5$, $p\text{-val}=0.11$) across findability judgments. The mean (median) depth of answer (or highlighted text) in difficult and easy documents is 0.43 (0.34) and 0.45 (0.39) respectively. We also observed a weak linear relationship between judging time and number of highlighted sentences in *Easy* ($R^2=0.007$, $b=0.12$, $t(173)=0.34$, $p\text{-val}=0.72$) and *Difficult* ($R^2=0.003$, $b=0.57$, $t(59)=0.42$, $p\text{-val}=0.66$) documents respectively. Similar weak linear relationship was observed between judging time and answer location

Feature	Document		Summary	
	μ (σ)	ρ	μ (σ)	ρ
# words	831.3 (832.4)	0.32*	630 (601.9)	0.27*
# sentences	146.6 (208.3)	0.30*	86.1 (106.1)	0.24*
LIX	33.3 (7.6)	-0.009	33.6 (7.4)	0.02
ARI	14.3 (6.3)	-0.09	15.7 (7.3)	-0.07
CLI	12.8 (2.9)	-0.02	13.11 (3.0)	0.007
outlinks	0.33 (0.13)	-0.11*	-	-
images	0.07 (0.08)	-0.01	-	-
tables	0.02 (0.04)	0.01	-	-

Table 5.4: Feature distribution and correlation with judging time

in *Easy* ($R^2=0.03$, $b=5.9$, $t(173)=0.25$, $p\text{-val}=0.79$) and *Difficult* ($R^2=0.07$, $b=9.8$, $t(59)=0.20$, $p\text{-val}=0.83$) webpages respectively. Both these linear curves are depicted in Figure 5.10a⁸ and Figure 5.10b respectively.

We use Pearson’s rho (ρ) to determine correlation between judging time and number of sentences/words in the answer and its location in the webpage. Pearson’s correlation between judging time and number of sentences for *Easy* and *Difficult* documents is 0.15 ($p\text{-val}<0.001$) and 0.08 respectively. Similarly, correlation between judging time and unique words in answer text of *Easy* and *Difficult* documents is 0.20 ($p\text{-val}<0.001$) and 0.07 respectively. The weaker correlation between judging time and the location/number of sentences in difficult documents indicates that location or length of the answer cannot clearly determine the effort required to find information in difficult documents.

Overall, we observe that the answers provided by annotators for difficult documents are shorter than those marked in easy documents. However, we did not find a significant difference in answer location between easy and difficult documents. A document can be represented using several features. It is important to investigate which features are correlated with findability labels and judging time. We mainly study relationship of document length and its readability⁹ in this section. We use the mean judging time of all assessors as each document is judged by three assessors. In the previous chapter, we also considered whether readability would be repre-

⁸We observed a similar relationship for the number of words in an answer.

⁹As defined in Equation 3.1 and Equation 3.2 respectively

sentative of effort per document. We briefly consider the correlation of findability judgments and judging time with readability to investigate whether these judgments are affected by readability of the document.

Distribution of the features (mean μ and standard deviation σ) and their Pearson's correlation (ρ) with judging time of all documents is given in Table 5.4. Pearson correlation between judging time and different document/summary level readability indices is also provided. Further division by majority findability label for document and summary specific features is given in Table 5.5 and Table 5.6 respectively. Statistically significant correlations are marked with *.

We observe that the document length (both number of words and sentences) in Table 5.4 has a significant correlation with its judging time. This is expected as longer documents would take more time to judge. Readability indices indicate that high-school level English language proficiency is required to judge these documents. However, we found all readability indices to have a very weak (zero to negative) correlation with judging time. This indicates that the readability of the document in this experiment does not impact judging time. We found that the number of outlinks in a document are negatively correlated with judging time, something observed before in [107, 157]. We found no correlation between percentage of images or tables in documents with judging time.

The summary of feature distribution and their correlation with judging time of documents with different findability grades in Table 5.5 and Table 5.6 highlights some key trends. Table 5.5 shows that difficult documents are indeed longer than documents where judges find no information (marked as *None*) or find information easily (marked as *Easy*).

The word distribution of *Difficult* is significantly different from *None* ($t(116)=-2.7$, $p\text{-val}=0.007$) and *Easy* ($t(285)=2.37$, $p\text{-val}=0.01$) documents respectively. Similarly, the sentence distribution of *Difficult* is significantly different from *None* ($t(116)=-2.15$, $p\text{-val}=0.03$) and *Easy* ($t(285)=2.15$, $p\text{-val}=0.03$) documents respectively. However, CLI and ARI of *Difficult* documents is only marginally higher (but not statistically different) than that of *Easy* documents (CLI= $t(285)=1.6$,

Feature	None		Difficult		Easy	
	μ (σ)	ρ	μ (σ)	ρ	μ (σ)	ρ
# words	667 (498)	0.28*	1123 (1019)	0.46*	749 (770)	0.27*
# sentences	107 (123.6)	0.57*	201.4 (256.4)	0.38*	132 (195.2)	0.24
LIX	34.8 (13.08)	-0.17	33.7 (6.4)	-0.01	32.8 (6.6)	0.06
ARI	13.7 (4.1)	-0.24	14.5 (4.1)	-0.11	14.3 (7.2)	-0.07
CLI	12.4 (7.2)	-0.14	13.2 (2.8)	-0.04	12.7 (2.7)	0.01
outlinks	0.33 (0.11)	-0.35*	0.31 (0.14)	-0.14	0.34 (0.12)	-0.03
images	0.10 (0.12)	0.06	0.06 (0.08)	-0.03	0.07 (0.08)	-0.03
tables	0.025 (0.05)	0.21	0.02 (0.03)	-0.08	0.022 (0.03)	-0.02

Table 5.5: Document features and judging time correlation for different labels

p-val=0.10), ARI=($t(285)=-0.38$, p-val=0.70)) or where judges could not find the required information (CLI=($t(116)=-1.37$, p-val=0.17), ARI=($t(116)=1.26$, p-val=0.20)) i.e. *None* documents. We did not observe any relationship between the number of outlinks, images or tables and judging time or findability grades. However, the distribution of images in *None* documents is statistically different ($t(249)=1.98$, p-val=0.04) from *Easy* documents.

Summary based features in Table 5.6 show that the readability of difficult document is the highest among all three groups. However, we observe a higher negative correlation (although not significant) between judging time and query-based summary readability for documents that were marked *None*. Summary CLI of *Difficult* documents was statistically different ($t(285)=2.22$, p-val=0.04) from *Easy* documents. Difficult documents have clearly more regions with query terms for a judge to read, since the number of words and sentences i.e. summary length in difficult documents is the highest amongst all labels. We also observe the highest correlation between judging time and *Difficult* document or summary length. The number of words in the query-based summary in *Difficult* documents is significantly different from *None* ($t(116)=-2.13$, p-val=0.03) documents and *Easy* ($t(285)=2.64$, p-val=0.004) documents computed using a two-tailed t-test. Overall, feature correlation with judging time clearly indicates that difficult documents are longer and take more time to judge than other documents.

Feature	None		Difficult		Easy	
	μ (σ)	ρ	μ (σ)	ρ	μ (σ)	ρ
# words	539.2 (100.6)	0.07	797.6 (165.4)	0.46*	582 (59.4)	0.23*
# sentences	69.3 (74.1)	0.12	106.5 (117.3)	0.38*	81.3 (105.6)	0.21*
LIX	33.5 (9.4)	-0.21	34.6 (6.6)	0	33.3 (7.2)	0.09
ARI	15.1 (5.3)	-0.22	16.2 (5.1)	-0.08	15.6 (8.3)	-0.05
CLI	13.0 (3.3)	-0.16	13.6 (2.8)	-0.03	12.9 (2.9)	0.05

Table 5.6: Summary features and judging time correlation for different labels

Label type	Annotation time		words per sec		char per sec	
	median	mean	median	mean	median	mean
ρ	0.36*	0.39*	0.31*	0.32*	0.29*	0.27*
κ	0.44	0.48	0.38	0.38	0.37	0.35

Table 5.7: Pearson’s Rho and Cohens Kappa

5.2.4 Inferring implicit labels from judging time

Previous work has investigated the impact of judging time on judging errors. For instance, [95] found that time to judge a document relative to other documents gives an indication of the difficulty of judging the document. However, they analyzed judging behavior and assigned labels for non-relevant and relevant documents. Instead, the focus of this chapter is to use judging time to distinguish between *low* and *high effort* documents. Carterette *et al.* [173] also investigated the impact of assessor errors on system rankings. They found low variance in assessor judging times. They found judges annotated non-relevant documents more quickly than relevant documents. This work is also different from ours in that all documents in this experiment are relevant.

Existing work only investigates judging time as an indicator of error *but not* effort. In previous sections we saw that the judges take more time to judge high effort documents. We now turn our attention to the conversion of judging time into effort labels. In the previous section, we observed that annotators took more time to judge documents where relevant information was ‘*difficult-to-find*’ compared to documents where they could locate relevant nuggets quickly. We build on this insight and explore the correlation between different ways of inferring effort labels and explicit effort labels obtained through our judging interface.

The effort label y_{fij} of document d_j with respect to query q_i is determined

using Equation 5.1, where \bar{t}_i is the mean (or median) judging time for the query q_i . This way, we can accommodate for the fact that different queries may require different amount of time to be judged as the derived effort labels would depend on the mean (or median) judging time per query.

$$y_{fij} = \begin{cases} 1 & \text{if } t_{ij} \leq \bar{t}_i \\ 0 & \text{if } t_{ij} > \bar{t}_i \end{cases} \quad (5.1)$$

Document length and an annotator's reading speed may also influence the judging time. Hence, we also explore different statistics related to reading speed i.e. words or characters read per second by the annotator to derive effort labels. If a document d_i with w_j words and c_j characters is judged in t_{ij} seconds for query q_j , we compute the reading speed as follows:

$$rs_{wij} = \frac{w_j}{t_{ij}} \quad rs_{cij} = \frac{c_j}{t_{ij}} \quad y_{fij} = \begin{cases} 0 & \text{if } rs_{wij} \leq r\bar{s}_j \\ 1 & \text{if } rs_{wij} > r\bar{s}_j \end{cases} \quad (5.2)$$

We derive the effort label y_{fij} of document d_i with respect to the mean (and median) reading speed $r\bar{s}_j$ for query q_j . We then analyze the usability of the inferred effort labels by comparing them with the explicit effort labels obtained through the crowdsourcing study described in Section 5.1. Since each document in our data is judged by at least three different judges, we use the majority vote as the actual effort label for each document. We then compute Pearson's correlation (ρ) and Cohen's Kappa κ between the explicit effort labels and the labels derived from different statistics of judging time to determine which statistic can be used to convert judging time to an effort label.

Cohen's kappa κ is commonly used to measure the agreement between different annotators. It was previously shown that Cohen's kappa between two random assessors is ≈ 0.4 , and ≈ 0.48 in the context of relevance assessments [149, 168, 174].

Table 5.7 shows the correlation between different statistics of implicit effort labels and manual effort labels. We observe the highest correlation between implicit

and explicit effort labels for *mean judging time*. Pearson's Rho (ρ) is 0.39 and Cohen's Kappa (κ) is 0.48. Even though our implicit labels do not come from a real assessor, the agreement between implicit labels and manual findability labels is comparable with the agreement rates reported in context of manual relevance assessments [149, 168, 174]. Thus, we use the mean judging time of the query to implicitly derive effort labels from judging time information in subsequent chapters.

We found that the average Kendall tau between time based ranking and label based ranking was 0.87 (± 0.13) which indicates high agreement between both ways of ordering documents for effort. We also found that the average change in DCG [58] (Equation 2.3) is 5.4% ($\pm 2.8\%$). This also indicates that judging time can be reliably used as a proxy to measure document-level effort. With judging time as a proxy for effort, we can use existing collections where both relevance grade of the document and its annotation time are present. In the next chapter we shall incorporate this finding into pairwise learning-to-rank models that leverage both relevance and time to optimize for effort without adversely affecting relevance.

Discussion and limitations

In this chapter, our aim was to determine if annotator, query, and document specific properties affect findability judgments. We gathered judgments using a crowdsourcing platform for two annotator specific properties: 1) language proficiency and 2) query topic expertise. We also analyzed the difference in findability judgments and annotation time for different types of information needs. Finally, we evaluated the relationship between different document properties and effort judgments. We asked the judges to highlight relevant information nuggets for a given search query, which were used to understand whether query-specific answer length or location affected findability judgments.

In our study, we found that 88.5% of the judges were fluent in English and that language proficiency did not have any correlation with effort judgments. However, this could be an outcome of the platform used to collect judgments and the procedure of judges selection. A potential follow-up study could be designed with

annotators from different regions (such as [162]) and language proficiency to better understand the relationship between language proficiency and findability judgments. Topic familiarity, on the other hand, may influence a judge's ability to locate relevant information in a webpage. We also found significant differences in annotation times of judges with different level of expertise in the query topic. Finally, the average time spent by the judges with *beginner* skills was the highest amongst all knowledge levels.

Users can issue different kinds of search queries [175] on a search engine. We specifically used *faceted* and *single* information needs in this study. We found that indeed the nature of information need (single or faceted) may influence findability judgments. We found no difference between judging time distribution of faceted and single queries. However, within faceted and single information needs, the judging time of documents where the relevant information was difficult to find (*Difficult*) was significantly higher than documents where it was either not found (*None*) or easier to find (*Easy*). This study could be expanded with several other kinds of information needs such as informational or ambiguous queries [175] in the future.

The location of relevant information nuggets in a webpage were not statistically different across findability judgments. We also observed a weak linear relationship between judging time and answer location in the webpage. This indicates that the location of query-specific information may not affect findability judgments. However, we observed that the annotators tend to highlight fewer sentences in *Difficult* documents which suggests that perhaps such pages force judges to focus their attention on small nuggets of information in the webpage for better understanding. An interesting extension of this experiment would be to manipulate answer location and length in *Difficult* webpages to check for differences in answer highlights. Finally, we saw that document length was positively correlated with judging time which is perhaps because longer documents take more time to judge. Finally, across all properties, we observed that judging time of *Difficult* documents was significantly higher than *Easy* documents, which is supported by prior work on task interleaving with word search puzzles [176] where participants invested significantly

more time in difficult tasks.

Overall, we found that our first hypothesis holds for certain properties i.e findability judgments are dependent on topic familiarity and the nature of the information need. Whereas, the length/location of relevant answer or readability of the document did not show a significant effect on the judgments. Our second hypothesis that judging time could be used as an indicator of effort was also supported by the judgments. This finding is crucial for large scale joint optimization of effort *and* relevance in retrieval systems proposed in the next chapter.

5.3 Conclusion

In this chapter, we investigated whether findability judgments are affected by factors associated with the annotator, query topic and judged document. We also performed an in-depth analysis of how effort judgments and annotation time vary with annotator's topical knowledge, query type and length/location of query specific answer in the document.

We analyzed each factor with respect to effort labels and judging time. Our analysis uncovers two key trends. First, effort judgments vary with annotators familiarity with query topic and the nature of underlying information need in the search query. We did not find any significant difference in length or location of relevant answer and readability of the document across different types of effort judgments. Finally, we observed that judges spend more time locating information in high effort or difficult documents as compared to low effort or easy documents.

