

A Study on Fine-Grained User Behavior Analysis in Web Search

Department of Social Informatics,
Graduate School of Informatics,
Kyoto University, Japan
Kazutoshi Umemoto

ABSTRACT

Web search engines are now used by people with diverse search intents for various purposes. Some tasks they conduct with the search engines cannot be fulfilled just by issuing a single search query or browsing a single document. There are several characteristics common to such *complex* search tasks. The first two are related to search intents: (1) even if the same query is issued to a search engine, the desired information may differ among searchers, since the search intents behind the query depend on their search contexts; and (2) searchers' intents can also shift from one to another during a series of their searches on a particular topic. The remaining two concern search outcomes: (3) as complex tasks require searchers to collect more than one relevant document, browsing a single relevant document does not necessarily indicate the success of their search sessions; and (4) this requirement also makes it difficult for searchers to decide when to stop searching. This thesis aims to establish methodologies for understanding diverse intents of searchers and for appreciating their search outcomes in order to help searchers complete complex tasks efficiently and effectively. Our key idea for achieving this aim is analyzing *fine-grained* user behavior data, where we use the phrase "user behavior" in a broader sense so that it can include search interaction data between searchers and search systems. More specifically, we address the following four research topics in this thesis:

1. Estimating Attractive Terms from Gazing Behavior

We address the problem of estimating hidden intents of Web search users from their observable search behavior. To this end, we focus on searchers' gazing behavior on their browsed Web pages and assume that terms drawing a high degree of the attention of a searcher are likely to be relevant to his/her topical intent. On the basis of this assumption, we propose a method of estimating a set of weighted terms that clarify the searcher's ambiguous topical intent. Through a laboratory experiment, the proposed method is evaluated from both qualitative and quantitative perspectives.

2. Predicting Query Reformulation from Current Search Behavior

We propose a method to detect the change of search scope in advance of the subsequent search. Following prior work that characterized searchers' query reformulation, we define our addressed task as a classification problem comprising five query reformulation types (*i.e.*, Generalization, Specialization, Parallel Move, Mission Change, and Error Correction). The proposed method utilizes various kinds of user interactions in the current search in order to predict the query reformulation type of the subsequent search. As well as evaluating the accuracy of the proposed method, we analyze which and to what extent each interaction feature contributes to the prediction.

3. Analyzing How Inconsistent Information Affects Search Behavior

We investigate how users handle inconsistent information by focusing on the effect of search expertise on their search behavior and satisfaction criteria. The research questions addressed in this work are: (1) how does the inconsistency of answers that users find relate to the degree of task satisfaction that they gain? (2) when do users complete search tasks where inconsistent answers exist? and (3) how do answers that users target for their searches change as time passes? To answer these questions, we select four tasks comprising factoid questions with inconsistent answers, extract answers found by study participants in these tasks, and analyze their answer-finding behavior in terms of the presence or absence of search expertise. On the basis of our findings, we discuss the design implications for making users without search expertise aware of the existence of inconsistent answers and helping them to search for supporting evidence for answers.

4. Visualizing the Amount of Missed Information in Search Results

We propose a query suggestion interface visualizing the amount of missed information for individual queries, with the goal of helping searchers discover unexplored aspects and find the appropriate timing for search stopping. We define the amount of missed information for a query as the additional gain that can be obtained from unclicked search results of the query, where gain is formalized as a set-wise metric based on aspect importance, aspect novelty, and per-aspect document relevance and is estimated by using a state-of-the-art algorithm for subtopic mining and search result diversification. We conducted a user study involving 24 participants to investigate the effect of showing missed information on users' search strategies and search outcomes. Experimental results showed that the proposed interface helped make users' search process more efficient, unless inaccurate estimates of missed information were visualized.

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my supervisor Professor Katsumi Tanaka. Since I joined Tanaka Laboratory, he always supported me with his brilliant insights into my research and his great breadth of mind.

I am deeply grateful to my thesis committee member Professor Toru Ishida, who was also my research advisor for five years. I was impressed with his comments about the abstraction of my research. I would like to thank my thesis committee member Professor Masatoshi Yoshikawa for his constructive suggestions on organizing my research. I really appreciate valuable comments from my research advisor Professor Hayato Yamana at Waseda University. He spent much time discussing my research when I visited his office.

I would like to show my deep appreciation to Dr. Ruihua Song at Microsoft Research Asia (MSRA). She was my mentor while I visited MSRA for six months. I received from her great care not only about our work at MSRA but also about my life at Beijing.

I would like to show my great appreciation to Associate Professor Satoshi Nakamura at Meiji University. I learned from him the fun of programming and the importance of presentation skills. Special thanks are given to Assistant Professor Takehiro Yamamoto for his numerous comments and suggestions. I always keep in mind his advice on how to improve the visual quality of research papers.

I am grateful to Professor Osami Kagawa at Osaka Gakuin University, Associate Professor Adam Jatowt, Associate Professor Hiroaki Ohshima, Lecturer Yusuke Yamamoto, and Assistant Professor Makoto P. Kato for their sharp comments and suggestions during research meetings.

I want to thank secretaries of Tanaka Laboratory, Ms. Ashiwa, Ms. Ikebe, Ms. Sato, and Ms. Shiraishi, for their kind help. I would also like to thank my current and former colleagues, especially Dr. Yoshiyuki Shoji and Dr. Kosetsu Tsukuda, and Mr. Tomohiro Manabe, for their cooperation and fruitful discussion.

Finally, I want to say thank you to my parents Takashi Umemoto and Seiko Umemoto for their strong support to my study for long years.

CONTENTS

1	Introduction	1
1.1	Background	1
1.1.1	Search Intents	2
1.1.2	Search Outcomes	3
1.2	Approaches	3
1.3	Thesis Structure	5
2	Related Work	7
2.1	Intent Mining	7
2.1.1	Direct Answers	7
2.1.2	Search Result Diversification	8
2.2	User Behavior Analysis	8
2.2.1	Search Satisfaction	8
2.2.2	User Factors	9
2.2.3	Search Stopping	9
2.3	Search Support Systems	10
2.3.1	Search Interfaces	10
2.3.2	Eye Tracking	10
3	Estimating Attractive Terms from Gazing Behavior	13
3.1	Introduction	13
3.2	Search Intent Model	14
3.3	Attractive Term Estimation	15
3.3.1	Method Outline	15
3.3.2	Weight Calculation for Gaze Terms	16
3.3.3	Gaze Term Aggregation Among Browsed Pages	17

3.4	Implementation	18
3.4.1	Preprocessing for Gaze Term Identification	19
3.4.2	Gaze Term Identification from Gaze Position Data	20
3.5	Experiment	20
3.5.1	Methods	20
3.5.2	Experimental Design	20
3.5.3	Accuracy of Estimated Terms	21
3.5.4	Uniqueness of Estimated Terms	23
3.6	Discussion	26
3.6.1	Attention to Visual Information	26
3.6.2	Eye Movement Patterns and Search Strategies	26
3.7	Summary	26
4	Predicting Query Reformulation from Current Search Behavior	29
4.1	Introduction	29
4.2	Problem Definition	31
4.2.1	Query Reformulation Type (QRT)	31
4.2.2	Prediction Task	32
4.3	Query Reformulation Type Prediction	32
4.3.1	Collection Procedure of Behavior Log Data	32
4.3.2	Labeling Behavior Log Data as QRT	32
4.3.3	Features	33
4.4	Evaluation	37
4.4.1	Overview of Collected Behavior Log Data	37
4.4.2	Classifier Setup	39
4.4.3	Classification by One-Versus-One Approach	39
4.4.4	Classification by One-Versus-Rest Approach	41
4.5	Discussion	44
4.5.1	Sufficient Number of Training Data	44
4.5.2	Influence of Preceding Search Behavior on Next Search	45
4.5.3	Predictability of Subsequent Query from Search Behavior	45
4.6	Summary	46
5	Analyzing How Inconsistent Information Affects Search Behavior	49
5.1	Introduction	49
5.2	Dataset	50

5.2.1	User Study	50
5.2.2	Task Selection	51
5.2.3	Answer Extraction	52
5.3	Analysis	53
5.3.1	How do Answers Relate to Satisfaction?	53
5.3.2	When do Users Finish Answer Searches?	55
5.3.3	How do Users Change Targeted Answers?	57
5.4	Discussion	59
5.4.1	Implications	59
5.4.2	Limitations	59
5.5	Summary	60
6	Visualizing the Amount of Missed Information in Search Results	61
6.1	Introduction	61
6.2	Proposed Interface	63
6.2.1	How It Works	63
6.2.2	Research Questions	64
6.3	Missed Information Estimation	65
6.3.1	Gain	65
6.3.2	Relation of Gain to Others	66
6.3.3	Missed Information	67
6.3.4	Gain Components	68
6.4	Experimental Design	69
6.4.1	Interfaces	70
6.4.2	Search System	71
6.4.3	Search Topics	71
6.4.4	Procedure	72
6.5	Results	73
6.5.1	Ground-Truth Data	74
6.5.2	Query Stopping	75
6.5.3	Query Selection	76
6.5.4	Gain Change	77
6.5.5	Session Stopping	78
6.5.6	Relationship Between Effort and Gain	79
6.5.7	Search Interaction	80

6.5.8	Questionnaires	81
6.6	Discussion	83
6.6.1	Implications	83
6.6.2	Limitations	84
6.7	Summary	85
7	Conclusions	87
7.1	Summary	87
7.2	Future Directions	89
	Bibliography	91
	Publications	103