Given these findings, we suggest that annotators be chosen carefully to obtain effort based judgments. Where manual judging of effort is difficult, we suggest use of judging time as an indicator of effort to perform ranking or evaluation. We use these findings in the next chapter to train rankers that optimize for relevance *and* effort.

Chapter 6

Incorporating effort in ranking

We have shown that besides relevance, effort also plays an important role in user satisfaction in Chapter 3. We also collected judgments to identify which factors may characterize effort when comparing two equally relevant documents in Chapter 4. Finally, in the previous chapter, we found that judging time can be used as an indicator of effort. However, it remains to be seen how judging time information can improve existing retrieval systems.

Existing IR systems are designed to optimize for document relevance. With the explosion in online content, there may be multiple *equally* relevant documents for a search query. However, systems trained to differentiate between relevant and non-relevant documents would not treat two *equally* relevant documents differently. Thus, we can use our work to distinguish between two equally relevant documents on the basis of effort.

As noted in Chapter 3, judges are not time bound to label documents for topical relevance. While a judge may patiently and thoroughly look for relevant content in a document, an impatient user may not spend as much time consuming it. In such a case, even if relevant (but high effort) documents are shown to the user, she may abandon such documents or abandon the query altogether. It is worth noting that existing implicit signals such as dwell time will *incorrectly imply* that such documents are *useful* for the user. This discrepancy can be addressed if document judging time is also used to train retrieval systems. Such systems would retrieve not only relevant but also low effort documents. Such systems would in turn address

the differences in evaluation caused by user effort.

Although there is sufficient evidence on the importance of incorporating effort [112, 113, 168] in retrieval, existing work falls short of addressing how effort could be incorporated into existing information retrieval systems. We attempt to address this shortcoming with this work and investigate the following hypothesis in this chapter:

Hypothesis 5. Existing learning-to-rank models can be optimized for both relevance *and* effort.

We rely on judging time as a proxy of ‘*effort*’ and propose multi-objective pairwise learning-to-rank models. We evaluate all these approaches on TREC Web data using relevance and time-biased gain metrics. The experiments show that significant fraction (30-50%) of top-10 search results (retrieved using existing ranking methods) consist of high effort documents, in turn, motivating the need of incorporating effort into ranking models. Our experiments indicate that relevance and effort can be effectively combined to tailor better search experience. Overall, our approach yields 25% reduction in high effort documents in top-10 results on TREC Web data. The remaining chapter is structured as follows. Section 6.1 gives an overview of the proposed approaches that incorporate both relevance and effort. We describe features, evaluation metrics and experimental results in Section 6.2 and Section 6.3 respectively. This is followed by conclusion in Section 6.4.

6.1 Effort aware ranking

Pairwise approaches such as RankSVM [177] and LambdaMart [170] have proved to be effective in optimizing rank biased metrics. Existing pairwise approaches, however, only account for relevance. For a fair comparison, we modify SVMRank and LambdaMart to account for both relevance and *effort*.

Given a set of queries Q , a collection of documents D , where each document is represented by a feature vector $x_i \in R^n$ where n indicates the dimension of the feature vector, the goal is to find a function $f \in \mathcal{F}$. The function $f \in \mathcal{F}$ scores each document d_i with respect to a query q_j . For this work, we assume that the relevance

of a document d_i with respect to a query q_j is denoted by y_{rij} . Similarly, its effort label is denoted by y_{fij} . We assume that the relevance labels are known a priori. In Section 6.2.2, we describe how to derive effort labels using the judging time y_{tij} of the document.

6.1.1 Effort aware SVMRank

SVMRank [177] (SVM_r), a pairwise max-margin approach is used to learn a function to rank documents for a query. Given a set of document labels $y_{ri} \in \mathcal{K}$ ¹, query $q_i \in Q$, document feature vectors $X \in R^n$, SVMRank learns a hyperplane that enforces ordering among relevant and non-relevant documents. Our aim is to incorporate effort into this loss function without affecting the performance on relevance. Thus, relevance is the primary factor to order documents and effort is the secondary criterion. We enforce the ordering based on relevance followed by an ordering based on effort. We capture this two-level ordering by modifying SVMRank (SVM_{rf}) to optimize for relevance *and* effort.

$$\begin{aligned}
\arg \min_{w, \xi_{i,j}} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i,j} \xi_{i,j} \quad \text{subject to} \\
& \forall \{(x_i, x_j, q_i = q_j) : y_{ri} > y_{rj} \in R\} : \\
& \forall \{(x_i, x_j, q_i = q_j) : y_{ri} = y_{rj} \text{ and } y_{fi} > y_{fj} \in R\} : \\
& w \cdot x_i \geq w \cdot x_j + 1 - \xi_{i,j} \\
& \forall \xi_{i,j} \geq 0
\end{aligned} \tag{6.1}$$

where (x_i, y_{ri}) and (x_j, y_{rj}) are n -dimensional feature vectors and relevance labels of documents d_i and d_j respectively. Similarly, y_{fi} and y_{fj} denote the effort label of documents d_i and d_j respectively.

The formulation in Equation 6.1 optimizes for effort when two documents have the *same relevance* grade. Given a pair of documents, we assume that a relevant document must always be ranked higher than an irrelevant document (first constraint) independent of its effort label. However, given two *equally relevant* documents, a ‘*low effort*’ document should always be ranked higher than a ‘*high effort*’ document

¹ \mathcal{K} = set of relevance labels

Type	Feature	Description
Query	QTerms	Query terms in title/URL
	QPos	Min/max/average query term position
	QTF	Min/max/average query term frequency
	QTF-idf	Min/max/average tf-idf of query terms
Document	DocCount	Number of char/words/sentences in document
	DocRead	ARI[154]/CLI[165]/LIX[114] of document
	DocQSent	#sentences with query terms
Summary	SumCount	Number of char/words/sentences in snippet
	SumRead	ARI[154]/CLI[165]/LIX[114] of snippet
	SumQSent	#sentences with query terms in snippet
Structure	HCount	Percentage of headings
	FCount	Percentage of bold/italics
	SCount	Percentage of table/list/images
	HPos	Min/max position of headings
Outlinks	DCount	Number of links with same (different) domains
	LinkPos	Min/max/average position of outlinks
	LinkQFreq	Number of outlinks with query terms
	LinkQPos	Avg/min/max position of outlinks with query terms

Table 6.1: Relevance and Findability based features

(second constraint). The objective function is only sensitive to differences in effort labels when relevance is the same.

6.1.2 Effort aware LambdaMart

LambdaMart [170], winner of Yahoo! learning-to-rank challenge [47] in 2010, optimizes non-smooth IR metrics using boosted regression trees. LambdaMart (**LMart**) is trained with λ -gradients from LambdaRank [13]. Given a ranked list of documents for a query, each document's scalar λ -gradient depends on its position in the list and on the positions of the other documents (that have different labels) in the sorted list. The gradient is a product of two factors (1) the cost for a pair of documents and (2) the NDCG (Equation 2.3) gained by swapping the pair i.e. $\Delta NDCG$. Formally, given two document scores s_i and s_j , λ -gradient is given in Equation 6.2, where C_{ij} is the cross entropy loss given by $C_{ij} = s_j - s_i + \log(1 + e^{s_i - s_j})$ and $o_{ij} = s_i - s_j$ is the difference in ranking scores.

$$\lambda_{ij} = s_{ij} \left| \Delta NDCG \frac{\delta C_{ij}}{\delta o_{ij}} \right| \quad (6.2)$$

$\Delta NDCG$ represents the gain obtained by swapping documents d_i and d_j , and $S_{ij} =$

$$\begin{cases} -1 & y_j \succ y_i \\ 1 & y_i \succ y_j \end{cases} \text{ where } y_i \text{ and } y_j \text{ are labels of documents } d_i \text{ and } d_j \text{ respectively.}$$

For each document, sum of all λ -gradients is computed using all pairs P in which it occurs.

We propose $LMart_{rf}$ such that λ -gradients can be modified to account for both relevance and effort. We compute the gradients for samples that have the same relevance grade and are incorrectly ordered for effort. Thus, our gradient update for each pair is as follows:

$$\lambda_{ij} = \begin{cases} S_{rij} \left| \Delta NDCG_r \frac{\delta C_{ij}}{\delta o_{ij}} \right| & \text{if } y_{ri} \neq y_{rj} \\ S_{eij} \left| \Delta NDCG_f \frac{\delta C_{ij}}{\delta o_{ij}} \right| & \text{if } y_{ri} = y_{rj}, y_{fi} \neq y_{fj} \end{cases} \quad (6.3)$$

where $\Delta NDCG_f$ and $\Delta NDCG_r$ denote the gain in effort based NDCG and relevance based NDCG respectively. Equation 6.3 incorporates effort in gradients depending on the relevance labels of the document pair. Here, we prioritize the optimization of relevance over effort.

6.2 Experimental setup

In this section we give a brief overview of the features, dataset and evaluation setup for the proposed methods.

6.2.1 Features

Each document is represented using several features. We rely on features proposed previously in Chapter 4 [168] in Section 4.4.1. We computed 64 features for each document and Table 6.1 contains the features divided in five categories.

While relevance oriented features are specific to query terms, effort based features focus on document structure and readability. Overall the features capture properties related to document structure and length, properties of snippet/tags containing query terms, text readability, and the presence of elements such as images, tables and lists which aid quick identification of information from a webpage.

6.2.2 Effort label generation

Existing retrieval systems are trained on labeled datasets, where the relevance of a document d_j with respect to a search query q_i is available. Relevance labels are either gathered from expert judges or inferred from user dwell time to generate a large number of training samples. Since large-scale manual labeling for effort would be costly and time-consuming, we generate effort labels from judging time.

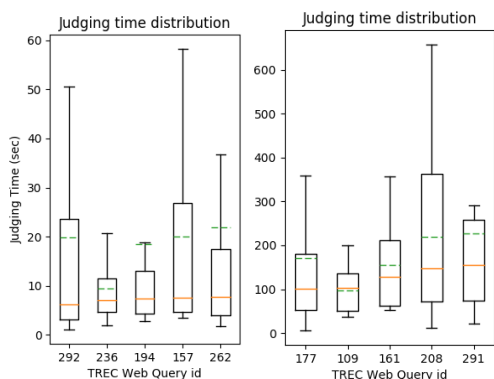
We posit that effort labels can be reliably inferred from collections built with the help of expert judges. Expert judges inspect a document thoroughly with respect to a search query. Thus, judging time represents an upper bound on time a user may spend on the page. Thus, for a query q_i , while ‘*low effort*’ documents could be judged quickly, a judge would take more time to judge ‘*high effort*’ documents thoroughly.

We generate effort labels from judging time as follows: For each query q_i , we use mean judging time q_{it} to determine effort label of each document d_j . Low effort documents have judging times lower than the mean, while high effort documents have judging times higher than the mean. This approach of deriving implicit labels for effort from judging time had the highest correlation with manual judgments of effort as described in Section 5.2.4. Thus, effort label y_{fij} of a document d_j with respect to a query q_i is generated using Equation 6.4, where \bar{t}_i is the mean judging time for the query q_i .

$$y_{fij} = \begin{cases} 0 & \text{if } t_{ij} \geq \bar{t}_i \\ 1 & \text{if } t_{ij} < \bar{t}_i \end{cases} \quad (6.4)$$

It is worth noting that judging time for some documents in the collection will not be available, as each document cannot be manually judged for relevance or effort with respect to a search query. In such cases, we posit that semi-supervised approaches can be used to deduce effort labels for unseen documents which is left for future work.

Year	#q	#docs	#rel	#high eff
2011	50	19381	3157	4692
2012	50	16102	3523	4938
2013	50	14531	4149	4288
2014	50	14610	5665	5338

Table 6.2: Query and label distribution of 2011-2014**Figure 6.1:** Judging time of low (left) and high (right) effort topics

Id	#d	Query
262	323	balding cure
236	173	symptoms of mad cow disease
194	375	designer dog breeds
157	329	the beatles rock band
292	368	electronic medical record history
177	249	best long term care insurance
109	377	mayo clinic jacksonville
161	255	furniture for small spaces
291	225	sangre de cristo mountains
208	249	doctor zhivago

Table 6.3: Low (high) effort queries

6.2.3 Datasets and evaluation metrics

We rely on the TREC Web collection (2011-2014) [140] for our experiments. We label each document for effort on the basis of NIST² judging time. We follow the labeling strategy described above. Table 6.2 shows the distribution of relevance and effort labels per year.

The judging time distribution of the top and bottom 5 queries (sorted on the basis of median judging time) is shown in Figure 6.1³ and corresponding queries are shown in Table 6.3. Figure 6.1 shows that judging time varies widely across queries which suggests that binary effort labels could be extended to incorporate judging time distribution directly. The table also shows that judging time is influenced by the nature of information need (faceted, ambiguous or informational) indicating that query dependent effort constraints could also be used to characterize effort.

We use 3 TREC Web datasets (150 queries)⁴ to train and one dataset (50 queries) to test its effectiveness. For example, results for 2014 TREC web queries

²http://ir.dcs.gla.ac.uk/test_collections/

³dashed line represents the mean judging time

⁴Further divided into 4:1 ratio for training and validation sets

are obtained by training on 2011-2013 TREC web queries as training data.

There are several metrics that can be used to evaluate the effectiveness of an IR system. In this work we rely on MAP (Equation 2.2), NDCG [58] (Equation 2.3) and Time Biased Gain (TBG) [16] to evaluate different systems. We evaluate each ranked list on two levels: 1) $NDCG$, MAP and TBG that evaluate retrieval quality based solely on relevance, and 2) $NDCG_{rf}$, MAP_{rf} and TBG_{rf} which incorporate both relevance and findability into evaluation.

To compute $NDCG_{rf}$, MAP_{rf} and TBG_{rf} we divide each relevance grade into two grades, one that corresponds to low-findability and the other that corresponds to high-findability (e.g., relevance grade 4 is mapped to grades 8 and 7 for ‘easy-to-find’ and ‘difficult-to-find’ documents in that relevance grade; relevance grade 3 is mapped to grades 6 and 5 for ‘easy-to-find’ and ‘difficult-to-find’ documents in that grade, etc). Since our primary criterion is still relevance, all non-relevant documents are mapped to grade 0 regardless of their findability labels. This enables a more refined evaluation of ranker performance.

We also report the average percentage of relevant documents where information is ‘difficult-to-find’ in top 20 results ($LF@20$) given in Equation 6.5 per year to gain better insight into ranker performance.

$$LF@k = \frac{\sum_{r=1}^k (\mathbb{I}y_r > 0) \cdot (\mathbb{I}y_f < 1)}{\sum_{r=1}^k \mathbb{I}y_r > 0} \quad (6.5)$$

Smucker et al. [16] proposed Time Biased Gain metric (TBG), which can be used to evaluate the quality of a ranking based on how much time a user has to spend on reading a document, as well as the relevance of retrieved documents. Since the Time Biased Gain metric both focuses on relevance and effort to find information (i.e., findability), we use TBG as another metric to evaluate the quality of our rankers.

$$TBG = \sum_{r=1}^{\infty} g(r) \exp(-T(r) \frac{\ln 2}{224}) \quad (6.6)$$

where the exponential factor is the time-based decay function and $T(r)$ is the esti-

mated time to reach rank r , computed as the time to read snippets plus the time to read clicked documents:

$$T(r) = \sum_{m=1}^{r-1} 4.4 + (0.018l_m + 7.8)P_{click}(m) \quad (6.7)$$

where l_m is the length of the document and $P_{click}(m)$ is the probability of click on the document at rank m . $P_{click}(m) = 0.64$ when document at rank m is relevant and $P_{click}(m) = 0.39$ otherwise as determined in [16]. TBG makes two assumptions, namely, that a user traverses the ranked list linearly and that the users reading speed is constant. The gain value was estimated to be $g(r) = 0.49$ for a relevant document and zero otherwise [16].

6.2.4 Baselines and systems summary

Several methods have been proposed to combine labels [178] or different rankers [14] to optimize for multiple criteria. We use a simple label aggregation method as baseline to map (relevance, findability) tuples to a hybrid label for training learning-to-rank models.

Linear aggregation

A simple mechanism is to map each (relevance, findability) tuple to a new grade as follows: $y_{rf} = 2 * y_r + y_f$. It transforms (relevance, findability) tuple to a grade $y_{rf} \in \{0, |r| * |f| - 1\}$, where $|r|$ and $|f|$ are number of relevance and findability grades respectively. We assume that a higher grade represents higher relevance and findability i.e. on 4-point Likert scale, label 4 would represent ‘*highly relevant*’ and ‘*easy-to-find*’ document respectively.

To summarize, we compare the following systems with rankers optimized only for relevance:

- *BM25*: A simple retrieval baseline that uses query terms to search a corpus of documents.
- *SVM_{r_{rel}}* and *LMart_{r_{rel}}*: Vanilla SVMRank⁵ and LambdaMart⁶ model trained

⁵Implementation at [://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html](http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html)

⁶Implementation at <https://sourceforge.net/p/lemur/wiki/RankLib/>

only on relevance.

- $SVMr_{lin}$ and $LMart_{lin}$: SVMRank and LambdaMart trained on a linear combination of (relevance, findability) labels.
- $SVMr_{rf}$ and $LMart_{rf}$: Models that account for both relevance and findability at time of training as described in Section 6.1.1 and Section 6.1.2 respectively.

6.3 Results and discussion

We evaluate all the models with metrics evaluated solely on relevance and jointly on relevance and effort. Table 6.4 shows the performance of the proposed methods and baselines on relevance. The best performing systems per year, metric and model type are highlighted. Statistical significance at different levels is computed with respect to rankers trained on relevance, i.e. $LMart_{rel}$ and $SVMr_{rel}$ using paired t-test respectively.

Relevance based evaluation

Our objective is to optimize the retrieval performance for findability without hurting relevance. Models trained on the linear combination of relevance and findability labels ($SVMr_{lin}$ and $LMart_{lin}$) show varied performance across different years and metrics. When evaluated with NDCG@20, both $SVMr_{lin}$ and $LMart_{lin}$ have the same performance as relevance based models $LMart_{rel}$ and $SVMr_{rel}$ in 2011-2012 but perform poorly in 2013-2014. However, these differences in performance of $LMart_{lin}$ over $LMart_{rel}$ in 2013 ($t(50)=1.46$, p-val=0.14) and 2014 ($t(50)=0.80$, p-val=0.42) were not significant. The difference in performance of $SVMr_{lin}$ over $SVMr_{rel}$ was not significant in 2013 ($t(50)=1.5$, p-val=0.12) but significant in 2014 ($t(50)=2.8$, p-val=0.02). In 2014, performance of $SVMr_{lin}$ on MAP@20 is also significantly lower ($t(50)=3.02$, p-val=0.01) than $SVMr_{rel}$ model.

Performance of $LMart_{lin}$ on $TBG@20$ shows that linear combination of relevance and findability labels yields little to no gain in time biased metrics over relevance based models. Linear model $LMart_{lin}$ performs only marginally better than $LMart_{rel}$ on $TBG@20$ in 2011 and 2014. However, $SVMr_{lin}$ shows 16% and 5% im-

Year	Metric	BM25	LMart			SVMr		
			<i>rel</i>	<i>lin</i>	<i>rf</i>	<i>rel</i>	<i>lin</i>	<i>rf</i>
2011	NDCG@20	0.27	0.28	0.28	0.28	0.29	0.30	0.30
	MAP@20	0.39	0.42	0.44	0.42	0.44	0.45	0.45
	TBG@20	0.38	0.35	0.36	0.40*	0.37	0.43*	0.43*
2012	NDCG@20	0.11	0.24	0.23	0.23	0.20	0.20	0.20
	MAP@20	0.22	0.47	0.43	0.46	0.40	0.42	0.41
	TBG@20	0.31	0.61	0.58	0.61	0.51	0.53	0.54*
2013	NDCG@20	0.27	0.38	0.36	0.38	0.39	0.37	0.39
	MAP@20	0.43	0.58	0.56	0.56	0.60	0.56	0.60
	TBG@20	0.43	0.55	0.54	0.57	0.56	0.54	0.59*
2014	NDCG@20	0.35	0.38	0.38	0.39	0.40	0.35*	0.40
	MAP@20	0.55	0.62	0.62	0.60	0.65	0.59*	0.65
	TBG@20	0.60	0.61	0.62	0.65*	0.62	0.59	0.62

⁷p-val: * ≤ 0.05 against *rel* baseline using paired t-test with bonferroni correction

Table 6.4: Relevance based evaluation of rankers for 2011-2014 Web Tracks

provement on *TBG@20* in 2011 ($t(50)=-3.4$, $p\text{-val}=0.005$) and 2012 respectively but performs poorly compared to relevance baseline $SVMr_{rel}$ in 2013-2014. Our experiments show that when relevance and findability labels are linearly combined, performance of $SVMr_{lin}$ and $LMart_{lin}$ is at par or worse than their relevance counterparts on most metrics. Overall, these experiments suggests that models trained on the linear combination of labels may sometimes hurt performance on relevance.

We proposed two pairwise learning-to-rank methods to incorporate both relevance and findability. We specifically incorporated findability labels in computation of λ -gradients in LambdaMart i.e. $LMart_{rf}$ and findability based constraints in SVMRank i.e. $SVMr_{rf}$. Performance of these models for MAP@20 and *NDCG@20* is at par with models trained on relevance i.e. *BM25*, $LMart_{rel}$ and $SVMr_{rel}$. It is interesting to note that $LMart_{rf}$ obtains lower (but not statistically significant) MAP@20 than $LMart_{rel}$ for 2011-2013 despite having same (or better in case of 2014) performance on *NDCG@20*. This is perhaps a result of how LambdaMart models are trained. In our implementation, we use *NDCG* to compute λ -gradients which leads to a better performance on *NDCG* over MAP.

Both models show improvements in *TBG@20* computed with relevance labels. $LMart_{rf}$ achieves significant improvements in *TBG@20* for 2011 ($t(50)=-2.5$, $p\text{-val}=0.04$) and 2014 ($t(50)=2.47$, $p\text{-val}=0.03$) respectively. We note that find-

ability based λ –gradient computation in $LMart_{rf}$ yields slight improvements over $LMart_{lin}$ model across all years. Time biased evaluation of $SVMr_{rf}$ shows statistically significant improvement over $SVMr_{rel}$ on three datasets. $SVMr_{rf}$ showed 16%, 3% and 5% improvements in $TBG@20$ in 2011-2013 respectively. However, we did not observe any increment in 2014 queries. Overall, our experiments indicate that the proposed reformulation of SVMRank can jointly optimize for relevance and findability better than LambdaMart reformulation when evaluated on metrics such as TBG and MAP .

Joint evaluation of relevance and findability

We also jointly evaluate the proposed models for relevance *and* findability. The objective is to determine whether these models perform better than the baselines if both relevance and findability labels are taken into consideration. Since we saw that the proposed models do not hurt relevance, we shall now test whether they show improvement when evaluated for findability. The performance of all the models on relevance *and* findability is shown in Table 6.5. We compute four metrics $NDCG_{rf}@20$, $MAP_{rf}@20$, $TBG_{rf}@20$ and $LF@20$ with labels derived from a linear combination of relevance and findability labels as described in Section 6.2.3.

Since the models trained to optimize for relevance do not account for the time it takes to find the required information from a webpage, our aim is to evaluate them on findability i.e. what percentage of top ranked *relevant* documents are labeled as ‘*difficult-to-find*’. $LF@20$ shows that models trained on relevance i.e. $BM25$, $LMart_{rel}$ and $SVMr_{rel}$ retrieve a significant number of documents in which users might struggle to find the relevant information which may hurt the user experience.

Our results indicate that models trained on relevance perform poorly on findability since $LF@20$ i.e. the average number of relevant ‘*difficult-to-find*’ documents is high, approximately 0.35~0.47 across 2011-2014. While, $LMart_{rf}$ showed no significant improvement in $LF@20$ except in 2014 ($t(50)=2.18$, p-val=0.03), $SVMr_{rf}$ obtained significant improvement in 2011 ($t(50)=2.47$, p-val=0.03) and 2013 ($t(50)=2.00$, p-val=0.04) respectively. Overall, linear combination models $LMart_{lin}$ and $SVMr_{lin}$ show smaller (but not significant) improvements in

Year	Metric	BM25	LMart			SVMr		
			<i>rel</i>	<i>lin</i>	<i>rf</i>	<i>rel</i>	<i>lin</i>	<i>rf</i>
2011	$NDCG_{rf}@20$	0.24	0.27	0.28	0.29	0.27	0.31*	0.33*
	$MAP_{rf}@20$	0.38	0.43	0.43	0.45	0.44	0.49*	0.50*
	$LF@20$	0.47	0.39	0.38	0.37	0.44	0.39	0.33*
	$TBG_{rf}@20$	0.40	0.42	0.43	0.48*	0.43	0.51*	0.52*
2012	$NDCG_{rf}@20$	0.12	0.25	0.23	0.25	0.20	0.20	0.21
	$MAP_{rf}@20$	0.22	0.47	0.43	0.47	0.40	0.41	0.42*
	$LF@20$	0.43	0.36	0.38	0.34	0.41	0.39	0.37
	$TBG_{rf}@20$	0.36	0.69	0.67	0.70	0.58	0.60	0.61
2013	$NDCG_{rf}@20$	0.27	0.39	0.36	0.38	0.38	0.38	0.40
	$MAP_{rf}@20$	0.43	0.58	0.56	0.58	0.60	0.56*	0.61
	$LF@20$	0.43	0.35	0.34	0.35	0.33	0.24*	0.26*
	$TBG_{rf}@20$	0.53	0.67	0.66	0.68	0.68	0.68	0.71*
2014	$NDCG_{rf}@20$	0.36	0.39	0.38	0.40	0.40	0.36*	0.40
	$MAP_{rf}@20$	0.55	0.62	0.62	0.62	0.65	0.58	0.65
	$LF@20$	0.44	0.38	0.33*	0.34*	0.39	0.35	0.32
	$TBG_{rf}@20$	0.70	0.75	0.75	0.78	0.74	0.72	0.75

⁸p-val: * ≤ 0.05 against *rel* baseline using paired t-test with bonferroni correction

Table 6.5: Joint relevance and effort based evaluation of rankers for 2011-2014

$LF@20$ across all years.

$LMart_{lin}$ shows no improvement in $NDCG_{rf}@20$ except in 2011. In fact, we observe a slight drop in $NDCG_{rf}@20$ for $LMart_{lin}$ in all years except 2011. $SVMr_{lin}$ obtains significant improvement of ($t(50)=-3.39$, $p\text{-val}=0.002$) in $NDCG_{rf}@20$ only in 2011. However, $SVMr_{lin}$ obtains significant increment ($t(50)=-3.39$, $p\text{-val}=0.005$) of 14% in 2011 and 10% of significant decrement ($t(50)=2.3$, $p\text{-val}=0.04$) in 2014 when compared to $SVMr_{rel}$ respectively. The lower performance of $SVMr_{lin}$ on joint metrics is caused by the poor performance ($NDCG@20=0.35$) on relevance metrics.

Clearly, any performance increment (or decrement) in relevance metrics translates into a good (or bad) performance on metrics computed with joint labels of relevance and findability. $LMart_{lin}$ achieves only slight improvement in $TBG_{rf}@20$ in 2011, while $SVMr_{lin}$ obtains significant improvement ($t(50)=-3.6$, $p\text{-val}=0.002$) in 2011 and a slight improvement in 2012 over $SVMr_{rel}$ model. Overall, both linear models $LMart_{lin}$ and $SVMr_{lin}$ do not yield high improvements in evaluation metrics computed using joint labels of relevance and findability.

The low performance of models based on linear combination of labels can be attributed to an increase in label space which in turn leads to label sparsity. Linear combination of relevance and findability increases the label space from $|r|$ to $|r| * |f| - 1$, which would make algorithms such as LambdaMart very sensitive to swapping documents at the time of λ -gradient computation. Linear combination of labels does not affect SVMRank as it does not rely on the *magnitude* of labels but on the relative order of the document pair to generate +1 or -1 at the time of training.

Despite similar performance as relevance baselines $SVMr_{rel}$ and $LMart_{rel}$ on metrics such as $NDCG@20$ and $MAP@20$, rankers with findability based constraints i.e. $SVMr_{rf}$ and $LMart_{rf}$ yield significant improvements in $LF@20$ and $TBG_{rf}@20$. $SVMr_{rf}$ shows improvements in $NDCG_{rf}@20$, $LF@20$ and $TBG_{rf}@20$ over the relevance based $SVMr_{rel}$ baseline. $LMart_{rf}$ achieves minor improvements in $NDCG_{rf}@20$, 7% and 2% over $LMart_{rel}$ for 2011 and 2014 queries respectively. Similarly, $SVMr_{rf}$ improves $NDCG_{rf}@20$ by 18% ($t(50)=-3.0$, $p\text{-val}=0.01$), 5% and 5% over $SVMr_{rel}$ for 2011-2013 respectively.

When evaluated using $TBG_{rf}@20$, $SVMr_{rf}$ gained 20% ($t(50)=-2.77$, $p\text{-val}=0.03$) in 2011, 5% in 2012 and 4% ($t(50)=-2.0$, $p\text{-val}=0.044$) in 2013 over $SVMr_{rel}$ in respectively. While $LMart_{rf}$ also obtained improvements in $TBG_{rf}@20$, they were not significant except in 2011 where it obtained significant improvement of 14% ($t(50)=-2.5$, $p\text{-val}=0.04$) over $LMart_{rel}$. Overall, we observed that findability based constraints in $SVMr_{rf}$ yield higher improvement than $LMart_{rf}$ on $NDCG_{rf}@20$, $LF@20$ and $TBG_{rf}@20$ for most datasets.

Joint evaluation of rankers for relevance and findability suggests that the models trained on linear combination of labels do not significantly outperform those trained only on relevance. However, the proposed models that incorporate both relevance and effort based constraints obtain highest improvements in $TBG_{rf}@20$ followed by $LF@20$ and $NDCG_{rf}@20$ over relevance baselines without hurting relevance based evaluation. These experiments reaffirm the benefit of jointly optimizing retrieval for relevance and effort.