LIST OF FIGURES

1.1	Search process model.	2
1.2	Overview of our approaches.	4
3.1	Our idea of attractive term estimation. We utilize searcher's gazing behavior during his/her browsing Web pages.	14
3.2	Example applications for attractive term estimation. Estimated terms enable search engines to re-rank search results and to suggest useful queries that reflect searcher's hidden topical intent.	15
3.3	Outline of our method to estimate attractive terms.	15
3.4	Screenshot of our prototype system for estimating attractive terms.	18
3.5	Interface used for annotating relevance of estimated attractive terms.	22
4.1	Our idea of query reformulation prediction. We leverage various kinds of user behavior observed in the current search to predict the type of query reformulation in the subsequent search.	30
4.2	Constructing training data from behavior log data.	33
4.3	Interface of our labeling system for constructing training log data.	33
5.1	Histograms for number of sessions with respect to time spent finding answers. . .	56
5.2	Changes in targeted answers over time.	58
6.1	Our idea of visualizing missed information.	62
6.2	How ScentBar works. The state of missed information after the initial search is shown in Figure 6.1.	63
6.3	Screenshot of ScentBar (topic: "global warming").	70
6.4	Mean change in oracle gain over time.	77
6.5	Relationship between effort and gain.	79

LIST OF TABLES

3.1	Tasks used in our experiment.	21
3.2	Results of nDCG@15 and MAP@15 for the ranking of terms estimated by each model. The best result is indicated by boldface for each task.	22
3.3	Mean unique rate (with standard deviation) of estimated terms among study participants. The best result is indicated by boldface for each task. Significant differences with the TF model at $p < 0.05$ and $p < 0.01$ are marked with “*” and “**”, respectively.	24
3.4	Actual examples of relevant term sets judged by study participants. Unique terms are shown in boldface.	25
4.1	Features used for constructing our QRT classifier.	36
4.2	Actual examples of search tasks and query reformulations included in the collected behavior log data. A symbol on an arrow means a QRT between two queries before and after the edge.	37
4.3	Distributions of QRTs in the collected behavior log data. The left side is about two kinds of distributions of our collected data: all the five QRTs (all-types) and the four QRTs except Mission Change (non-MC). The QRT distributions reported in [8] are shown on the right side. UK means the QRT distributions in Yahoo! UK, and US means those in Yahoo! US.	38
4.4	Classification accuracy for each method.	40
4.5	Useful/Useless features for QRT prediction with accuracy after ablation. A feature that has low/high accuracy makes a great/small contribution to the prediction.	40
4.6	MRR for the ranking of the predicted QRTs.	42

4.7	Ablated features for each QRT classifier. In this table, f denotes the feature name shown in Table 4.1, r denotes the ablation order, d denotes the difference between r and the mean ablation order. A feature having a low value for $r + d$ in QRT t makes a great contribution to classification of t . As for the features indicated in boldface, the reason why these worked well could be inferred.	44
4.8	Classification accuracy among different sampling sizes.	45
4.9	MAP for the ranking of predicted subsequent queries.	46
5.1	Task descriptions given to participants of user study.	51
5.2	Statistics on task answers found by study participants with answer samples. . . .	52
5.3	Features related to search effort and found answers. Value on the left of each cell denotes mean (with standard deviation) for feature values, and that on the right denotes Pearson's r (with p -value) between the feature and task satisfaction. . . .	54
6.1	Search topics (with oracle aspects) used in our user study (translated from Japanese).	72
6.2	Correlation coefficients between oracle and estimated gains. The top three values for each measure are shown in bold face.	75
6.3	Mean change (with SD) in oracle missed information at the query level. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.	76
6.4	Mean (with SD) of oracle missed information on query issuing. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.	77
6.5	Top: mean change (with SD) in oracle missed information at the session level. Bottom: mean (with SD) of oracle gain obtained through task sessions. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face. .	78
6.6	Results of search interaction analysis. The mean (with SD) values are reported. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.	81
6.7	Mean rating scores (with SD) of post-task and exit questionnaires in which five-point scale was used for each question. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.	82

INTRODUCTION

1.1 BACKGROUND

With the increase in the amount and quality of information on the Web, search engines are now used by people with diverse intents for various purposes. Searching is driven by information needs or search intents and starts with formulating queries. While search engine can observe issued queries, the real intents of searchers are hidden from them. A single intent can be represented in different queries. It is also possible that queries formulated by searchers do not precisely represent their search intents. Users will fail in achieving their information needs if search results returned by search engines are irrelevant to their search intents. Thus, providing appropriate contents within search engine results pages (SERPs) is one of the perpetual challenges to be addressed.

While search intents cannot be observed by search engine, users with search intents take various actions (*e.g.*, issuing queries, examining snippets, and browsing documents), which is observable. The model of search process is shown in Figure 1.1, where observable actions that searchers take are represented by rectangles, and unobservable decisions that they make are represented by ellipses. This model is inspired by recent work on interactive information retrieval [7, 63, 77, 95], while a similar model has been proposed by Marchionini [62] in the 1990s in the context of information seeking. To improve the retrieval performance of Web search engines, great effort has been devoted to understand search intents and appreciate search outcomes by analyzing users' search behavior (*e.g.*, [1, 8, 20, 47, 59, 70, 76]).

Searchers sometimes conduct *complex* tasks, where their information needs are not satisfied just by issuing a single search query or browsing a single document. Unlike simple search tasks

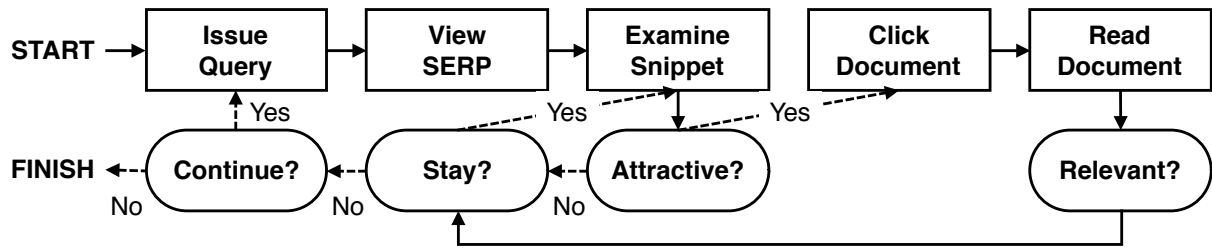


Figure 1.1: Search process model.

(e.g., navigational search [9], where searchers desire to visit a particular Web site), there are mainly four issues common to such complex search tasks. The first two concern search intents, while the remaining two concern search outcomes. We describe each of them in detail below.

1.1.1 Search Intents

We first describe two issues, *intent ambiguity* and *intent transitivity*, which concern topical search intents [51, 52, 72]. Topical search intents form the basis of users' information needs as they are search intents modeled by topics that users want to find through searches.

Intent Ambiguity

Information that searchers desire to obtain is sometimes ambiguous in the beginning of complex search tasks. This is mainly due to the fact that a large fraction of searches are conducted to obtain information in topics about which users do not know much. In such cases, users cannot formulate specific queries in the beginning. Their ambiguous topical intents are clarified through the search process. Therefore, even if exactly the same query is issued to a search engine, the desired information may differ among searchers. Let us consider the query “Kyoto history” as a concrete example. While some users may intend to survey the history of Kyoto in order to complete a report, others may issue this query for making a travel plan to visit historical places in Kyoto. To assist searchers with ambiguous topical intents, search engines need to grasp what kind of specific information users desire to search for.

Intent Transitivity

In complex search tasks, information of interest to searchers can shift from one to another during a series of their searches. Let us again take the query example “Kyoto travel”. A user issuing this query to a search engine may first search for sightseeing spots in Kyoto and then try to decide which hotel they will stay during the travel. Another searcher may use this query to find a good restaurant in the first search and to examine the means of transportation appropriate to their travel plan in the subsequent search. If search engines can detect the change of search scope in advance,

they can assist users in proactive manners (*e.g.*, suggesting queries that are likely to be used in future searches).

1.1.2 Search Outcomes

More than one document should be collected in complex search tasks. This requirement raises two issues concerning search outcomes, namely *search satisfaction* and *search stopping*.

Search Satisfaction

To be succinct, the goal of search engines is to satisfy searchers by providing appropriate search results. Various studies have been conducted to predict whether searchers get satisfaction from their search process [1, 35, 54, 81]. While these studies assume browsing relevant documents is a good indicator of search satisfaction, this assumption does not always hold, especially in complex search tasks where searchers encounter inconsistent information while browsing documents. Even if two documents are topically relevant, searchers may become dissatisfied when finding different answers from them. How do searchers act to complete such tasks with satisfaction in complex tasks? What factors do affect search outcomes in these tasks? These are important but unexplored research problems.

Search Stopping

Some complex tasks require searchers to collect exhaustive information covering a variety of aspects of search topics. In these tasks, when to stop searching is a very difficult decision for searchers. Quitting the tasks too early without in-depth exploration prevents searchers from finding essential information. On the other hand, if they have already obtained extensive information, continuing these tasks for too long wastes time and effort, as they will not be able to acquire much more gain. Techniques for evaluating the value of their collected documents and systems for guiding them in effective decision-making on search stopping should be developed in order to help users stop searching at the appropriate timing.

1.2 APPROACHES

Existing studies on search behavior analysis has focused on *coarse-grained* user behavior¹, *i.e.*, query- and/or document-level behavior, with the objective of mining the representative intents of general searchers and estimating the per-document relevance. However, analyzing such coarse-grained data is not sufficient in complex search tasks due to the above issues of these tasks. Given this background, understanding users' search behavior in complex tasks is one of the most

¹ We use the phrase "user behavior" in a broader sense throughout this thesis. The phrase may include not only actions that searchers take but also contents that they obtain from search systems.