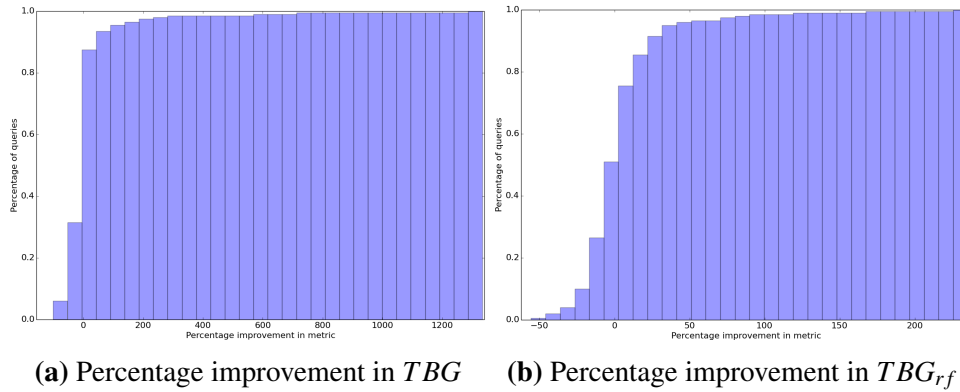


Figure 6.2: Percentage improvement of $LMart_{rf}$ over $LMart_{rel}$ for all queries

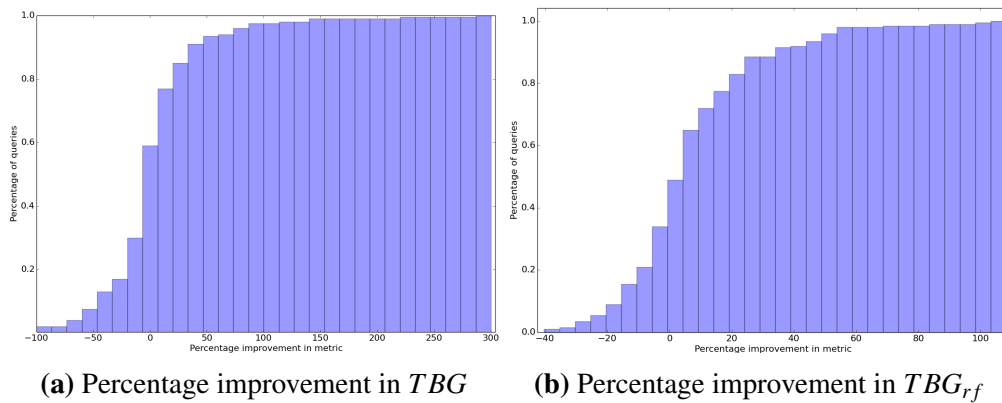


Figure 6.3: Percentage improvement of SVM_{rf} over SVM_{rel} for all queries

Query performance and feature importance

We now focus on the relative improvements in queries obtained by the proposed models. We specifically focus on improvements in $TBG@20$ and $TBG_{rf}@20$ to determine the fraction of queries and the scale of improvement over relevance baselines. Figure 6.2 shows the percentage improvements obtained by $LMart_{rf}$ in $TBG@20$ (Figure 6.2a) and $TBG_{rf}@20$ (Figure 6.2b) against $LMart_{rel}$ on x-axis along with the cumulative frequency of queries on y-axis for 2011-2014. $LMart_{rf}$ improves performance of 98 queries over $LMart_{rel}$ when evaluated by $TBG@20$. Similarly, it improves the performance of 111 queries over $LMart_{rel}$ when evaluated using $TBG_{rf}@20$. $LMart_{rf}$ gets an improvement of $> 50\%$ for 27 queries when evaluated using $TBG@20$ and 34 queries for $TBG_{rf}@20$ respectively.

Similarly, Figure 6.3 shows the improvements obtained by SVM_{rf} in

Type	$LMart_{rf}$ over $LMart_{rel}$		$SVMr_{rf}$ over $SVMr_{rel}$	
	$NDCG_{rf}@20$	$TBG@20$	$NDCG_{rf}@20$	$TBG@20$
Faceted	2.78%	15.01%	5.13%	16.27%
Single	1.39%	7.21%	4.9%	15.95%
Ambiguous	0.701%	7.56%	-1.84%	8.58%

Table 6.6: Gain of rel+find joint models per query type

$TBG@20$ (Figure 6.3a) and $TBG_{rf}@20$ (Figure 6.3b) respectively. $SVMr_{rf}$ improves the performance of 101 and 123 queries when evaluated by $TBG@20$ and $TBG_{rf}@20$ over $SVMr_{rel}$ baseline. $SVMr_{rf}$ shows an improvement of $> 50\%$ for 24 queries when evaluated using $TBG@20$ and 38 queries for $TBG_{rf}@20$ respectively. Overall, $SVMr_{rf}$ improves the performance of more queries than $LMart_{rf}$, however, the magnitude of improvements are larger in $LMart_{rf}$.

In Section 5.2.2, we analyzed how findability labels may differ across different types of information needs. Users may take longer to find relevant information for some queries (for example, locating weight of a product in the webpage) compared to others. TREC Web track queries consist of three types of information needs: faceted, single and ambiguous. We report the percentage improvement of $LMart_{rf}$ over $LMart_{rel}$ and $SVMr_{rf}$ over $SVMr_{rel}$ for all three information need types in Table 6.6 respectively.

Overall, both $LMart_{rf}$ and $SVMr_{rf}$ obtain the highest improvements for faceted queries. $LMart_{rf}$ achieves 15% and $SVMr_{rf}$ gets 16% improvement over relevance baseline for $TBG@20$. This suggests that faceted queries benefit from joint modeling of relevance and findability. While $LMart_{rf}$ gains only 7%, $SVMr_{rf}$ shows an improvement of 15% for topics with single information needs. Both models do not show significant improvements on ambiguous queries. Ambiguous queries are underspecified or vague information needs (or tail queries) which perhaps need more features or query specific parameters to train a model that jointly optimizes for both relevance and findability.

It is important to know which feature categories are most important for each model. We analyze whether relevance based models assign different weights to different set of features when compared to joint relevance and findability models.

Feature Group	$LMart_{rel}$	$LMart_{rf}$	$SVMr_{rel}$	$SVMr_{rf}$
Query	0.80	0.820	0.86	0.88
Outlinks	0.44	0.006	0.6	0.54
Summary	0.36	0.41	0.75	0.39
Readability	0.23	0.23	0.001	0.002
Document	0.19	0.15	0.15	0.07
Structure	0.001	0.39	0.17	0.34

Table 6.7: Feature weight determined using $NDCG_{rf}@20$

Since joint relevance and findability models also take into account time, we determine whether feature importance differs from relevance baselines. The weight of each feature indicating its importance is given in Table 6.7. For all models, features associated with the *query* are most important. For relevance based models, $LMart_{rel}$ and $SVMr_{rel}$ feature groups *summary* and *outlinks* are also important. On the other hand, for $LMart_{rf}$, features derived from *summary* and *structure* are also important and features associated with *structure* and *outlinks* are important for $SVMr_{rf}$ respectively. Our experiments with two learning-to-rank models showed that we can exploit judging time information and jointly optimize for relevance and findability. SVMRank based models yield more improvement than LambdaMart based models in ranking. We also found that joint models obtained highest improvements over relevance baselines for faceted queries but led to smaller gains on ambiguous queries.

Despite promising results, our work has some shortcomings. We posit that while judging time distribution may vary across datasets (some dataset might have different judging times), we would still observe a difference between judging time of ‘*difficult-to-find*’ and ‘*easy-to-find*’ documents in practice. Thus, it would be interesting to test these models on different datasets where judging time is available.

We also did not model the complexity of a webpage with embeddings or train neural network based rankers [179] to jointly score for relevance and findability in this work. However, we believe that findings of this work are generalizable and model independent, with more sophisticated models, it should only become easier to jointly model relevance and findability. In this work we focused only on pairwise rankers, however in the future findability constraints could also be integrated with

listwise rankers that optimize the overall time spent on task completion.

6.4 Conclusion

In the previous chapter, we showed that judging time could be a reliable proxy of effort required to locate relevant information from a webpage. This finding could be useful in building retrieval systems that jointly optimize for both relevance *and* effort. Another advantage is that judging time can be gathered at the time of relevance assessments at no extra cost which further reduces the overhead of collecting explicit effort judgments from judges. In this chapter, we investigated the hypothesis that learning-to-rank models can be used to optimize for relevance *and* effort. We proposed two pairwise learning-to-rank rankers that jointly optimized for relevance and effort with explicit labels of relevance and document judging time.

We proposed two effort aware pairwise learning-to-rank approaches and test their effectiveness with four years of TREC Web track queries. We evaluated the rankers using several metrics based on relevance and jointly for relevance *and* effort. We focused on NDCG, MAP and time-sensitive metric TBG. The proposed models outperform relevance baselines for two time biased metrics without hurting relevance. There was also a significant drop in '*difficult-to-find*' documents in the ranked list retrieved by effort aware rankers. Finally, we observed the highest improvement in performance for faceted queries i.e. queries that have multiple subtopics. Overall, our experiments indicate that indeed effort can be effectively combined with relevance by different means. We found effort aware formulation of SVMRank to be more effective than LambdaMart. Our proposed approach improved effort based NDCG by 33% over model trained only on relevance. In future, the proposed models can be tested with real users to measure their effectiveness.

Chapter 7

Search effort vs. satisfaction on mobiles devices

Search is an extremely popular means of finding information online. Users repeatedly interact with a search engine to satisfy their information needs which makes Interactive Information Retrieval (IIR) an active area of research. In the previous chapter, we investigated how effort can be incorporated in learning-to-rank models but did not elaborate on the effort required from an end user to search for information that requires her to issue multiple queries or consume multiple documents. To this end, in this chapter, we investigate the role of effort via formal models that encode actions (as costs) a user must perform and user's information gain while addressing an information need interactively.

Recently, some formal models [42, 43, 44] have been proposed that capture user cost (or effort) and benefit by incorporating several user actions. Users incur some cost for each of these actions: input a search query, read snippets, click results or scroll up/down the search engine result page (SERP). At present, cost of each action is measured in time, keystrokes or the number of documents clicked. For instance, query cost can be estimated via $W * c_w$ [42] where W is the number of words in a query and c_w is the average time it takes a user to type each word. Several models have been proposed [43], simulated [45] or empirically evaluated [44] on real datasets.

However, existing work only provides an *estimate* of user cost or benefit per

action, it does not explore how these costs or user effort are correlated with explicit labels of user satisfaction. It remains to be seen what cost functions correlate best with user satisfaction. Existing research in IIR is also limited to a desktop setting. User models of search and interaction have been developed for desktop environments and lab studies have been conducted to empirically evaluate and learn these models. However, today users have easy access to information on several devices such as desktops, mobiles, and tablets. Whether these models highly correlate with user satisfaction on different devices needs further investigation. Prior work [180, 134] has shown differences in how users interact with mobile search result pages. Thus, we posit that existing cost-benefit models may not be feasible for modeling search interactions on smaller devices. Therefore, in this chapter we investigate the following hypothesis:

Hypothesis 6. Existing cost-benefit analysis models designed for desktops cannot be directly used to model user behavior in mobile search

We begin by introducing a mobile search dataset collected during a lab study. We explore different actions and their costs across 25 users and 193 sessions. We also investigate how these cost functions correlate with explicit labels of user satisfaction provided by the participants of the study. Our experiments show that once trained, cost-benefit model parameters are different for desktop and mobile search. We also found that correlation of satisfaction with net benefit varied across different cost functions proposed in the literature. In the following sections, we briefly explain cost functions proposed in the literature, followed by examining the correlation between user satisfaction and search costs (or effort), benefit (or gain) and profit respectively as proposed in the previous work.

We begin with a brief overview of different cost-benefit analysis models in Section 7.1. We provide details of how the user study was instrumented in Section 7.2. Section 7.3 investigates two IIR models for mobile search. We summarize our findings in Section 7.3 and conclude in Section 7.4.

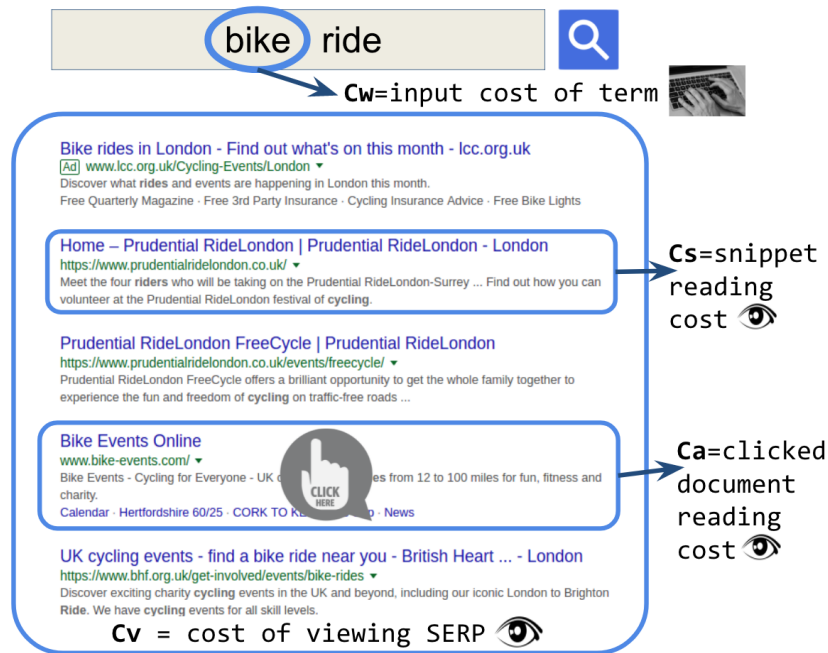


Figure 7.1: An example of different types of costs used in economic models of interaction.

7.1 Overview of economic models in IIR

Searching for information requires effort (or a cost) and for each user action, there is some gain (or reward) associated with it. We can use the principles of microeconomic theory to model IIR. If the search process is posed as an economic problem, we can further seek answers to questions such as, what search strategy (i.e. the combination of inputs) will minimize effort (i.e. user cost) for a given level of utility/gain (i.e. output) when using a particular retrieval system?

The models encode the utility or gain obtained from the relevant documents and the inputs to such models are: (i) total number of queries Q , (ii) length of the query L , and (iii) the depth of assessment per query D in a search result page. An example of all these interactions is shown in Figure 7.1. We can then devise a search production function $f(Q, D)$ which will quantify the maximum amount of Cumulative Gain that could be obtained if the user issued Q queries of length L and assessed D documents per query. In this section, we give a brief overview of two models that encode query interaction and SERP interactions. Several models have been proposed, we refer the reader to [42, 45, 43, 181] for in-depth overview. We chose the query interaction model in next section as it is the only model that

captures query specific cost and gain via economic models from the perspective of a user that has been evaluated on desktops. We chose the search interaction model in Section 7.1.2 as it was also evaluated for desktops.

7.1.1 Query interaction

We begin by describing how querying can be modeled using cost-benefit framework proposed by Azzopardi [42]. Assume that the user issues a query of length W and obtains a benefit determined by $b(W)$ and incurs a cost (or effort in querying) which is defined by the cost function $c(W)$. Since issuing multiple queries follows the law of diminishing returns [182, 183], we choose a benefit function such that they receive decreasing benefit as the length of the query increases. This is modeled with the function: $b(W) = k * \log_a(W + 1)$ where k represents a scaling factor and a determines how quickly the user experiences diminishing returns. Let us also assume that the cost of entering a query is a linear function based on the number of words such as: $c(W) = W * c_w$.

We can employ more complex cost functions. However, this model provides a simple abstraction. We can compute the profit (net benefit) π that the user receives for a query of length W as follows: $\pi = b(W) - c(W) = k * \log_a(W + 1) - W * c_w$. To find the query length that maximizes user's net benefit, we can differentiate and solve the equation which results in $W^* = \frac{k}{c_w * \log_a} - 1$.

This model has only been evaluated using simulations. However, in subsequent sections, we empirically determine whether this function correlates with explicit user satisfaction labels and what values of hyper-parameters are suitable for mobile search.

7.1.2 SERP interaction

To model user interaction at the SERP level, we would have to consider costs and rewards of more actions. Let us assume that a user issues Q queries, examines V search result pages per query, inspects S snippets per query and with probability p_a assesses A documents per query during a session. Each interaction has an associated cost where c_q is the cost of a query, c_v is the cost of viewing a page, c_s is the cost of

inspecting a snippet and c_a is the cost of assessing a document as shown in Figure 7.1.

In this work, costs are defined in terms of time. A linear cost function such as $c(Q, V, S, A) = c_q * Q + c_v * V * Q + c_s * S * Q + c_a * A * Q$ can be used to model search costs. To reduce the number of unknown parameters in the model, we use the average number of pages examined per query (v) in place of V . If we let the probability of assessing a document given the snippet be p_a , then the expected number of assessments viewed per query would be $A = S * p_a$. The benefit or gain function is defined as follows: $b(Q, A) = k * Q^\alpha * A^\beta$.

By taking partial derivatives and then solving for A and Q , we obtain the following expressions for the optimal number of assessments per query A^* : $\frac{\beta(c_q + c_v * v)}{(\alpha + \beta)(\frac{c_s}{p_a} + c_a)}$.

The above formulation has been empirically evaluated for desktop in [45] but not for mobile search. Different user actions have been previously [92, 110, 19] considered as a proxy for effort. However, to the best of our knowledge cost models that use user actions have not been applied to understand the net gain for a user in mobile. Hence, in this chapter we focus on understanding user effort required in issuing queries and overall session interaction in mobile search.

7.2 User Study and data statistics

The primary objective of this study is to understand how explicit labels of user satisfaction correlate with different costs-benefit models in mobile. We designed a study where users were asked to perform 10 search tasks. We engineered the search result pages (SERPs) using the results of a popular commercial search engine and also instrumented the SERPs to show image panels, videos and wiki results. The objective was to instrument the search results shown by existing search engines. While news is a fairly common result in desktop search, users have access to dedicated news applications on their phones that serve news content which reduces the utility of news results on mobile SERPs. We begin by explaining the parameters we control and variables we observe in this study. We also explain the application interface and

Topic Query	Topic Description
BMW C1 motorcycle	Your friend has just bought a new BMW motorcycle. You want to know what a BMW C1 motorcycle looks like and information about its price and mileage.
Blonde jokes	You are having a conversation with friends about blonde jokes. You are looking for some good blonde jokes.
Kim Kardashian	You just saw a news item about Kim Kardashian. You want to know everything and latest gossip about Kim Kardashian.
Inertia of sphere	Your younger sibling just asked you about how to calculate inertia of square. You want to know how to calculate it.
Varese ionisation	You heard this song play in a caf and would like more about it. You to see/hear it again online and find out about its lyrics.
Kobe Bryant	You overheard a conversation about Kobe Bryant being a great basketball player. You want to find more information about him.
Long beach California	You are planning a vacation around Long beach California. You are especially interested in knowing which cities, surrounding Long Beach, are worth paying a visit.
Bachelor party ideas	You just had an idea of throwing bachelor party for a friend. You want to find inspiration on some fun activities you could use.
Dewar Flask	People at work are talking about Dewar flask. You do not know what it is. You would like to know how a Dewar flask works.
Xmen sequel	You heard news about x-men sequel. You want to find out if there is going to come a sequel to the X-Men film series, and if so, when it can be expected.

Table 7.1: Topic descriptions

how we generated different SERPs.

7.2.1 Search topics

Since this is a small scale study, we sampled 10 search topics from publicly available FedWeb greatest hits collection [184]¹ to study user interaction with mobile SERPs. We selected ten topics and modified description of each topic for mobile search. The topic queries and corresponding descriptions are given in Table 7.1.

7.2.2 SERP population and presentation

We used the Bing Search API² with fixed parameters to search results for each user query. We only fetched results from English speaking market of United states and

¹<https://fedwebgh.intec.ugent.be/>

²<http://www.bing.com/toolbox/bingsearchapi>

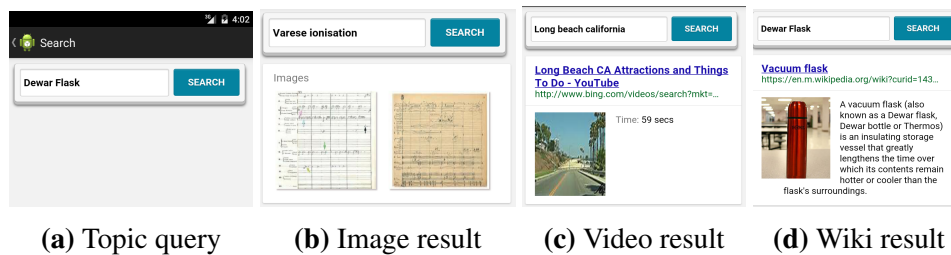


Figure 7.2: Search result page samples

filtered adult content from the results. We pre-fetched and cached search results of every topic query shown in Table 7.1. Since we stored search results for every user query, if two users were to issue lexically similar queries for a topic, they were shown the same results. This is to ensure that different users are shown exactly the same results for the same queries.

We customized different layouts for each vertical in line with existing search engines. While images are shown in a horizontally scrollable panel, wiki results are shown with an image along with first two sentences of the Wikipedia page. Video results are shown with an image and duration of the video. An example of all three layouts is shown in Figure 7.2. For each task and a user, a vertical is uniformly selected to be shown (or not) on the first position. Once a vertical was selected for user's search topic, the first result of subsequent queries for that search topic were of the same vertical.

7.2.3 App interface

We built an android app³ for our experimental study. The app interface is shown in Fig 7.3a. Each participant could perform as many search tasks as she liked. Participants were asked to complete post-task feedback once they felt they had found enough information regarding the search topic.

Each participant was required to register with a unique id. This was followed by a screen with tasks list in Figure 7.3a. The participants could begin with any task of their choice. Selection of a task led them to the task description window whose example is shown in Figure 7.3b. Participants could execute the task and provide

³Topics, results and app are available at <http://www0.cs.ucl.ac.uk/staff/M.Verma/app.html>

task feedback once they had finished the search task. The execute button would lead them to a screen with a search box as shown in Figure 7.2a. Participants were shown a pre-determined sample topic query (from Table 7.1) at the beginning of each search task. Participants could either use that topic query or issue a new query.

On returning to the SERP from a clicked page, participants are asked to provide page feedback, as shown in Figure 7.3c, on a Likert scale of 1 (non-relevant/not-satisfied) to 5 (highly-relevant/highly-satisfied), both for page relevance and their satisfaction. Thus, for all the experiments in this chapter, we obtained explicit labels of relevance and satisfaction from the participants in real time. They had the option to cancel, in case they did not want to provide any feedback. We also asked them to provide relevance and satisfaction labels for the *entire* SERP.

7.2.4 Participants

Participants were recruited via several mediums. We recruited some participants via university mailing lists and some via social media websites. Overall we recruited 25 participants (7 females and 18 males) for this study whose ages were between 22-55. We ensured that participants owned an android phone and were familiar with searching for information on mobiles. We briefed each participant about different screens in the app with one sample task. We also asked them to perform one sample task to become familiar with the app. We did not impose any time restrictions on any task so as to collect interaction data in the natural setting.

7.2.5 Observed variables

We track the following information for our analysis:

- SERP Behavior: We track user's behavior on the SERP for each query.
- Page and Task Feedback: We provide users the option to provide explicit feedback for pages they visit as well as for the overall task. We explicitly ask them to assign relevance and satisfaction grades to SERP and visited pages. Participants could also submit task feedback on completion of the task. This is to compare explicit feedback for each topic with implicit search behavior.

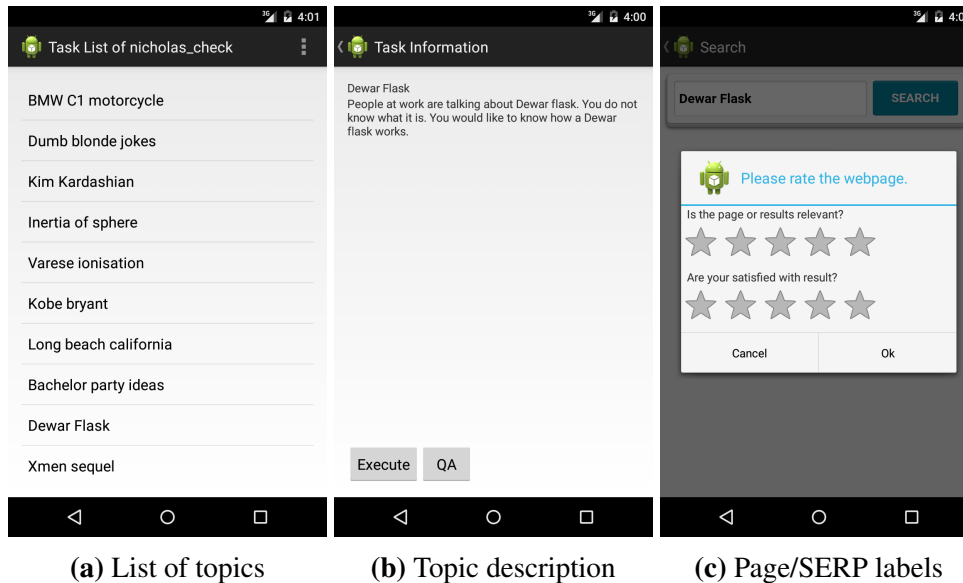


Figure 7.3: Topic layout and feedback screens

Search interactions on mobiles can be logged with the help of several events. In particular, we log user taps, pinches, query reformulations and dwell time on each page. We also record swipe (or pan) actions in four directions: up and down (for SERP scrolling) and left or right (for image panel interaction). Finally, we also record what items were visible on SERP during the search session.

We collected data for 10 tasks spanning 193 search sessions, 104 unique queries issued and 161 unique SERP result (URL) clicks. In total, we received 221 relevance/satisfaction labels for SERPs and 506 relevance/satisfaction labels for clicked URLs respectively. Finally, there were 205 responses for post-task survey. The distribution of SERP satisfaction labels is 1=13, 2=12, 3=32, 4=54, 5=81 respectively.

Aggregated statistics are shown in Table 7.2. It contains the average μ (and standard deviation σ) number of queries, clicks, page relevance and satisfaction ratings submitted by the participants. Since the post-task survey was optional for each topic, we consider only those sessions for analysis which had responses for post-task questionnaire. We chose to keep the first response for tasks that were executed twice by a participant.

sessions	queries	clicks	page rel	page sat	task sat
193	1.44 (0.92)	2.1 (2.16)	3.97 (1.2)	3.5 (1.42)	2.5 (0.58)

Table 7.2: Data summary

7.3 Cost/Benefit vs. satisfaction analysis

Cost (or effort) and benefit can be analyzed in multiple ways. Existing work [42] investigates user costs on a per-action basis. In this chapter, we limit our investigation to two types of costs: query cost and click/scroll cost. The cost of querying solely depends on user’s input query i.e. it is directly proportional to query length. However, click/scroll costs are relatively more complex as they depend on factors such as the number of snippets read, clicked and number of SERPs examined by the user. We explain different cost/benefit functions, discuss their correlation with SERP satisfaction labels from our study and finally estimate their parameters by optimizing different cost functions in the following subsections.

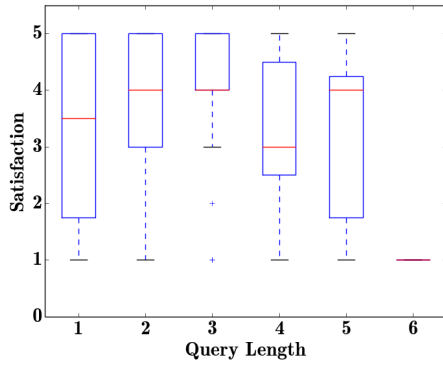
7.3.1 Query cost-benefit and user satisfaction

Users rely on keywords to formulate their information needs. They may incur different costs for issuing the query on different mediums. For instance, users can issue a query via keyboard or touch screens on desktop and mobile respectively. Users of our app were required to touch type their queries and we did not provide query auto-completion, to ensure that users explicitly type each query.

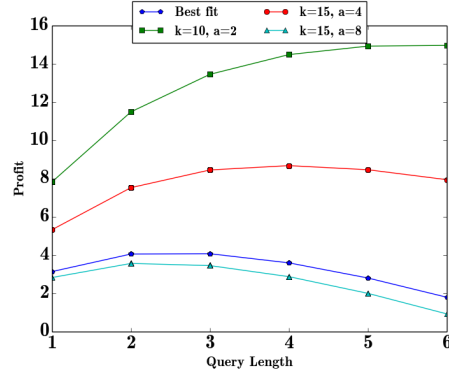
Given that a user enters a query with W words and c_w captures the effort required to input each word, we use the model from [42], in Equation 7.1, to compute net profit (π), benefit $b(W)$ and cost $c(W)$ for each query:

$$\begin{aligned}
 b(W) &= k \cdot \log_{\alpha}(W + 1) \\
 c(W) &= W \cdot c_w \\
 \pi(W) &= b(W) - c(W)
 \end{aligned}
 \tag{7.1}$$

Here, k represents a scaling factor and α captures diminishing returns of typing subsequent words. Distribution of satisfaction labels for queries of varying length is shown in Figure 7.4a. We use the same values for $k \in \{10, 15\}$ and $\alpha \in \{2, 4, 8\}$



(a) satisfaction vs. query length



(b) profit curves

Figure 7.4: Query length and user satisfaction (left) and query interaction net profit (right)

	k		
α	10	15	20
2	-0.10	-0.14	-0.15
4	0.312*	-0.009	-0.10
6	0.271*	0.27*	-0.02
8	0.256*	0.312*	-0.09
10	0.248*	0.295*	0.23

Table 7.3: Pearson's ρ between satisfaction & net query profit

	k			
β	2.0	5.0	10.0	16.0
0.03	0.16*	0.14*	0.10*	0.09
0.3	0.17*	0.13*	0.09*	0.08
0.43	0.16*	0.12	0.08	0.08
1.0	0.11	0.08	0.07	0.06

Table 7.4: Pearson's ρ between satisfaction & net search profit

as in [42] to compute Pearson correlation (ρ) between query profit and satisfaction. Correlation between satisfaction and profit for each combination of k and α is given in Table 7.3. We obtain values of c_w, k, α by optimizing the objective function in Equation 7.2 which minimizes the difference between user satisfaction ($\hat{\pi}$) and net user profit.

$$\min_{c_w, k, \alpha} \sum_{i=1}^n (\hat{\pi} - \pi(W))^2 \quad (7.2)$$

We can estimate the parameters c_w, k and α by minimizing the squared loss on satisfaction labels from our study. Parameter values $c_w = 2.18, k = 8.5$ and $\alpha = 3.0$ yield best fit on our data. When substituted, net profit has Pearson's ρ of 0.314 (p-val < 0.001) with satisfaction. Profit curves for different parameter settings are shown in Figure 7.4b. We observe that as the length of query increases, overall profit of user decreases which was also observed in [42].