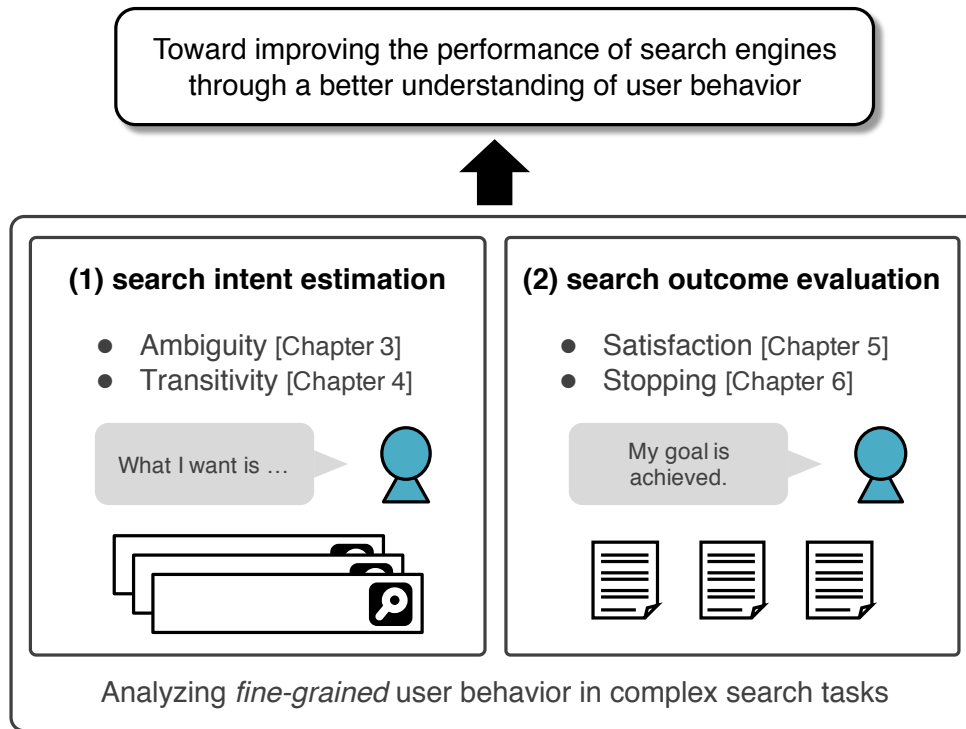


Figure 1.2: Overview of our approaches.

important research problems in the field of information retrieval.

With the aim of make users' search process efficiently and effectively, we address the problem of establishing methodologies for understanding their search behavior especially in complex search tasks. Our key idea for achieving this aim is analyzing *fine-grained* user behavior data observed in those tasks. Figure 1.2 shows the overview of our approaches, which are described in detail below.

First, we need to understand diverse search intents of users who address complex tasks. As described in Section 1.1, the desired information of searchers is not obvious from the initial queries in complex search tasks and is clarified through a series of their searches. Estimating such information in an online manner can be used to provide search results appropriate to search intents of individual users, not of the general population. To this end, we focus on searchers' *eye-gazing* behavior during their browsing Web pages and analyze those data to estimate a set of words relevant to searchers' intents. The transitivity of search topics is another issue concerning search intents in complex tasks. If the shift of search topics is predictable before searchers reformulate their search queries, search engines can proactively suggest queries that users are likely to use in future searches. To make this feasible, we propose a method of predicting the type of query reformulation that will be performed in the subsequent search by utilizing various fine-grained user behavior in the current search, such as *eye-gazing* and *cursor-scrolling* data.

Second, we need to appreciate search outcomes that users achieve through their searches in complex tasks. To do this, we attempt to reveal how searchers gain satisfaction in complex tasks where different *answers* are found from different documents for a factoid question. While the effort expended to complete tasks have been shown to have negative correlation with user satisfaction [29, 32, 93], other factors may play an important role in searchers' perceiving satisfaction in such tasks. We focus on search expertise, which is defined as specialized knowledge on the internal workings of search engines, and analyze how the presence or absence of this factor affects search behavior and outcomes of users. As described in Section 1.1, search stopping is also an important issue concerning search outcomes, especially in tasks where collecting extensive information from different aspects of a topic is required. With the objective of helping searchers find the appropriate timing for search stopping in such tasks, we propose a query suggestion interface visualizing the amount of important information that a user potentially misses collecting from the search results of individual queries. To this end, we propose a metric for evaluating the gain that searchers acquire from a set of their collected documents. This metric is based on *aspects* of the search topic covered by a given document set.

1.3 THESIS STRUCTURE

The remainder of this thesis is structured as follows:

- Chapter 2

We overview existing studies related to our research presented in this thesis.

- Chapter 3

We address the problem of estimating hidden intents of Web search users from their observable search behavior. To this end, we focus on searchers' gazing behavior on their browsed Web pages and assume that terms drawing a high degree of the attention of a searcher are likely to be relevant to his/her topical intent. On the basis of this assumption, we propose a method of estimating a set of weighted terms that clarify the searcher's ambiguous topical intent. Through a laboratory experiment, the proposed method is evaluated from both qualitative and quantitative perspectives.

- Chapter 4

We propose a method to detect the change of search scope in advance of the subsequent search. Following prior work that characterized searchers' query reformulation, we define our addressed task as a classification problem comprising five query reformulation types (*i.e.*, Generalization, Specialization, Parallel Move, Mission Change, and Error Correction). The proposed method utilizes various kinds of user interactions in the current search

in order to predict the query reformulation type of the subsequent search. As well as evaluating the accuracy of the proposed method, we analyze which and to what extent each interaction feature contributes to the prediction.

- Chapter 5

We investigate how users handle inconsistent information by focusing on the effect of search expertise on their search behavior and satisfaction criteria. The research questions addressed in this work are: (1) how does the inconsistency of answers that users find relate to the degree of task satisfaction that they gain? (2) when do users complete search tasks where inconsistent answers exist? and (3) how do answers that users target change as time passes? To answer these questions, we select four tasks comprising factoid questions with inconsistent answers, extract answers found by study participants in these tasks, and analyze their answer-finding behavior in terms of the presence or absence of search expertise. On the basis of our findings, we discuss the design implications for making users without search expertise aware of the existence of inconsistent answers and helping them to search for supporting evidence for answers.

- Chapter 6

We propose ScentBar, a query suggestion interface visualizing the amount of missed information for individual queries, with the goal of helping searchers discover unexplored aspects and find the appropriate timing for search stopping. We define the amount of missed information for a query as the additional gain that can be obtained from unclicked search results of the query, where gain is formalized as a set-wise metric based on aspect importance, aspect novelty, and per-aspect document relevance and is estimated by using a state-of-the-art algorithm for subtopic mining and search result diversification. We conducted a user study involving 24 participants to investigate the effect of showing missed information on users' search strategies and search outcomes. Experimental results showed that ScentBar helped make users' search process more efficient when our gain estimation algorithm worked reasonably.

- Chapter 7

We summarize the research addressed in this thesis and offer several directions to be explored in future work.

RELATED WORK

Existing work related to our study covers the following research areas: (1) mining searchers' hidden intents, (2) analyzing and modeling search behavior logs, and (3) developing systems for supporting users' search activities. This chapter overviews each of these areas in turn.

2.1 INTENT MINING

Many researchers have addressed the task of mining intents hidden behind *ambiguous* or *underspecified* queries [59, 70, 76], where ambiguous queries have multiple distinct interpretations while underspecified ones have multiple aspects [22]. Mined intents can be used to satisfy searchers with diverse intents by, for example, providing answers to a given query within the SERP and returning diversified search results that cover popular subtopics underlying the query.

2.1.1 Direct Answers

Recently, commercial search engines provide a search feature that directly returns answers within the SERP in response to a given query [20]. Thanks to this feature, users with simple search intents (*e.g.*, checking the local weather forecast in this week) can be satisfied without clicking any search results, which is known as *good abandonment* [57].

People use the Web to find answers to their questions [58, 64]. To foster research on systems that directly return answers for a given question, the QA track was held in Text REtrieval Conferences (TREC)s¹ from 1999 to 2007. The most dominant questions used in the TREC 2007 QA track were categorized into the *factoid* type, which asked for a fact-based short answer. As discussed by Dang *et al.* [23], different documents might support contradictory answers as being

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

correct.

2.1.2 Search Result Diversification

Researchers have proposed various search result diversification algorithms [2, 25, 71] to satisfy different user intents behind ambiguous or underspecified queries. Most of these algorithms depend on subtopic mining, which is a technique for mining the intents underlying a given query, in order to find a set of documents that cover as many important intents as possible. Approaches to subtopic mining can be classified into two categories: subtopics that are modeled explicitly or implicitly [37]. Maximal Marginal Relevance (MMR) proposed by Carbonell and Goldstein [13] is based on the implicit modeling of subtopics; it iteratively selects a document that is relevant to a given query and dissimilar from already selected documents as the next element of a diversified document list. Approaches belonging to explicit subtopic modeling include IA-Select proposed by Agrawal *et al.* [2], who defined the problem of search result diversification as discovering a set of documents covering important subtopics. Tsukuda *et al.* [79] proposed one of the state-of-the-art algorithms for explicitly mining subtopics. Their algorithm utilizes different resources to collect subtopics for a given query and clusters the resulting subtopics to identify intents behind the query.

2.2 USER BEHAVIOR ANALYSIS

Much effort has been devoted to analyzing search behavior logs in order to better understand how users engage in search activities [3, 36, 41, 83, 85]. Query logs and clickthrough data are popular resources due to ease of tracking and can be analyzed for various purposes such as predicting search engine switching [33, 84], detecting searcher frustration [27] and modeling user behavior in successful searches [34].

2.2.1 Search Satisfaction

Batch-style evaluation, which is based on assessors' relevance judgments, has been reported to produce different results from user-oriented evaluation, in which the effectiveness of a system is measured through its use by real users [80, 96]. In light of this mismatch, researchers have recently directed their attention to evaluating search outcomes in more direct ways. Particularly, predicting user satisfaction from search behavior is becoming a trending research topic [1, 35, 54, 81], while the granularity of satisfaction to be predicted differs from work to work (*e.g.*, task-level [1] and click-level [54]). Wang *et al.* [81] modeled latent satisfaction for individual actions to predict task satisfaction. Their model outperformed state-of-the-art satisfaction prediction models, suggesting the importance of considering fine-grained actions to predict task satisfaction.

Effort expended to complete tasks has also shown to be related to user satisfaction [29, 32, 93]. Guo *et al.* [32] reported a negative correlation between task satisfaction and effort features (*e.g.*, the numbers of issued queries and browsed pages). Xu and Mease [93] found that time spent completing tasks was negatively correlated with user satisfaction.

2.2.2 User Factors

Some analytical studies have pointed out the effect of user-dependent factors on their search behavior. Hölscher and Strube [41] analyzed the difference in query formulation strategies between users with high search skills and those without. Their analysis showed that the former users used advanced search options more frequently than the latter. Knowledge of search domains, or domain expertise, is known to influence the likelihood of search success [85]. White [83] investigated how the beliefs and biases of users affected their search behavior. His surveys, targeting yes-no questions in the medical domain, revealed that users had the search biases of preferring information affirming their beliefs, which led them to settle on incorrect answers about half the time.

2.2.3 Search Stopping

Various studies have been conducted to understand how users decide when to stop searching [61, 67, 89, 97] and to modeling searchers' stopping behavior [55], while different terminologies have been used for describing this behavior (*e.g.*, search stopping [89] and search patience [61]). Most studies have been based on interviews that clarify the qualitative characteristics of search stopping. Prabha *et al.* [67], for example, interviewed people in academia to analyze how much information is enough to meet their information needs. They reported that study participants had qualitative criteria for search stopping, including whether they feel that sufficient information has been collected (the sense of "good enough" [97]).

Toms and Freund's work [78] is one of the few quantitative studies on search stopping. They analyzed the characteristic actions that preceded people stopping their searches. Wu *et al.* [90] conducted a user study in which they varied the number and distribution of relevant search results and found that these factors had different effects on when searchers left the SERPs. There have also been attempts to mathematically model search stopping behavior [6] using various theories, including Search Economic Theory [4, 5]. Maxwell *et al.* [63] very recently investigated different search stopping rules to find which one approximated actual search behavior most closely.

Dostert and Kelly [24] demonstrated through a user study the gap between cognitive and actual recall: while the study participants believed that they had found about 51–60% of relevant documents when they stopped their search tasks, the actual recall they obtained was less than

10% on average. These results suggest that searchers have difficulty in accurately estimating the amount of relevant information they have (or have not) found.

2.3 SEARCH SUPPORT SYSTEMS

A great number of search interfaces have been developed by researchers with the objective of gaining a better understanding of users' search behavior and/or making their searches more effective. Eye tracking technology have also received great attention from many researchers in recent years, since eye-gazing data have a great potential as a resource of implicit relevance feedback from searchers.

2.3.1 Search Interfaces

The concept of *information scent* was introduced in Information Foraging Theory [66], which explains the information seeking behavior of human beings by making an analogy to the food foraging behavior of animals. In this theory, information scent indicates proximal cues from which searchers perceive the value of distal information sources [18]. As summarized in Hearst's book [39], considerable research has been conducted on developing search interfaces that show users information scent [38, 40, 53, 73, 75].

To help searchers make quick relevance assessments, Hearst [38] proposed the TileBars interface visualizing the query term occurrence as a rectangle for each search result, where the horizontal bar represents the document length and the vertical one represents the query terms. Iwata *et al.* [46] also proposed a tile-based visualization interface, called AspecTiles, which was designed to help users issue queries with multiple aspects. Zha *et al.* [99] proposed an interface that suggests queries with their representative images so that users can efficiently convey their specific search intents. Aiming to help searchers understand inter-query relationships, Kato *et al.* [50] proposed SParQS, which presents query suggestions classified into automatically generated categories. Qvarfordt *et al.* [68] proposed a query preview interface that visualizes three kinds of information on the search results of the current search query in the form of a stacked bar chart: (1) the number of newly retrieved search results, (2) the number of re-retrieved but not clicked ones, and (3) the number of re-retrieved ones that have already been clicked by the searcher.

2.3.2 Eye Tracking

Recently, a wide variety of studies on information retrieval have focused on tracking users' eye movements. Affordable eye tracking devices² and eye tracking softwares with common web-

² <https://theeyetribe.com/>

cams³ gave a big boost to research on utilizing eye tracking technology for building search support systems. Buscher *et al.* [10, 11] proposed a method to extract keywords for query expansion based on users' gaze positions. Xu *et al.* [91] proposed a method of applying implicit relevance feedback to search results for documents, images, and videos on the Web by using eye tracking data. Xu *et al.* [92] also proposed a method to summarize documents on the basis of where a user pays attention when reading.

³ <http://www.inference.phy.cam.ac.uk/opengazer/>

ESTIMATING ATTRACTIVE TERMS FROM GAZING BEHAVIOR

3.1 INTRODUCTION

When users conduct searches on an unfamiliar topic, the underlying topical search intents are initially ambiguous and become clear through their search process. In these cases, queries issued to search engines at the beginning of searches tended to be short and underspecified. Suppose, as an example, users who try to buy the best digital camera to buy. Even if they issue the same query “digital camera” to a search engine in their initial search, the desired information may differ among the users [42]; while some may end up deciding a high-performance camera, others may think price is the first priority and choose another camera sold cheaply. To help users with diverse topical intents, it is essential for search engines to well understand what kind of specific information is desired for individual users.

In this work, we address the problem of clarifying users’ ambiguous topical search intents. Let us consider the above example again. When users are interested in a portable camera, we expect them to gaze frequently at terms such as “compact” and “lightweight” that appear in their browsed Web pages. Therefore, we focus on searchers’ gazing behavior on their browsed Web pages and assume that terms drawing a high degree of the attention of a searcher are likely to be relevant to his/her search intent. On the basis of this assumption, we propose several methods of estimating terms that clarify users’ ambiguous topical search intents by utilizing terms at which they gaze during browsing Web pages, as shown in Figure 3.1.

We conduct a user study to evaluate the estimation accuracy of our proposed methods. In

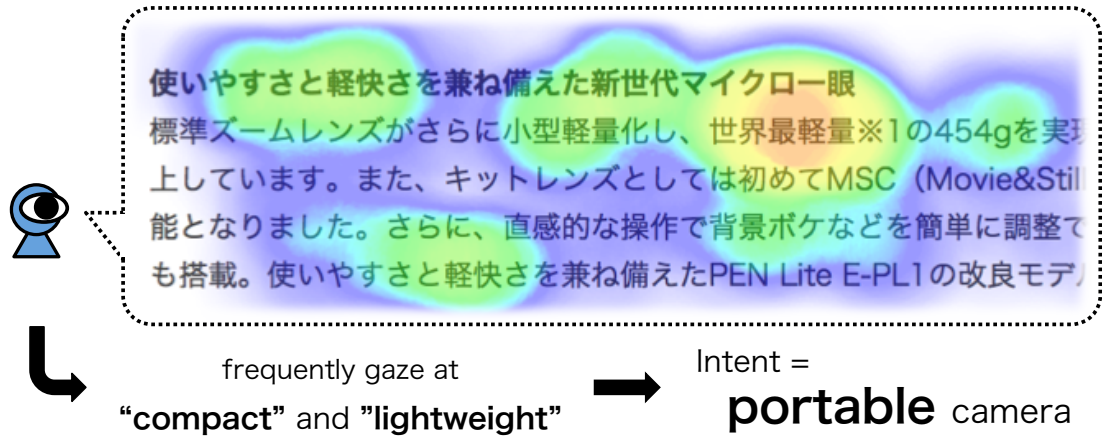


Figure 3.1: Our idea of attractive term estimation. We utilize searcher’s gazing behavior during his/her browsing Web pages.

addition, we analyze the qualitative characteristics of those methods by comparing the overlap of estimated terms between the methods and baseline methods.

3.2 SEARCH INTENT MODEL

Every searcher has a *topical search intent* in his/her mind in the form of an image of desirable information that he/she wants to find through searches. A topical search intent forms the basis of an *information need* [74], which is defined as the perceived need for information that leads to someone using an information retrieval system in the first place. As it is difficult to model topical search intents explicitly, we represent a user’s topical search intent as *a set of pairs of a term and its weight*, where the weight for each term expresses the extent to which the term is relevant to the user’s intent. This idea is based on the vector space model [60]. For example, suppose a user u_1 wants to buy a digital camera that is popular and inexpensive, and another user u_2 wants to buy one that is popular and has high-performance. In this case, topical search intents of these two users may be represented as follows in our model: $\{(\text{camera}, 0.5), (\text{popular}, 0.4), (\text{price}, 0.3), (\text{discount}, 0.3), \dots\}$ for u_1 and $\{(\text{camera}, 0.5), (\text{popular}, 0.4), (\text{wide-angle-lens}, 0.3), (\text{face-recognition}, 0.3), \dots\}$ for u_2 .

Estimating searchers’ topical intents in this manner has a wide range of applications as shown in Figure 3.2. For example, search engines can present appropriate search results to individual searchers by sorting the original search results on the basis of term overlap between those captions and the searchers’ intents. Query suggestion is another application of topical search intent estimation. Furthermore, Web browsers can utilize terms related to a searcher in order to highlight relevant areas in his/her browsing Web pages.

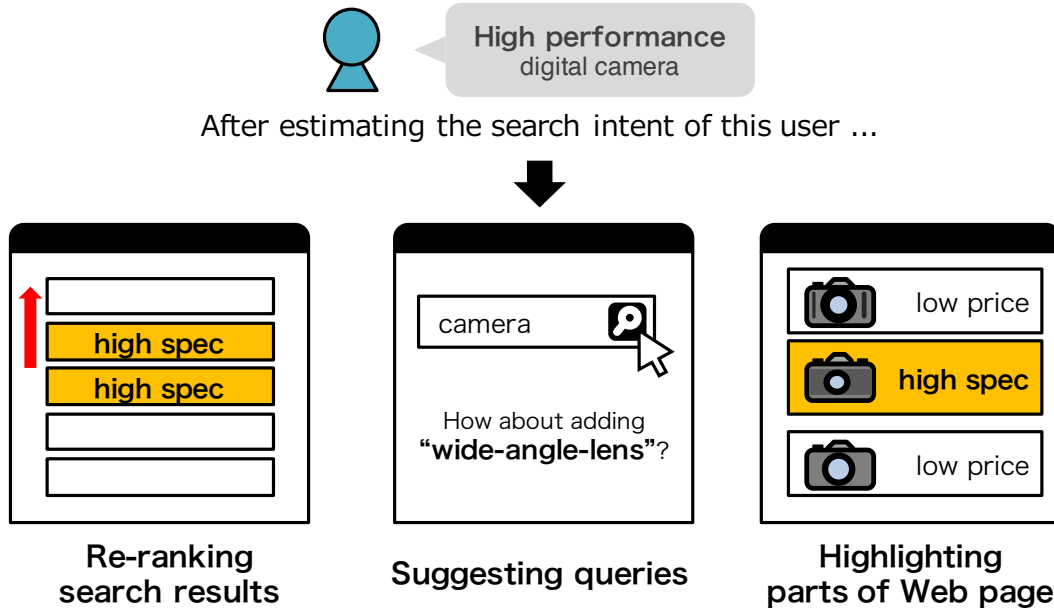


Figure 3.2: Example applications for attractive term estimation. Estimated terms enable search engines to re-rank search results and to suggest useful queries that reflect searcher’s hidden topical intent.