We also observe a similar trend in our data where profit is highest for three

word queries and rapidly drops thereafter. Table 7.3 shows that higher α yields stronger correlation between satisfaction and user profit which indicates rapid diminishing returns of typing subsequent words. While query cost does not model entire search process, experiments on our data suggest that query costs (in Equation 7.2) can affect overall user satisfaction.

7.3.2 Search cost-benefit and user satisfaction

A user can choose from several actions once she submits any query to the search engine. She can either choose to examine a snippet, click a result, go to the search result next page or issue a new query. We assume that a user submits Q queries, reads S snippets, views V SERP pages per query and reads A clicked documents per query. If the cost of querying is c_w , the cost of viewing a SERP page is c_v , the cost of reading a snippet is c_s and the cost of reading a clicked document is c_a , we can use cost $c(Q, V, S, A)$ and gain/benefit $b(Q, A)$ function from [45] to compute the net profit π given in Equation 7.3. Here, α and β capture a user's frequency of issuing multiple queries and reading documents respectively.

$$\begin{aligned}
 c(Q, V, S, A) &= (c_w + c_v \cdot V + c_s \cdot S + c_a \cdot A) \cdot Q \\
 b(Q, A) &= k \cdot Q^\alpha \cdot A^\beta \\
 \pi &= b(Q, A) - c(Q, V, S, A)
 \end{aligned} \tag{7.3}$$

The distribution of satisfaction with respect to the time spent on reading (or examining) A clicked documents, viewing S snippets, cost of reading each snippet (c_s) and clicked document (c_a) is shown in Figure 7.5a, 7.5b, 7.6a and 7.6b respectively. Some users in our study, despite clicking on more than 10 documents for a query, have assigned higher satisfaction grades to SERP. It is worth noting that the median cost of reading a snippet (in milliseconds) is higher on low satisfaction SERPs than on high satisfaction SERPs. However, the trend reverses in the curve depicting examination cost of clicked documents i.e. Figure 7.5a where users spend less time reading a document clicked on low satisfaction SERP than on high satisfaction SERP.

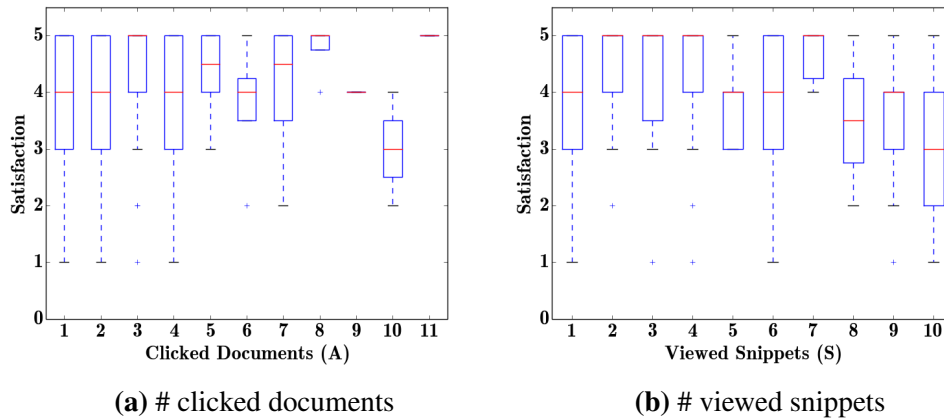


Figure 7.5: Clicked documents, viewed snippets and user satisfaction

We optimize the function in Equation 7.2 with satisfaction labels and net profit for each SERP. Since our satisfaction labels are gathered on a per-SERP basis, we set $Q=1$ to compute the cost and benefit function for each SERP. We perform the same optimization as shown in Equation 7.2 where we minimize the difference between satisfaction labels and net profit obtained from total SERP interaction. We obtained lower values for $k = 2.0$ and $\beta = 0.30$ than previously reported values $k = 5.3$ and $\beta = 0.43$ as given in [181]. The variation in profit curves for different combinations of k and β for clicked documents and viewed snippets is given in Figure 7.7a and Figure 7.7b, respectively. Pearson correlation (ρ) between net profit and satisfaction for different values of k and β is shown in Table 7.4⁴.

Best fit ($k = 2.0$ and $\beta = 0.30$) net profit curve in Figure 7.7a shows that *change* in net user gain is highest when only one document is clicked. Net profit gradually increases as more documents are clicked. The kink in curve for two clicked documents suggest that other costs (such as viewing snippets or issuing multiple queries) dominate the cost function, thereby lowering net profit. We did not observe a significant drop in the profit with increase in the number of clicked documents. However, net profit when $k = 5.3$ and $\beta = 0.43$ (from [181]) rapidly increases as more documents are clicked.

Our data suggests that a lower number of clicked documents yields higher user satisfaction on mobile. Profit curves for the number of viewed snippets in

⁴* indicates $p\text{-val} < 0.05$

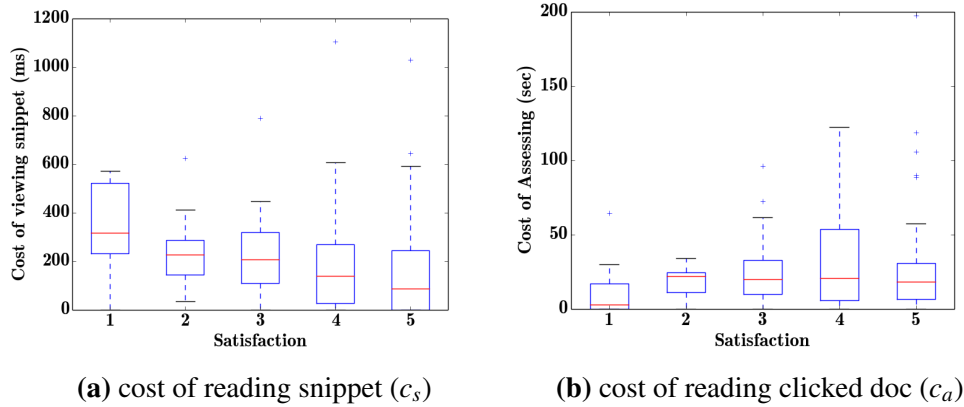


Figure 7.6: Cost of reading a snippet and clicked document

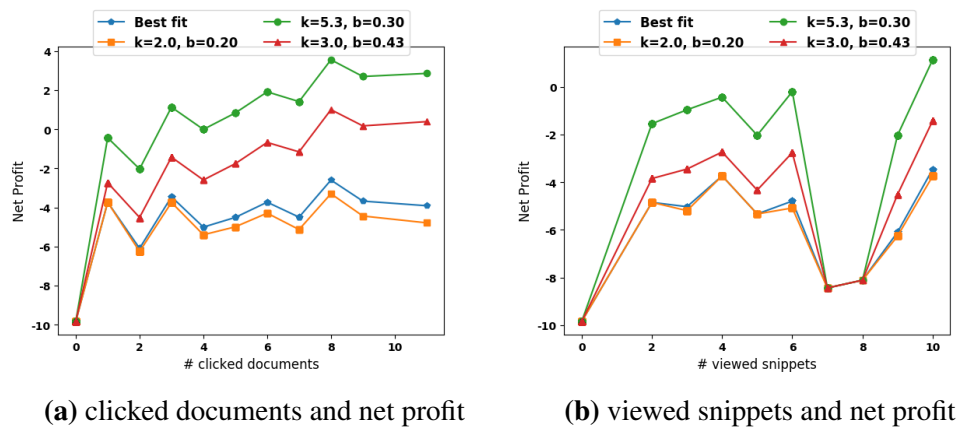


Figure 7.7: Net profit of reading a snippet and clicked document

Figure 7.7b shows a different trend. Net profit rapidly increases as users view more snippets but drops when they read between six to eight snippets. Best fit curve shows highest profit when a user views four snippets and declines thereafter. The best fit profit curve is similar to curve with $k = 5.3$ and $\beta = 0.43$ (from [181]) when plotted against viewed snippets. Table 7.4 shows that correlation between satisfaction and net profit weakens as k and β increase.

Pearson correlation (ρ) between satisfaction and net search profit on our data, for parameters obtained by optimizing objective function in Equation 7.3 ($k = 2.0$ and $\beta = 0.30$) was significantly low, only 0.17 ($p\text{-val} < 0.05$) which indicates that a linear combination of query, snippet examination and clicked document examination costs may not be optimal for mobile search. Pearson correlation (ρ) of each variable with satisfaction is as follows:

- $c_w * w = -0.33^*$
- $c_v = 0.03, v = -0.02, c_v.v = 0.03$
- $c_a = 0.07, A = 0.06, c_a.A = 0.09$
- $c_s = -0.13^*, S = -0.17^*, c_s.S = -0.16^*$

It is worth noting that each variable is correlated differently with satisfaction. While snippet (c_s) and query (c_w) costs are negatively correlated with satisfaction, the cost of examining clicked document (c_a) and search result pages (c_v) are positively (but not significantly) correlated with user satisfaction.

Overall, for both query and search cost-benefit functions, we observed a different optimal value for each parameter on mobile. We observed higher correlation between net query profit and satisfaction on mobile search data. However, the correlation with satisfaction and net search profit was relatively low, which suggests that linear combination of search costs may not be suitable for a mobile setting.

7.4 Conclusion

Existing models of cost-benefit analysis in IIR estimate how users maximize their net gain while minimizing search costs. These models do not provide any insight into how these strategies correlate with user satisfaction. Empirical study of these models is also limited to desktop setting. This chapter was an investigation of correlation between cost-benefit of querying/searching and user satisfaction in mobile search. We found that optimal parameters of these models differ in desktops and mobiles. We also found satisfaction to be highly correlated with net query profit but weakly correlated with net search profit. Our study motivates further investigation of non-linear cost models to better capture user behavior on mobile devices.

Chapter 8

Conclusion

Information retrieval systems have been traditionally designed to optimize for relevance. It is believed that showing more relevant documents to users would yield higher user satisfaction. One would expect that system evaluation based on relevance judgments would reflect the satisfaction of real users that interact with the search engine. However, it has been shown that batch evaluation does not always agree with user-based evaluation indicating that factors besides relevance that affect user satisfaction are absent from existing judgments.

We believe that the primary source for disagreement between batch evaluation and user-based online experiments is due to the disagreements between what judges consider as relevant versus the utility of a document to an actual user. In other words, existing relevance judgments and retrieval systems do not account for the '*effort*' an end user must put forth to locate and consume relevant information. Hence, in this thesis, we investigate the role of effort in information retrieval in greater depth. Primarily, we focus on what constitutes effort, the collection of effort judgments, how these judgments can be used to optimize IR systems and how effort varies across devices.

In Chapter 3, we investigated the extent of mismatch between explicit relevance judgments and implicit relevance judgments derived from dwell time information. Our objective was to empirically evaluate the extent of mismatch and the role of effort in explaining this mismatch between batch and online evaluation. Our hypothesis was that the existing relevance judgments do not account for *effort* and

that features associated with effort could be used to predict the difference in dwell time and judging time.

Empirical evaluation with three datasets showed that a fraction of documents exists which is labeled relevant by the judges but elicit very low dwell time from the users. We noted that such documents tend to have a very high judging time but very low dwell time. This suggests that users do not extract the same utility from the document as per judges expectation. To further test our hypothesis, we fit two regression models to predict the dwell time and the difference between dwell time and judging time with some effort related features. We focused mainly on document length, query-based snippets, and readability features to represent each document.

Our experiments indicate that users optimize for properties such as readability or length besides document relevance which is not considered by relevance assessors. Positive coefficients of readability features also indicate that the gap between judging time and dwell time would widen as query-based nuggets in the document become difficult to read i.e. users will spend far less time examining such documents than judges. Our first research question was to empirically evaluate the role of effort in explaining the mismatch between batch and online evaluation. Our experiments indeed showed that effort related features can explain the mismatch between judging time and dwell time.

In Chapter 3, we relied on features to approximate effort but did not have explicit labels for documents. Thus, in Chapter 4, we investigated whether it would be feasible to collect judgments associated with effort. Since several factors could be associated with effort, we focused on three potential factors, mainly *understandability*, *readability*, and *findability*. Analysis of explicit effort judgments and preference based judgments showed that the users prefer documents in which it is easy to locate relevant information. Thus, users distinguish two equally relevant documents on the basis of *findability* or ‘*ease of finding information*’ which addresses our second research question.

We also investigated the utility of several features in predicting both findability and relevance. Features such as the minimum position of query terms and the length

of query-based summary and number of images, lists or tables were useful in predicting *findability*. However, features such as document length, query terms in title and number of headings were important for predicting relevance. The differences in feature importance suggest that features associated with findability should also be used in building and optimizing the retrieval systems. Finally, we also showed that ranking of retrieval systems on the basis of findability *and* relevance is different from that generated solely on the basis of relevance. This further suggests the design of metrics that consider both relevance as well as effort to evaluate retrieval systems.

Effort judgments may, however, be affected by several annotator, query, and document specific properties. Thus, Chapter 5 investigates the relationship of some annotator, query, and document specific properties with findability judgments. Analysis of crowdsourced judgments and highlighted answers provided two insights. Our third research question was to investigate whether annotator, query or document specific properties affect effort judgments of relevant documents. We found that findability judgments vary with annotators prior knowledge of query topic and the nature of underlying information need in the search query. Finally, we found that judges spend more time locating information in high effort or difficult documents as compared to low effort or easy documents. We also found that rankings derived from explicit effort labels are similar to those generated from effort labels derived from the *mean judging time* of documents which answers our fourth research question.

Our fifth research question was to determine whether learning-to-rank models could jointly optimize for relevance and effort. To that end, we proposed two pairwise learning-to-rank models that account for both relevance *and* effort. Time and rank biased evaluation of these models on four years of TREC Web track queries showed that one can optimize for effort without compromising on relevance. The proposed models outperformed relevance baselines for two time-biased metrics without hurting relevance. There was also a significant drop in '*difficult-to-find*' documents in the ranked list retrieved by effort aware rankers. Finally, we

observed the highest improvement in performance for faceted queries i.e. queries that have multiple subtopics.