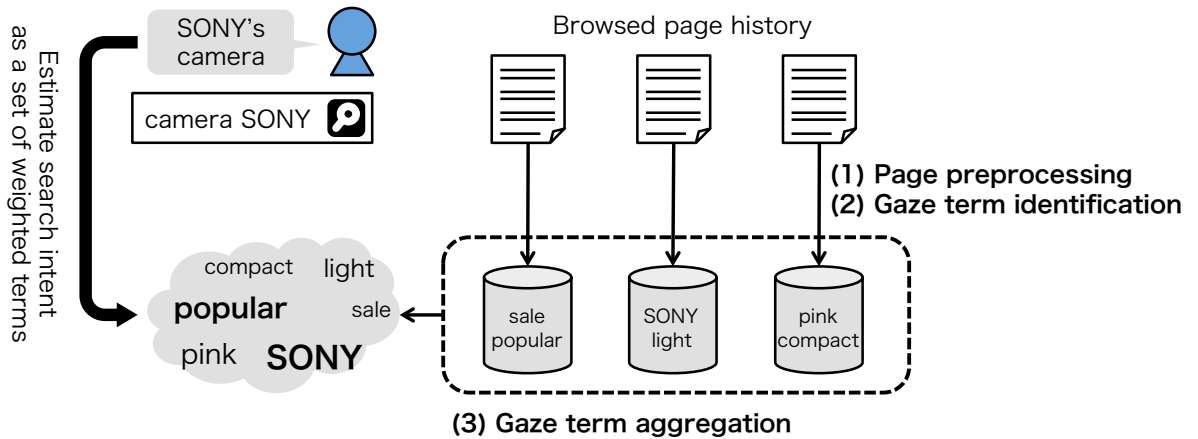


Figure 3.3: Outline of our method to estimate attractive terms.

3.3 ATTRACTIVE TERM ESTIMATION

3.3.1 Method Outline

Figure 3.3 shows the outline of our proposed method of estimating attractive terms. When a searcher visits a new Web page, we start to identify terms at which the searcher gazes (referred to as *gaze terms*) using his/her eye movement data. This step continues until the searcher moves

to another page. We then aggregate gaze terms for each browsed page to estimate the searcher's topical intent through the search session. We describe each process of our method in detail below. Note that how to identify gaze terms from eye movement data is independent from our method of estimating attractive terms. The identification algorithm we implemented is described in Section 3.4.

3.3.2 Weight Calculation for Gaze Terms

For each gaze term, we need to calculate its weight that represents a degree of relevance to a searcher's topical intent. Searchers are more likely to gaze at terms appearing more times in a Web page than less appearing terms. Therefore, there could be any relationship between the number of appearance of a term (*i.e.*, term frequency) and the number of a user's gaze for the term (*i.e.*, gaze frequency). In this work, we make two hypotheses on this relationship in order to calculate the weight for each gaze term.

- **Positive correlation between term frequency and gaze frequency:**

The tf-idf model, a kind of vector space model, regards terms appearing more times in a Web page as more important for characterizing the page. In other words, the more times a term appears in the page, the more relevance the term has to the topical intent of a searcher who visits the page. In this view, we should regard terms as highly relevant to the searcher's topical intent when both term frequency and gaze frequency are high for those terms.

- **Negative correlation between term frequency and gaze frequency:**

If a user gazed at terms many times in spite of their low term frequency, these terms would attract a higher degree of his/her interest than those having the same degree of gaze frequency and the higher degree of term frequency. For example, when a searcher who wants a popular digital camera compares multiple products at an online shopping Web site, both terms "camera" and "hot-selling" are likely to be gazed at many times by him/her. However, the former would have much higher term frequency than the latter, while the latter might be a more specific term relevant to the searcher's topical intent. In this view, we should impose some penalties on terms having high term frequency when estimating their relevance.

On the basis of the above hypotheses, we define four equations that calculate the weight for

each term in a Web page as follows:

$$\text{Score}_{\text{MGT}}(t, p) = \text{gf}(t, p) \cdot \text{tf}(t, p), \quad (3.1)$$

$$\text{Score}_{\text{DGT}}(t, p) = \frac{\text{gf}(t, p)}{\text{tf}(t, p)}, \quad (3.2)$$

$$\text{Score}_{\text{nMGT}}(t, p) = \frac{\text{gf}(t, p)}{\max_{t_i \in p} \text{gf}(t_i, p)} \cdot \frac{\text{tf}(t, p)}{\max_{t_i \in p} \text{tf}(t_i, p)}, \quad (3.3)$$

$$\text{Score}_{\text{nDGT}}(t, p) = \frac{\text{gf}(t, p)}{\max_{t_i \in p} \text{gf}(t_i, p)} / \frac{\text{tf}(t, p)}{\max_{t_i \in p} \text{tf}(t_i, p)}, \quad (3.4)$$

where $\text{gf}(t, p)$ denotes gaze frequency of a term t in a page p , and $\text{tf}(t, p)$ denotes term frequency of t in p .

The *MGT* (Multiply Gaze by Tf) model assumes the positive correlation between term frequency and gaze frequency. It gives a high weight score to a term where both term frequency and gaze frequency are high. In contrast, the *DGT* (Divide Gaze by Tf) model is based on the negative correlation hypothesis and performs score calculation by giving heavy weight to a term that has high gaze frequency and low term frequency. The *nMGT* and *nDGT* models are normalized versions of *MGT* and *DGT*; these models absorb the difference in the range of values that $\text{gf}(t, p)$ or $\text{tf}(t, p)$ takes among Web pages.

3.3.3 Gaze Term Aggregation Among Browsed Pages

After a searcher browsed several Web pages, we aggregate gaze terms for each of these pages in order to estimate which terms attracted the searcher through his/her search session. In this work, we assume that a searcher's topical intent does not shift drastically during his/her session: *e.g.*, a user who searches for digital cameras are not interested in comic books in the search session. We calculate, under this assumption, the relevant score of a term t to a user u 's topical search intent by the following equation:

$$\text{Intent}_m(t, P_u) = \sum_{p_u^{(i)} \in P_u} \text{Score}_m(t, p_u^{(i)}),$$

where m denotes a weight score calculation model described above, $P_u = \{p_u^{(1)}, p_u^{(2)}, \dots\}$ denotes a set of Web pages browsed by u . As shown in the above equation, we calculate the relevance score of a term by summing its weight score for each browsed page uniformly. This model gives a high score to a term that constantly attracts a user's interest throughout his/her search session.

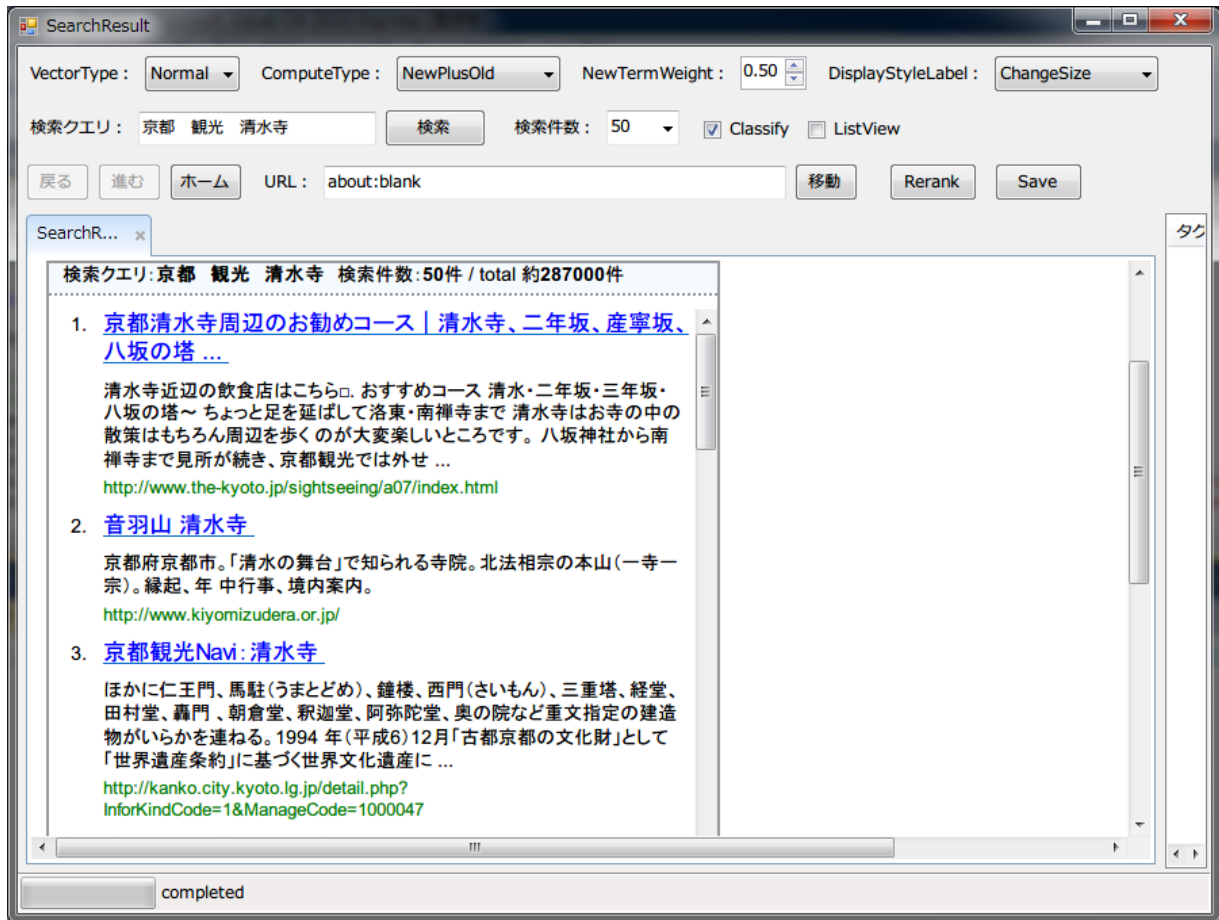


Figure 3.4: Screenshot of our prototype system for estimating attractive terms.

3.4 IMPLEMENTATION

To evaluate the effectiveness of our proposed method, we implemented a prototype system as a desktop application using the programming language C#. Figure 3.4 shows the prototype interface. Though its appearance differs little from existing Web browsers, it continues to identify gaze terms during the search session and estimates a searcher's topical intent on the fly. When implementing this system, we used the Web search API provided by Yahoo! JAPAN¹ to obtain Web search results. We also used the eye tracking device Tobii T60² and the software development kit Tobii SDK to obtain searchers' gazing data every 16 ms (= 60 Hz). We describe the details of the implemented system below.

¹ <http://developer.yahoo.co.jp/webapi/search>

² <http://www.tobii.com/>

Algorithm 3.1 EMBEDTAGS(n, l_{cls})**Input:** a DOM node n , a label l_{cls}

```

1: for each  $n_c$  such that a child node of  $n$  do
2:   if  $n_c$  is text node then
3:      $M \leftarrow$  term set obtained from morphological analysis for  $n_c$ 
4:     for each  $m$  such that an element in  $M$  do
5:        $n'_c \leftarrow$  a node of font element whose text value is  $m$ 
6:       if not  $m$  is a stopword then
7:         set  $l_{cls}$  as a class name of  $n'_c$  // used for gaze term identification
8:       end if
9:       insert  $n'_c$  before  $n_c$ 
10:    end for
11:    remove  $n_c$  from DOM tree
12:  else
13:    if not ( $n_c$  is script element or  $n_c$  is style element) then
14:      EMBEDTAGS( $n_c, l_{cls}$ ) // process recursively
15:    end if
16:  end if
17: end for

```

3.4.1 Preprocessing for Gaze Term Identification

To estimate a searcher's topical intent using the proposed method, we need to identify which terms the searcher gazes at when his/her browsing Web pages. However, conventional Web browsers do not offer any function that tells us a term on a given coordinate; we can only know an HTML element on the coordinate. In the prototype system, we dealt with this problem by embedding an original tag in each candidate term. Specifically, we apply the procedure EMBEDTAGS(n_{body}, l_{cls}) shown in Algorithm 3.1 for a body element n_{body} of a Web page and a label l_{cls} ³. We used a morphological analyzer MeCab⁴ for parsing text in a page. Using the above algorithm, we wrap terms that are not stop-words with the original tag.

³ We tested various HTML elements for embedding terms and adopted the `font` element, which had the least effect on the original page style.

⁴ <http://mecab.sourceforge.net/>

3.4.2 Gaze Term Identification from Gaze Position Data

Having embedded a tag for each term appearing in a Web page, we can know whether a user gaze at any term or not. Specifically, gaze term identification is achieved in the following manner:

1. Obtain an HTML element on the searcher's gaze position returned by the eye tracker;
2. If the element's class name is l_{cls} , the inner text of the element is the term at which the user gazes; otherwise the user does not gaze at any term.

In this work, we use the `GetElementFromPoint` method offered by C# for accessing the HTML element on a given coordinate.

3.5 EXPERIMENT

We conducted an experiment to evaluate the estimation accuracy of our proposed method and to investigate which model among Equations 3.1 through 3.4 worked best for calculating the weight for each gaze term.

3.5.1 Methods

To confirm the effectiveness of our method of estimating terms relevant to searchers' topical intents, we prepared two baselines in this experiment: the *TF* model calculates the relevance score of a term to a searcher's topical intent using only term frequency of the term, which corresponds to the equation $\text{Score}_{\text{TF}}(t, p) = \text{tf}(t, p)$; on the other hand, the *GAZE* model calculates the relevance score using only gaze frequency of the term, which corresponds to the equation $\text{Score}_{\text{GAZE}}(t, p) = \text{gf}(t, p)$.

Comparing the result of our proposed models to that of the TF model enables us to validate the advantage of ours over the existing feature extraction approaches based solely on term frequency. We also evaluate the effectiveness of combining gaze frequency and term frequency by comparing our models and the GAZE model. In summary, we performed attractive term estimation by using six models (four are proposed ones and two are baselines) in the experiment and compared the difference in estimated terms between these models.

3.5.2 Experimental Design

We prepared two types of search tasks for our experiment as shown in Table 3.1. The following criteria were used for task selection:

- Popular task often performed by many searchers;
- Task where relevant information depends on individual searchers.

Table 3.1: Tasks used in our experiment.

Task	Search context
Task 1	You intend to buy a smartphone, so please search for information about it.
Task 2	You intend to go on a trip to Okinawa, which is one of the most popular tourist resorts in Japan, so please search for information to make a trip plan.

We asked eight volunteers to take part in this experiment, all of whom were male undergraduate or graduate students familiar with information seeking on the Web. Prior to beginning the experiment, we explained the usage of our prototype system and the context behind the search tasks to each participant. Each study participant took the following procedure for each task.

1. The participant performs calibration to identify individual characteristics of his/her eye movements (less than 10 seconds);
2. The participant searches for information on the basis of the specified search context for 10 minutes by freely modifying search queries and browsing Web pages;
3. The system continues to identify gaze terms and estimates the participant's topical search intent using different estimation models;
4. After the task ended, the system extracts up to 15 terms that are regarded as most relevant by each model, merges all terms into a list in random order, and presents the list to the participant as shown in Figure 3.5;
5. The participant rates each term in the list on a three-point scale on the basis of its relevance to the participant's topical search intent (zero, one, and two correspond to irrelevant, somewhat relevant, and relevant, respectively).

3.5.3 Accuracy of Estimated Terms

To evaluate the accuracy of our proposed method, we calculated nDCG@15 (normalized Discounted Cumulative Gain) [47] and MAP@15 (Mean Average Precision) [60] for the ranking of terms scored by each model. The results are shown in Table 3.2. When calculating MAP scores, we regarded terms rated as relevant as correct answers, and other terms as incorrect.

As is clear from Table 3.2, nMGT, one of our proposed models, obtained the highest values in terms of nDCG and MAP; it improved the MAP@15 score by about 11% compared to the

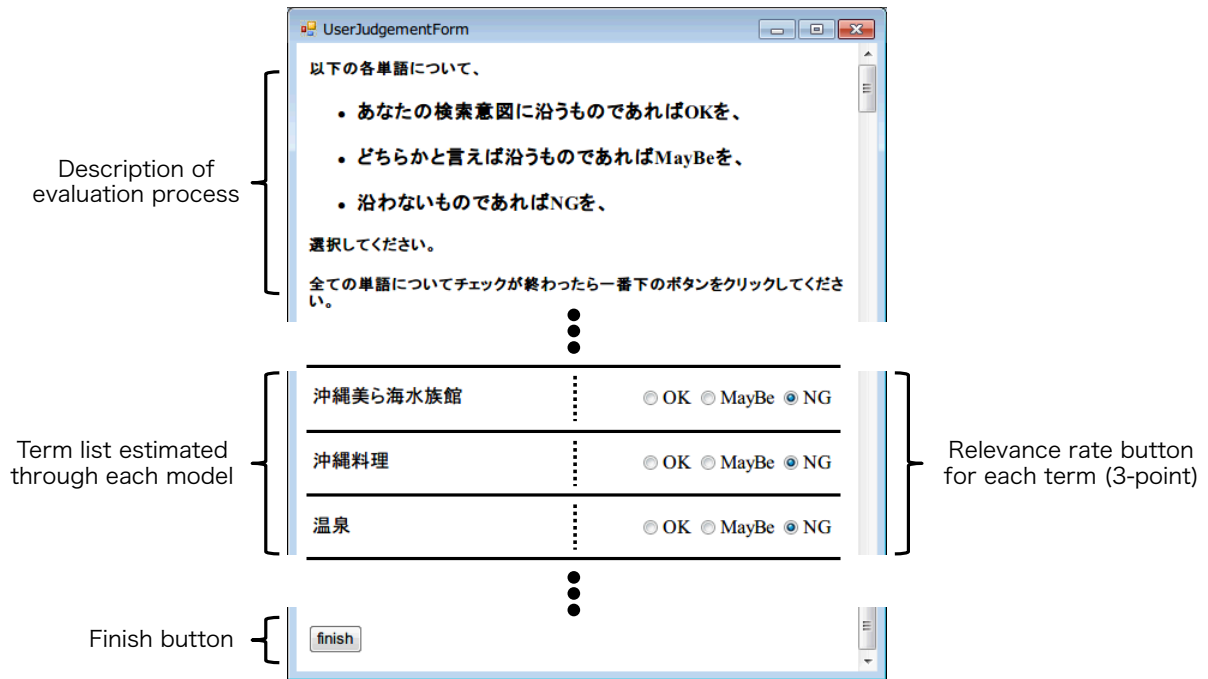


Figure 3.5: Interface used for annotating relevance of estimated attractive terms.