Effort becomes more critical when users access information via small devices like mobile. It has been shown [185, 135] that user satisfaction is significantly impacted by limited text input and touch interactions in mobiles. Cost-benefit models have also been previously [88, 42] used to study user behavior in information seeking tasks on desktop. However, the correlation of these models and their parameters with user satisfaction has not been investigated. There is also no prior investigation of how these models would perform on devices other than desktops such as mobiles. Thus, our final research question was to evaluate whether existing desktop-based cost-benefit analysis models empirically correlate with user satisfaction in mobile search. In Chapter 7, we showed that the optimal parameters of cost-benefit models in mobile search differ from desktop search. We also found that explicit labels of user satisfaction were highly correlated with the net query profit but weakly correlated with the net search profit. Our study motivates further investigation of non-linear cost models to better capture user behavior on mobile devices.

Future Work

This thesis aims to understand the role of effort in information retrieval. With the help of several crowdsourcing studies, we tried to determine characteristics associated with effort and their relationship with different annotator, query, and document specific properties. We also proposed models to incorporate effort into ranking and explored how existing cost-benefit models explain user interaction in mobile search. However, these problems constitute a small subset of the potential future research directions.

We conducted several studies to gather effort specific judgments from annotators but only conducted one user study on mobile in the last chapter. It would be interesting to conduct more user studies that study the effect of different parameters such as query difficulty, document length or readability on user effort. We could collect explicit judgments from the user such as those collected in [186] for

an in-depth understanding of effort in information seeking. In our studies, we did not exploit the interaction sequences [38, 131] of annotators to understand differences between easy and difficult documents. It would be interesting to mine event sequences that pertain to difficult (or easy) documents for early detection of user frustration for instance. We can also use these sequences to help annotators at the time of judging by asking them to judge documents that are suited to their expertise.

We investigated the effectiveness of only pairwise learning-to-rank models in this thesis. However, one could also explore listwise models or systems that optimize for session level effort by using submodular nature of effort. One can also use deep learning [187] methods to create embeddings or rankers that jointly optimize for relevance and effort. Finally, given the expense of gathering labels at scale, exploration of semi-supervised [188] approaches to rank for effort and relevance would be extremely useful in practise.

We also did not explore evaluation metrics that would account for effort and relevance. Recently, some metrics have been proposed to account for document-level effort [112]. However, more refined metrics could be designed to account for session or task oriented effort in the future. We explored the use of cost-benefit models in the last chapter. Our experiments showed that existing models cannot be directly applied to mobile search. In future, these models could be extended to incorporate user behavior on smaller devices. It would also be interesting to study information foraging models [84, 109, 83] in the context of mobile search in the future.

Finally, there are several applications that could potentially benefit from effort modeling. For instance, we can study the role of document level effort on prediction of task difficulty [189, 190]. We could also use effort specific information to generate document summaries that could quickly satisfy an information need. We could also build models that would translate difficult documents into easy documents for search queries. This would be extremely useful for children or people that suffer from autism. This can also be used to simplify legal or medical documents for non-technical people.

Bibliography

- [1] Evgeniy Gabrilovich, Shaul Markovitch, David C Howell, RA Fisher, James H Steiger, and Guang Y Zou. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, volume 12, pages 399–413. Cengage Learning.
- [2] Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st national conference on Artificial intelligence-Volume 2*, pages 1419–1424. AAAI Press, 2006.
- [3] Ian H Witten and David N Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. 2008.
- [4] N. J. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, (5):133–143, 1980.
- [5] Marcia J Bates. Information search tactics. *Journal of the American Society for information Science*, 30(4):205–214, 1979.
- [6] Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.
- [7] Linda Schamber and Michael Eisenberg. Relevance: The search for a definition. 1988.

- [8] Linda Schamber, Michael B. Eisenberg, and Michael S. Nilan. A re-examination of relevance: toward a dynamic, situational definition. *Information Processing & Management*, 26(6):755 – 776, 1990.
- [9] Stefano Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.
- [10] Pia Borlund. The concept of relevance in ir. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, 2003.
- [11] Tefko 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. *Journal of the American Society for Information Science and Technology*, 58(13), 2007.
- [12] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. Multi-dimensional 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.
- [13] Christopher J. Burges, Robert Ragno, and Quoc V. Le. Learning to rank with nonsmooth cost functions. In P. B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, 2007.
- [14] Qiang Wu, Christopher J. Burges, Krysta M. Svore, and Jianfeng Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3), June 2010.
- [15] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola Tonello, and Rossano Venturini. Quickscore: A fast algorithm to rank documents with additive ensembles of regression trees. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 73–82. ACM, 2015.

- [16] Mark D. Smucker and Charles L.A. Clarke. Time-based calibration of effectiveness measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12*, New York, NY, USA, 2012. ACM.
- [17] Nicola Ferro, Gianmaria Silvello, Heikki Keskustalo, Ari Pirkola, and Kalervo Järvelin. The twist measure for ir evaluation: Taking user's effort into account. *Journal of the Association for Information Science and Technology*, 2015.
- [18] Mark D Dunlop. Time, relevance and interaction modelling for information retrieval. In *ACM SIGIR Forum*, volume 31, pages 206–213. ACM, 1997.
- [19] Paavo Arvola, Jaana Kekäläinen, and Marko Junkkari. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13(5):460–484, Oct 2010.
- [20] Arjen P De Vries, Gabriella Kazai, and Mounia Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. *RIAO Conference Proceedings*, pages 463–473, 2004.
- [21] Carol L. Barry. User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3):149–159, 1994.
- [22] Peiling Wang and Marilyn Domas White. A cognitive model of document use during a research project. study ii. decisions at the reading and citing stages. *Journal of the American Society for Information Science*, 50(2):98–114, 1999.
- [23] Anastasios Tombros, Ian Ruthven, and Joemon M. Jose. How users assess web pages for information seeking. *Journal of the American Society for Information Science and Technology*, 56(4):327–344, 2005.

- [24] Yunjie (Calvin) Xu and Zhiwei Chen. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7), 2006.
- [25] Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2), 2007.
- [26] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 154–161, New York, NY, USA, 2005. ACM.
- [27] Georges Dupret and Benjamin Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of the Annual international ACM SIGIR conference on Research and development in information retrieval*, Singapore, 2008.
- [28] William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, and Daniel Olson. Do batch and user evaluations give the same results? In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 17–24, New York, NY, USA, 2000. ACM.
- [29] Andrew H. Turpin and William Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 225–231, New York, NY, USA, 2001. ACM.
- [30] James Allan, Ben Carterette, and Joshua Lewis. When will information retrieval be "good enough"? In *Proceedings of the 28th Annual International*

- ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 433–440, New York, NY, USA, 2005. ACM.
- [31] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 555–562, New York, NY, USA, 2010. ACM.
- [32] William Hersh, Andrew Turpin, Lynetta Sacherek, Daniel Olson, Susan Price, Benjamin Chan, and Dale Kraemer. Further analysis of whether batch and user evaluations give the same results with a question-answering task. In *Proceedings of the Ninth Text REtrieval Conference (TREC 9)*, pages 40–7.
- [33] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. The relationship between ir effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 773–774. ACM, 2007.
- [34] Falk Scholer and Andrew Turpin. *Metric and Relevance Mismatch in Retrieval Evaluation*, pages 50–62. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
- [35] Azzah Al-Maskari, Mark Sanderson, Paul Clough, and Eija Airio. The good and the bad system: Does the test collection predict users' effectiveness? In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, Singapore, 2008.
- [36] Mark D Smucker and Chandra Prakash Jethani. Human performance and retrieval precision revisited. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602. ACM, 2010.
- [37] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *Proceedings of the third*

- ACM international conference on Web search and data mining*, pages 221–230. ACM, 2010.
- [38] Qi Guo, Dmitry Lagun, and Eugene Agichtein. Predicting web search success with fine-grained interaction data. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2050–2054. ACM, 2012.
- [39] Jin Young Kim, Jaime Teevan, and Nick Craswell. Explicit in situ user feedback for web search results. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 829–832. ACM, 2016.
- [40] Michael D Cooper. A cost model for evaluating information retrieval systems. *Journal of the Association for Information Science and Technology*, 23(5):306–312, 1972.
- [41] Peter Pirolli. *Information foraging theory: Adaptive interaction with information*. Oxford University Press, 2007.
- [42] Leif Azzopardi and Guido Zuccon. An analysis of the cost and benefit of search interactions. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pages 59–68. ACM, 2016.
- [43] Leif Azzopardi and Guido Zuccon. Two scrolls or one click: A cost model for browsing search results. In *European Conference on Information Retrieval*. Springer International Publishing, 2016.
- [44] Leif Azzopardi, Diane Kelly, and Kathy Brennan. How query cost affects search behavior. In *Proceedings of the Annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013.
- [45] Leif Azzopardi. Modelling interaction with economic models of search. In *Proceedings of the 37th International ACM SIGIR Conference on Research*

- & *Development in Information Retrieval*, SIGIR '14, pages 3–12, New York, NY, USA, 2014. ACM.
- [46] Leif Azzopardi. The economics in interactive information retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 15–24. ACM, 2011.
- [47] O. Chapelle and Y. Chang. Yahoo! learning to rank challenge overview. In Olivier Chapelle, Yi Chang, and Tie-Yan Liu, editors, *Proceedings of the Learning to Rank Challenge*, volume 14 of *Proceedings of Machine Learning Research*, pages 1–24, Haifa, Israel, 25 Jun 2011. PMLR.
- [48] William S. Cooper. On selecting a measure of retrieval effectiveness, part i. *Journal of the Association for Information Science and Technology*, 24(2):87–100, 1973.
- [49] W.S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19 – 37, 1971.
- [50] Arthur Taylor. User relevance criteria choices and the information search process. *Information Processing and Management*, 48, 2012.
- [51] Diane Kelly, Xiao-jun Yuan, Nicholas J. Belkin, Vanessa Murdock, and W. Bruce Croft. Features of documents relevant to task- and fact- oriented questions. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages 645–647, New York, NY, USA, 2002. ACM.
- [52] Célia da Costa Pereira, Mauro Dragoni, and Gabriella Pasi. Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting. *Information processing & management*, 48(2), 2012.
- [53] Howard Greisdorf. Relevance thresholds: a multi-stage predictive model of how users evaluate information. In *Information Processing and Management*, pages 403–423, 2003.

- [54] Kira Radinsky, Fernando Diaz, Susan Dumais, Milad Shokouhi, Anlei Dong, and Yi Chang. Temporal web dynamics and its application to information retrieval. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 781–782. ACM, 2013.
- [55] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on World wide web*, pages 331–340. ACM, 2010.
- [56] Anlei Dong, Yi Chang, Zhaohui Zheng, Gilad Mishne, Jing Bai, Ruiqiang Zhang, Karolina Buchner, Ciya Liao, and Fernando Diaz. Towards recency ranking in web search. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 11–20. ACM, 2010.
- [57] Fernando Diaz. Integration of news content into web results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 182–191. ACM, 2009.
- [58] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48. ACM, 2000.
- [59] Falk Scholer, Milad Shokouhi, Bodo Billerbeck, and Andrew Turpin. Using clicks as implicit judgments: Expectations versus observations. In *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 28–39. Springer Berlin Heidelberg, 2008.
- [60] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, 23(2):147–168, April 2005.

- [61] Mark Sanderson et al. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval*, 4(4):247–375, 2010.
- [62] Martha Evens. Information retrieval experiment. *Computational Linguistics*, 11(4), 1985.
- [63] Georges Dupret and Ciya Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10*, pages 181–190, New York, NY, USA, 2010. ACM.
- [64] Diane Kelly and Nicholas J. Belkin. Display time as implicit feedback: Understanding task effects. In *Proceedings of the Annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, UK, 2004.
- [65] Georg Buscher, Ludger van Elst, and Andreas Dengel. Segment-level display time as implicit feedback: A comparison to eye tracking. In *Proceedings of the Annual international ACM SIGIR conference on Research and development in information retrieval*, Boston, USA, 2009.
- [66] Mark Claypool, Phong Le, Makoto Wased, and David Brown. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pages 33–40. ACM, 2001.
- [67] Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2):18–28, September 2003.
- [68] Masahiro Morita and Yoichi Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the Annual international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, 1994.

- [69] D. Sculley, Robert G. Malkin, Sugato Basu, and Roberto J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, France, 2009.
- [70] Kuansan Wang, Toby Walker, and Zijian Zheng. Pskip: Estimating relevance ranking quality from web search clickthrough data. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1355–1364, Paris, France, 2009.
- [71] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 193–202. ACM, 2014.
- [72] Qi Guo and Eugene Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the 21st international conference on World Wide Web*, pages 569–578. ACM, 2012.
- [73] Joeran Beel and Stefan Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In *International Conference on Theory and Practice of Digital Libraries*, pages 153–168. Springer, 2015.
- [74] Claudia Hauff, Diane Kelly, and Leif Azzopardi. A comparison of user and system query performance predictions. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, pages 979–988, New York, NY, USA, 2010. ACM.
- [75] Xiao Hu and Noriko Kando. User-centered measures vs. system effectiveness in finding similar songs. In *ISMIR*, pages 331–336, 2012.

- [76] Diane Kelly, Xin Fu, and Chirag Shah. Effects of rank and precision of search results on users evaluations of system performance. *University of North Carolina*, 2007.
- [77] Scott B. Huffman and Michael Hochster. How well does result relevance predict session satisfaction? In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 567–574, New York, NY, USA, 2007. ACM.
- [78] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 15–24, New York, NY, USA, 2017. ACM.
- [79] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Incorporating variability in user behavior into systems based evaluation. In *Proceedings of the ACM conference on Information and knowledge management*, Maui, USA, 2012.
- [80] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the ACM conference on Information and knowledge management*, Glasgow, UK, 2011.
- [81] Justin Kruger, Derrick Wirtz, Leaf Van Boven, and T William Altermatt. The effort heuristic. *Journal of Experimental Social Psychology*, 40(1):91–98, 2004.
- [82] Wai-Tat Fu and Peter Pirolli. Snif-act: A cognitive model of user navigation on the world wide web. *Human–Computer Interaction*, 22(4):355–412, 2007.
- [83] Peter Pirolli and Wai-Tat Fu. Snif-act: A model of information foraging on the world wide web. In *International Conference on User Modeling*, pages 45–54. Springer, 2003.

- [84] Peter Pirolli and Stuart Card. Information foraging. *Psychological review*, 106(4):643, 1999.
- [85] Eric L Charnov. Optimal foraging, the marginal value theorem. *Theoretical population biology*, 9(2):129–136, 1976.
- [86] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the Association for Information Science and Technology*, 59(7):1041–1052, 2008.
- [87] Olga Arkhipova and Lidia Grauer. Evaluating mobile web search performance by taking good abandonment into account. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1043–1046. ACM, 2014.
- [88] Leif Azzopardi and Guido Zuccon. An analysis of theories of search and search behavior. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 81–90. ACM, 2015.
- [89] Marilyn Hughes Blackmon, Muneo Kitajima, and Peter G Polson. Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 31–40. ACM, 2005.
- [90] Paul Thomas, Falk Scholer, and Alistair Moffat. What users do: The eyes have it. In *Asia Information Retrieval Symposium*, pages 416–427. Springer, 2013.
- [91] Zhiwei Guan and Edward Cutrell. An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 417–420. ACM, 2007.