Table 3.2: Results of nDCG@15 and MAP@15 for the ranking of terms estimated by each model. The best result is indicated by boldface for each task.

	nDCG@15			MAP@15
	Task 1	Task 2	Avg.	Avg.
TF	0.758	0.747	0.753	0.586
GAZE	0.806	0.763	0.784	0.594
MGT	0.823	0.752	0.788	0.583
nMGT	0.826	0.805	0.816	0.652
DGT	0.683	0.722	0.702	0.394
nDGT	0.700	0.656	0.678	0.393

baseline TF model, and by about 10% compared to the baseline GAZE model. The nMGT model also improved the nDCG score by about 8% compared to the TF model, and 4% compared to the GAZE model, though the improvement in nDCG@15 was smaller than that in MAP@15. These results indicated that considering gaze frequency feature had an advantage over the existing feature extraction approaches based solely on term frequency. By comparing the result of the GAZE model to that of the nMGT model, we also found that combining gaze frequency and

term frequency features enabled us to estimate searchers' topical intents with higher accuracy than using each feature alone. We infer from these results that searchers' eye-gazing data (related to readers' intents) bear a complementary relationship to term distribution data in Web pages (related to writers' intents).

Our experimental results also showed that the estimation accuracy for the DGT and nDGT models were substantially lower than that for the baseline models. As is clear from the definitions (*i.e.*, Equations 3.2 and 3.4), these models give high scores to terms at which a searcher frequently gaze in spite of low term frequency. As an extreme example, when a searcher happens to see a term appearing only once in a Web page, this term obtains a high score from these models regardless his/her intent. This drawback of the DGT and nDGT models (*i.e.*, low robustness for noisy data) could be considered as the main cause of their low estimation accuracy in our experiment.

In Section 3.3.2, we proposed the nMGT and nDGT models, which normalize the range of term frequency and gaze frequency that can be distributed differently in different Web pages, as variants for the MGT and DGT models, respectively. The normalization was effective in attractive term estimation for the positive correlation model (*i.e.*, MGT), while the performance was worsened for the negative correlation model (*i.e.*, DGT). We could not reveal the cause of this result, which suggests that additional experiments would be needed for further investigation.

3.5.4 Uniqueness of Estimated Terms

To evaluate whether our proposed method can estimate searcher-dependent topical intents, we looked into terms estimated by our best model (*i.e.*, nMGT) and those by the baseline models. To this end, we introduce a notion of the uniqueness of a searcher's topical intent. The score $\text{Unique}_m(u, U)$ denotes the uniqueness of a user u 's intent compared to other users U , where m is a model of attractive term estimation. This score indicates the extent to which the model m can estimate the unique topical intent that the user u has, that is, any other user $u_i \in U$ does not have such intent. In this work, we define the uniqueness score of a user u as the ratio of the number of terms relevant to u 's intent estimated by a model m and not included in the set of relevant terms for any other user $u_i \in U$ to the number of all terms relevant to u 's intent estimated by m . Using this definition, we calculate mean unique rate among users U_s as follows:

$$\overline{\text{Unique}}_m(U_s) = \frac{\sum_{u_i \in U_s} \text{Unique}_m(u_i, U_s - \{u_i\})}{|U_s|}.$$

Obtaining a high score from $\text{MeanUnique}_m(U_s)$ indicates that the model m can estimate unique topical intents for individual searchers. We use this metric for evaluating the uniqueness of estimated search intents.

Table 3.3: Mean unique rate (with standard deviation) of estimated terms among study participants. The best result is indicated by boldface for each task. Significant differences with the TF model at $p < 0.05$ and $p < 0.01$ are marked with “*” and “”, respectively.**

	Task 1	Task 2	Avg.
TF	0.35 (0.32)	0.54 (0.24)	0.44 (0.23)
GAZE	0.57 (0.26)	0.75* (0.22)	0.66** (0.22)
nMGT	0.65* (0.15)	0.79* (0.08)	0.72** (0.10)

Table 3.3 shows the result of the mean unique rate for each task. We can find from this table that the uniqueness of the nMGT and GAZE models outperformed the result of the TF model. Applying the paired t -test revealed that the nMGT model improved the uniqueness score significantly for both Task 1 and Task 2 ($p < 0.05$), and whole tasks ($p < 0.01$), compared to the TF model. Furthermore, the nMGT model obtained the higher uniqueness score than the GAZE model in every task, though we could not find a significant difference between these models. These results verify that searchers have different topical intents even if they conduct searches on the same topic and that such searcher-dependent intents can be inferred by leveraging their eye-gazing behavior on Web pages.

Table 3.4 shows the actual examples of relevant term sets judged by three participants in our experiment. Terms relevant to a single participant are shown in boldface. The TF model tended to extract abstract or generic terms such as “buy”, “cell-phone” or “travel” to many participants. In contrast, many terms extracted by our nMGT model were related to specific concepts or entities and differed among participants. For example, the nMGT model could extract terms of historic places to one participant and those related to whale watching to another participant in Task 2.

Table 3.4: Actual examples of relevant term sets judged by study participants. Unique terms are shown in boldface.

Task 1	
u_1	TF スマートフォン, iPhone, アプリ
	nMGT スマートフォン, Openroid, おサイフケータイ, Android アプリマーケット, アプリケーション, REGZA
u_2	TF スマートフォン, iPhone, 買う, Android
	nMGT スマートフォン, iPhone, 買う, スマートフォン遊び, DesireHD, Android, モバイルインターネット, おすすめスマートフォンレビュー
u_8	TF スマートフォン, 携帯電話, Xperia, 機能, NTT ドコモ, ドコモ
	nMGT スマートフォン, Xperia, XPERIA, 購入, BlackBerryRCurve, BlackBerry, 機能
Task 2	
u_6	TF 沖縄, 本部, 沖縄県, 地図, 海, 旅行, クチコミ, 自然, 観光, 旅行記
	nMGT 沖縄美ら海水族館, 沖縄, 旧海軍司令部壕, 古宇利島, 体験工房, 沖縄首里城公園, 水族館, コメント
u_7	TF 沖縄, 旅行
	nMGT ホテル, 沖縄, 沖縄旅行, ビーチ, お得
u_8	TF 沖縄, 国際通り, 海, 深海, 観光, 居酒屋, 那覇
	nMGT コマカ無人島, 深海, マリンスポーツ, 沖縄料理, 国際通り県庁前商店街, 石焼ステーキ, 魚, 水族館エリアガイド, 沖縄美ら海水族館, 国際通り

3.6 DISCUSSION

In the previous section, we confirmed the advantage of our proposed method in terms of both estimation accuracy and uniqueness, though there remains some room for our method to improve, which is discussed in this section.

3.6.1 Attention to Visual Information

When checking the Web pages that study participants had browsed in our experiment and the log files in which their relevance judgment data were recorded, we found that our proposed method had not worked well when they had often browsed pages containing Adobe Flash⁵ contents or many images (*e.g.*, maps and smart phones images). Study participants tended to gaze more at those visual contents than at textual contents due to the attractiveness. As our proposed method is based on gaze frequency of terms appearing in their browsed Web pages, searchers' high attention to visual contents would lead to inaccurate estimation of attractive terms. This is a possible reason why there was no great difference in estimation accuracy between our nMGT model and the baseline models. Some solution has to be proposed for this problem in order to improve the accuracy and robustness of our estimation method.

3.6.2 Eye Movement Patterns and Search Strategies

This work assumed that terms drawing a high degree of a searcher's attention were likely to be relevant to his/her topical intent. However, this assumption may be doubtful in some cases. For example, when search results returned by a search engine are very noisy, a searcher may filter out irrelevant results by skimming them one by one. In that case, the searcher would gaze at irrelevant terms many times when investigating search results. To deal with this problem, we need to identify which strategy (*e.g.*, filtering phase and comparing phase) users employ in the current search. Utilizing eye movement patterns could play an important role in search strategy identification. If a searcher's gaze points go back and forth between two areas, he/she may compare two items. Reading speed is another factor worth considering; searchers are likely to read Web pages at high speed when the pages contain much noisy information. Thus, we need to incorporate eye-gazing patterns in our proposed method for more accurate intent estimation.

3.7 SUMMARY

In this work, we addressed the problem of estimating terms that clarify ambiguous topical intents by using searchers' eye-gazing behavior on their browsed Web pages. We proposed four models

⁵ <http://www.adobe.com/products/flash/>

and evaluated through a user study involving eight participants, each of whom conducted two search tasks. Experimental results showed that the nMGT model, which assumes positive correlation between term frequency and gaze frequency, performed best among our proposed four models and two baseline models. Our qualitative analysis also revealed that the nMGT model had an advantage over other models in extracting terms capturing the individual differences in user interests.

While the present work focused on users' attention to textual contents in Web pages, it would also be worth addressing how to incorporate their attention to visual contents in our proposed method. Furthermore, consideration of contextual information, such as eye movement patterns and appearance positions of words in Web pages, would be needed for more accurate intent estimation.

PREDICTING QUERY REFORMULATION FROM CURRENT SEARCH BEHAVIOR

4.1 INTRODUCTION

About 28% of search queries are reused in the following searches after some reformulation [44]. Suppose a user searches for information about trip to a country in Asia. This task consists of multiple sub-tasks such as determining the place of trip, sightseeing spots and a hotel to stay at. Therefore, he/she may have to conduct Web searches from the various point of view using multiple queries. A user's search intent could change within such tasks. In the above example, the user's first intent may be to obtain popular information about a trip to Japan if a search query is like "Japan trip famous". Then if the user finds Kyoto being a famous tourist area in Japan for foreign travelers and gets interested in obtaining more detailed information about Kyoto, he/she may change the search query accordingly. In this case, it can be seen that his/her search intent changes from a broad one to narrow one. Similarly, if this user wants to know which hotel is good to stay at after making the search on the trip destination, we can say that the focus point of the search changes from the destination to the hotel.