- [92] Jacek Gwizdka and Irene Lopatovska. The role of subjective factors in the information search process. *Journal of the American Society for Information Science and Technology*, 60(12):2452–2464, 2009.
- [93] Robert Villa and Martin Halvey. Is relevance hard work?: evaluating the effort of making relevant assessments. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013.
- [94] Eero Sormunen. Liberal relevance criteria of trec -: Counting on negligible documents? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, New York, NY, USA, 2002. ACM.
- [95] Mark D Smucker and Chandra Prakash Jethani. Time to judge relevance as an indicator of assessor error. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1153–1154. ACM, 2012.
- [96] Timothy Jones, David Hawking, Paul Thomas, and Ramesh Sankaranarayana. Relative effect of spam and irrelevant documents on user interaction with search engines. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011.
- [97] Martin Halvey and Robert Villa. Evaluating the effort involved in relevance assessments for images. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14. ACM, 2014.
- [98] Praveen Chandar, William Webber, and Ben Carterette. Document features predicting assessor disagreement. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 745–748. ACM, 2013.

- [99] Duncan P Brumby and Andrew Howes. Strategies for guiding interactive search: An empirical investigation into the consequences of label relevance for assessment and selection. *Human-Computer Interaction*, 23(1):1–46, 2008.
- [100] Melanie Kellar, Carolyn Watters, Jack Duffy, and Michael Shepherd. Effect of task on time spent reading as an implicit measure of interest. *Proceedings of the Association for Information Science and Technology*, 41(1):168–175, 2004.
- [101] Kevyn Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135, 2014.
- [102] Jin Young Kim, Kevyn Collins-Thompson, Paul N Bennett, and Susan T Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 213–222. ACM, 2012.
- [103] Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Sebastian De La Chica, and David Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 403–412. ACM, 2011.
- [104] Kevyn Collins-Thompson. Enriching the web by modeling reading difficulty. In *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval, ESAIR '13*, New York, NY, USA, 2013. ACM.
- [105] Paul Kidwell, Guy Lebanon, and Kevyn Collins-Thompson. Statistical estimation of word acquisition with application to readability prediction. *Journal of the American Statistical Association*, 2011.

- [106] Azzah Al-Maskari and Mark Sanderson. The effect of user characteristics on search effectiveness in information retrieval. *Information Processing & Management*, 47(5), September 2011.
- [107] Diana DeStefano and Jo-Anne LeFevre. Cognitive load in hypertext reading: A review. *Computers in human behavior*, 23(3):1616–1641, 2007.
- [108] Mauro Mosconi, Marco Porta, and Alice Ravarelli. On-line newspapers and multimedia content: an eye tracking study. In *Proceedings of the 26th annual ACM international conference on Design of communication*, pages 55–64. ACM, 2008.
- [109] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4, 2005.
- [110] William S Cooper. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *Journal of the Association for Information Science and Technology*, 19(1):30–41, 1968.
- [111] Ben Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11. ACM, 2011.
- [112] Jiepu Jiang and James Allan. Adaptive effort for search evaluation metrics. In *European Conference on Information Retrieval*, pages 187–199. Springer, 2016.
- [113] Guido Zuccon. Understandability biased evaluation for information retrieval. In *European Conference on Information Retrieval*, pages 280–292. Springer, 2016.

- [114] Jonathan C. Brown and Maxine Eskenazi. Student, text and curriculum modeling for reader-specific document retrieval.
- [115] Na Dai, Milad Shokouhi, and Brian D Davison. Multi-objective optimization in learning to rank. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1241–1242. ACM, 2011.
- [116] Na Dai, Milad Shokouhi, and Brian D Davison. Learning to rank for freshness and relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 95–104. ACM, 2011.
- [117] Changsung Kang, Xuanhui Wang, Yi Chang, and Belle Tseng. Learning to rank with multi-aspect relevance for vertical search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12*, pages 453–462, New York, NY, USA, 2012. ACM.
- [118] Krysta M Svore, Maksims N Volkovs, and Christopher JC Burges. Learning to rank with multiple objective functions. In *Proceedings of the 20th international conference on World wide web*, pages 367–376. ACM, 2011.
- [119] Timothy Sohn, Kevin A Li, William G Griswold, and James D Hollan. A diary study of mobile information needs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 433–442. ACM, 2008.
- [120] Jeonghee Yi, Farzin Maghoul, and Jan Pedersen. Deciphering mobile search patterns: A study of yahoo! mobile search queries. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, pages 257–266, New York, NY, USA, 2008. ACM.
- [121] Karen Church and Nuria Oliver. Understanding mobile web and mobile search use in today's dynamic mobile landscape. In *Proceedings of the*

- 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 67–76, New York, NY, USA, 2011. ACM.
- [122] Karen Church, Barry Smyth, Keith Bradley, and Paul Cotter. A large scale study of european mobile search behaviour. In *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '08, pages 13–22, New York, NY, USA, 2008. ACM.
- [123] Karen Church, Barry Smyth, Paul Cotter, and Keith Bradley. Mobile information access: A study of emerging search behavior on the mobile internet. *ACM Transactions on the Web (TWEB)*, 1(1), May 2007.
- [124] Karen Church and Barry Smyth. Understanding the intent behind mobile information needs. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, pages 247–256, New York, NY, USA, 2009. ACM.
- [125] Yang Song, Hao Ma, Hongning Wang, and Kuansan Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1201–1212, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- [126] Maryam Kamvar and Shumeet Baluja. A large scale study of wireless search behavior: Google mobile search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 701–709, New York, NY, USA, 2006. ACM.
- [127] Maryam Kamvar and Shumeet Baluja. Deciphering trends in mobile search. *Computer*, 40(8):58–62, aug 2007.

- [128] Maryam Kamvar, Melanie Kellar, Rajan Patel, and Ya Xu. Computers and iphones and mobile phones, oh my!: A logs-based comparison of search users on different devices. In *Proceedings of the 18th International Conference on World Wide Web, WWW '09*, pages 801–810, New York, NY, USA, 2009. ACM.
- [129] Chad Tossell, Philip Kortum, Ahmad Rahmati, Clayton Shepard, and Lin Zhong. Characterizing web use on smartphones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, pages 2769–2778, New York, NY, USA, 2012. ACM.
- [130] Shuguang Han, Zhen Yue, and Daqing He. Understanding and supporting cross-device web search for exploratory tasks with mobile touch interactions. *ACM Transactions on Information Systems (TOIS)*, 33(4):16, 2015.
- [131] Fernando Diaz, Qi Guo, and Ryen W White. Search result prefetching using cursor movement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 609–618. ACM, 2016.
- [132] Dmitry Lagun, Mikhail Ageev, Qi Guo, and Eugene Agichtein. Discovering common motifs in cursor movement data for improving web search. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM '14*, pages 183–192, New York, NY, USA, 2014. ACM.
- [133] Jeff Huang, Ryen W White, and Susan Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1225–1234. ACM, 2011.
- [134] Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. Mining touch interaction data on mobile devices to predict web search result relevance. In *Proceedings of the 36th International ACM SIGIR Conference*

- on Research and Development in Information Retrieval*, SIGIR '13, pages 153–162, New York, NY, USA, 2013. ACM.
- [135] David Nicholas, David Clark, Ian Rowlands, and Hamid R. Jamali. Information on the go: A case study of europeana mobile users. *Journal of the American Society for Information Science and Technology*, 64(7):1311–1322, 2013.
- [136] Jane Li, Scott Huffman, and Akihito Tokuda. Good abandonment in mobile and pc internet search. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 43–50, New York, NY, USA, 2009. ACM.
- [137] Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabza. Detecting good abandonment in mobile search. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.
- [138] Milad Shokouhi, Rosie Jones, Umut Ozertem, Karthik Raghunathan, and Fernando Diaz. Mobile query reformulations. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1011–1014, New York, NY, USA, 2014. ACM.
- [139] George Buchanan, Sarah Farrant, Matt Jones, Harold Thimbleby, Gary Marsden, and Michael Pazzani. Improving mobile internet usability. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 673–680, New York, NY, USA, 2001. ACM.
- [140] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. Trec 2014 web track overview. Technical report, DTIC Document, 2015.

- [141] Omar Alonso and Stefano Mizzaro. Using crowdsourcing for trec relevance assessment. *Information processing & management*, 48(6):1053–1066, 2012.
- [142] Ian Ruthven, Mark Baillie, and David Elsweiler. The relative effects of knowledge, interest and confidence in assessing relevance. *Journal of Documentation*, 63(4):482–504, 2007.
- [143] Ben Carterette, Paul D. Clough, Evangelos Kanoulas, and Mark Sanderson. Report on the ecir 2011 workshop on information retrieval over query sessions. *SIGIR Forum*, 45(2):76–80, January 2012.
- [144] Georg Buscher, Ralf Biedert, Daniel Heinesch, and Andreas Dengel. Eye tracking analysis of preferred reading regions on the screen. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 3307–3312. ACM, 2010.
- [145] Jacek Gwizdka and Michael J Cole. Inferring cognitive states from multi-modal measures in information science. In *ICMI 2011 Workshop on Inferring Cognitive and Emotional States from Multimodal Measures (ICMI2011 MMCogEmS)(Alicante:)*, 2011.
- [146] Kasper Hornbæk and Erik Frøkjær. Reading patterns and usability in visualizations of electronic documents. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 10(2):119–149, 2003.
- [147] Gabriella Kazai, Nick Craswell, Emine Yilmaz, and S.M.M Tahaghoghi. An analysis of systematic judging errors in information retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*. ACM, 2012.
- [148] Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 205–214. ACM, 2011.

- [149] Tyler McDonnell, Matthew Lease, Tamer Elsayad, and Mucahid Kutlu. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, page 10, 2016.
- [150] Ryen W White and Diane Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 297–306. ACM, 2006.
- [151] William R Reader and Stephen J Payne. Allocating time across multiple texts: Sampling and satisficing. *Human–Computer Interaction*, 22(3):263–298, 2007.
- [152] Jacek Gwizdka and Michael J Cole. Towards human-information system interaction models derived from eye-tracking data. *Studia Ekonomiczne*, 158:65–80, 2013.
- [153] Georg Buscher, Andreas Dengel, and Ludger van Elst. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the Annual international ACM SIGIR conference on Research and development in information retrieval*, pages 387–394, Singapore, 2008. ACM.
- [154] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.
- [155] Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. Relevance and effort: An analysis of document utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14*. ACM, 2014.

- [156] Harald Weinreich, Hartmut Obendorf, Eelco Herder, and Matthias Mayer. Not quite the average: An empirical study of web use. *ACM Transactions on the Web (TWEB)*, 2(1):5:1–5:31, March 2008.
- [157] Oleg Rokhlenko, Nadav Golbandi, Ronny Lempel, and Limor Leibovich. Engagement-based user attention distribution on web article pages. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 196–201. ACM, 2013.
- [158] Jonathan Brown and Maxine 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.
- [159] Sandy JJ Gould, Anna L Cox, and Duncan P Brumby. Diminished control in crowdsourcing: an investigation of crowdworker multitasking behavior. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(3):19, 2016.
- [160] Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.
- [161] Omar Alonso and Stefano Mizzaro. Using crowdsourcing for {TREC} relevance assessment. *Information Processing and Management*, 48(6), 2012.
- [162] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval*, 16(2):138–178, 2013.
- [163] Philipp Schaer. Better than their reputation? on the reliability of relevance assessments with students. In Tiziana Catarci, Pamela Forner, Djoerd Hiemstra, Anselmo Peas, and Giuseppe Santucci, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, volume 7488 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012.

- [164] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or there: Preference judgments for relevance. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval, ECIR'08*. Springer-Verlag, 2008.
- [165] Meri Coleman and Ta Lin Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.
- [166] RJ Senter and EA Smith. Automated readability index. Technical report, DTIC Document, 1967.
- [167] Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000.
- [168] Manisha Verma, Emine Yilmaz, and Nick Craswell. On obtaining effort based judgements for information retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 277–286. ACM, 2016.
- [169] Kenneth A Kinney, Scott B Huffman, and Juting Zhai. How evaluator domain expertise affects search result relevance judgments. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 591–598. ACM, 2008.
- [170] Chris J.C. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical report, June 2010.
- [171] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.
- [172] Gabriella Kazai and Natasa Milic-Frayling. On the evaluation of the quality of relevance assessments collected through crowdsourcing. In *SIGIR 2009 Workshop on the Future of IR Evaluation*, 2009.

- [173] Ben Carterette and Ian Soboroff. The effect of assessor error on ir system evaluation. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 539–546. ACM, 2010.
- [174] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 2583–2586, New York, NY, USA, 2012. ACM.
- [175] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.
- [176] Stephen J Payne, Geoffrey B Duggan, and Hansjörg Neth. Discretionary task interleaving: heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology: General*, 136(3):370, 2007.
- [177] Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. Adapting ranking svm to document retrieval. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06*. ACM.
- [178] Lidan Wang, Jimmy Lin, and Donald Metzler. Learning to efficiently rank. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 138–145. ACM, 2010.
- [179] Bhaskar Mitra and Nick Craswell. Neural models for information retrieval. *arXiv preprint arXiv:1705.01509*, 2017.
- [180] Shuguang Han, I-Han Hsiao, and Denis Parra. A study of mobile information exploration with multi-touch interactions. In *SBP*, pages 269–276. Springer, 2014.

- [181] Leif Azzopardi. Economic models of search. In *Proceedings of the 18th Australasian Document Computing Symposium, ADCS '13*, pages 1–1, New York, NY, USA, 2013. ACM.
- [182] Leif Azzopardi. Query side evaluation: an empirical analysis of effectiveness and effort. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 556–563. ACM, 2009.
- [183] Nicholas J Belkin, Diane Kelly, G Kim, J-Y Kim, H-J Lee, Gheorghe Muresan, M-C Tang, X-J Yuan, and Colleen Cool. Query length in interactive information retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 205–212. ACM, 2003.
- [184] Ke Zhou, Thomas Demeester, Dong Nguyen, Djoerd Hiemstra, and Dolf Trieschnigg. Aligning vertical collection relevance with user intent. In *Proceedings of the ACM conference on Information and knowledge management*, pages 1915–1918. ACM, 2014.
- [185] Tomi Heimonen and Mika Käki. Mobile index: Supporting mobile web search with automatic result categories. In *Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services, MobileHCI '07*, pages 397–404, New York, NY, USA, 2007. ACM.
- [186] Jiepu Jiang, Daqing He, Diane Kelly, and James Allan. Understanding ephemeral state of relevance. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 137–146. ACM, 2017.
- [187] Aliaksei Severyn and Alessandro Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15*, New York, NY, USA, 2015. ACM.

- [188] Jason Weston, Frédéric Ratle, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, 2008.
- [189] Jaime Arguello. Predicting search task difficulty. In *European Conference on Information Retrieval*, pages 88–99. Springer, 2014.
- [190] Chang Liu, Jingjing Liu, and Nicholas J Belkin. Predicting search task difficulty at different search stages. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 569–578. ACM, 2014.