In this way, transitions of users' search intents cause reformulations of search queries. The existing work [8] introduced the following five query reformulation types (QRTs): *Specialization*, *Parallel Move* and so on. For example, when a user reformulates the current search query "movie thrilling" to "Japanese movie thrilling", the reformulated query is more specific than the one before the reformulation. This type is called Specialization. In the case of a query reformulation from "Kyoto travel hotel" to "Kyoto travel transportation", on the other hand, both queries are

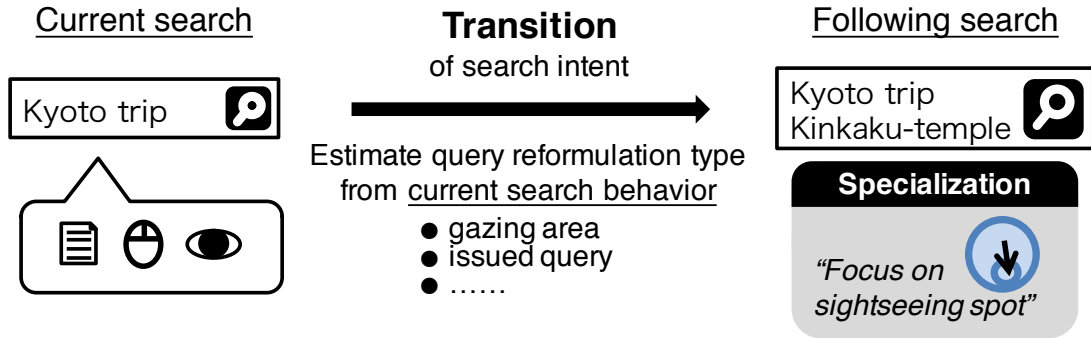


Figure 4.1: Our idea of query reformulation prediction. We leverage various kinds of user behavior observed in the current search to predict the type of query reformulation in the subsequent search.

used to make a search about a trip to Kyoto, while the search aspect changes before and after the reformulation. This reformulation type is called Parallel Move.

In this work, we address the task of prediction about the reformulation type between the current query and one that will be issued in the subsequent search. An overview of our approach is shown in Figure 4.1. We collect as search behavior logs not only *coarse-grained* data (e.g., search queries, browsed Web pages and clicked search results) used in the existing research [84, 27], but also *fine-grained* data (e.g., eye-gazing behavior, mouse cursor positions, and fired mouse events). Using the collected behavior logs, we calculate various kinds of features and construct a Support Vector Machine (SVM) classifier. This classifier predicts the type of query reformulation that will be occurred in the subsequent search on the basis of features in the current search. In addition to investigating the classification accuracy, we also analyze and discuss which kinds of features are useful for predicting users’ query reformulation types.

Our method and findings are important for search engines to develop interfaces and interaction designs that support searchers on the basis of their behavior. Conventional search engines provide information only after search queries are issued to them. If the dynamics of searchers’ intents were known in advance of query issuing events, search engines could support searchers in various ways (e.g., suggesting queries that are likely to be issued for the next search and changing the ranking of search results on the basis of the shift of searchers’ intents). Moreover, more appropriate advertisements could be provided in SERPs according to searchers’ intent transitions. For example, when a searcher’s target changes from restaurants to hotels during his/her search with the query “Kyoto famous restaurant”, the search engines can draw his/her attention to the advertisement area by changing its content to information about hotels in Kyoto. These functions can respond to users’ variable search intents and help their searches by offering more appropriate

information.

The contributions of this work can be summarized as follows.

- We proposed a method of predicting the type of query reformulation occurred in the subsequent search using user logs of the current search. Our proposed method can be applied to supporting searchers on the fly by changing information dynamically when the shift of their intents is detected.
- We performed in-depth analyses of the relationship between user behavior in the current search and query reformulation in the next search. Our results are beneficial to better understanding and modeling search behavior.

4.2 PROBLEM DEFINITION

As described in Section 4.1, we address the task of prediction about the type of query reformulation from their search behavior data. Before formalizing the problem addressed, we first describe the *query reformulation type (QRT)*, which is the output of our tackled problem.

4.2.1 Query Reformulation Type (QRT)

Boldi *et al.* [8] classified the relationship between two queries issued successively into the following five categories, so called QRTs: (i) *Generalization*, (ii) *Specialization*, (iii) *Parallel Move*, (iv) *Mission Change*, and (v) *Error Correction*.

Generalization arises when a user wants broader information than that obtained from the current search. For example, if a user modifies the search query from “Kyoto famous Japanese food” to “Kyoto famous food”, it can be regarded that he/she is relaxing the search condition about food in Kyoto by omitting the term “Japanese”. Specialization, as opposed to Generalization, arises when a user narrows down the current search. When the query is modified from “Kyoto sightseeing temple” to “Kyoto sightseeing Kiyomizu-temple”, the corresponding QRT is Specialization. When a user modifies the current query to change the search aspect with the same context, its QRT is regarded as Parallel Move. Specifically, in the situation about planning a trip in the summer vacation, reformulation from a hotel-related query to a food-related one is an example of Parallel Move. Mission Change arises when a user changes the search topic to an entirely different one. The last one is Error Correction, which arises when a user’s search intent does not change before and after the reformulation. Correcting a misspelled term and issuing a paraphrased query belong to this QRT.

4.2.2 Prediction Task

In this work, we denote a set of QRTs described above as $T = \{G, S, P, M, E\}$ (each type is expressed by a heading character), a sequence of queries issued by a user u as $Q_u = \{q_1^u, q_2^u, \dots\}$, behavior logs of u regarding the search query q_i^u as $BL(q_i^u)$, and the corresponding QRT between q_i^u and q_{i+1}^u as $QRT(q_i^u, q_{i+1}^u) \in T$. Using this notation, we define the problem addressed in this work as predicting a QRT $t = QRT(q_c^u, q_n^u)$ from behavior logs about the current search $BL(q_c^u)$, where q_c^u is a query that a user u issued in the current search, and q_n^u is a query that u will issue in the subsequent search.

4.3 QUERY REFORMULATION TYPE PREDICTION

4.3.1 Collection Procedure of Behavior Log Data

For predicting users' QRTs, we first need to collect their behavior log data during information seeking. We collect not only *coarse-grained* behavior logs (*e.g.*, query and click information used in the existing work [84, 27]) but also *fine-grained* ones (*e.g.*, mouse cursor movements and eye-gazing positions). Each log is recorded with a timestamp when its event is occurred. We use an eye-tracker called Tobii T60 and a software development kit Tobii SDK, both of which are Tobii products¹, for obtaining users' eye-gazing positions in real time. We track searchers' mouse cursor movements and eye-gazing positions at 100 ms and 25 ms intervals, respectively. These sampling intervals are determined by reference to the work by Huang *et al.* [45].

4.3.2 Labeling Behavior Log Data as QRT

After collecting behavior log data in the above manner, we then construct training data for setting up a QRT classifier by splitting the collected data into subgroups. Specifically, training data consist of behavior log data $BL(q_i^u)$ for each query q_i^u in the issued query set Q_u of a user u . As shown in Figure 4.2, the label for this data corresponds to a QRT between the search regarding q_i^u and the following one. For example, suppose there is a consecutive log data pair $(BL(q_i^u), BL(q_{i+1}^u))$ of a user u , and q_i^u and q_{i+1}^u corresponding to “Kyoto sightseeing” and “Kyoto sightseeing temple” respectively. In this case, the QRT label of the data $BL(q_i^u)$ is Specialization in accordance with the query reformulation type between the two queries. If no search follows the query q_{i+1}^u (*i.e.*, the last query in each day), we regard it as the end of the search session, label $BL(q_{i+1}^u)$ as Mission Change.

As it is sometimes not easy for a person to label the behavior data of anyone else, we decided to leave the above labeling task up to the searcher himself/herself. Figure 4.3 shows the interface

¹ <http://www.tobii.com>

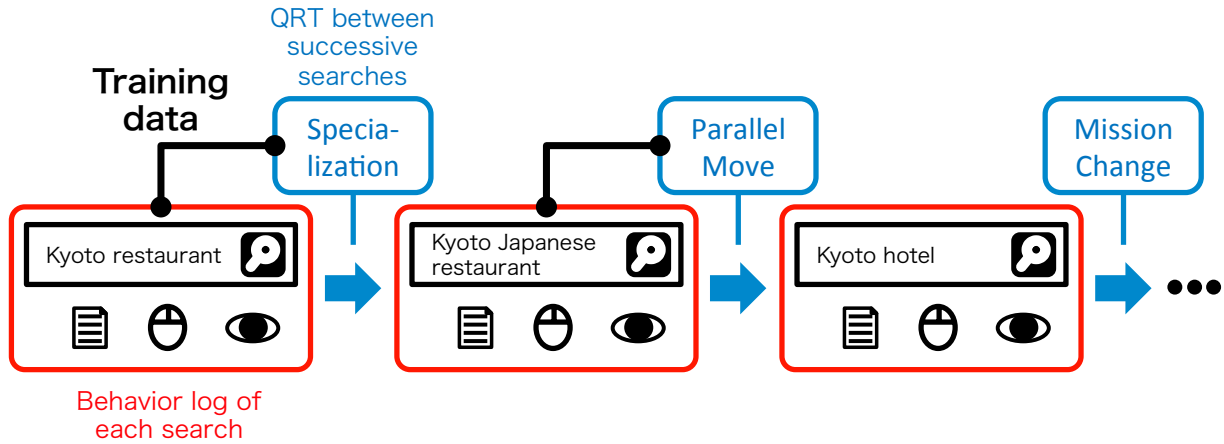


Figure 4.2: Constructing training data from behavior log data.



Figure 4.3: Interface of our labeling system for constructing training log data.

of our labeling system. This system shows a searcher a record of his/her behavior including issued queries, browsed Web pages, and clicked search results. Searchers use the system to label the training data at the end of each day.

4.3.3 Features

We adopt Support Vector Machine (SVM), one of major machine learning techniques, for constructing a classifier that predicts users' QRTs from their behavior data during a Web search. In

this work, we calculate various kinds of features from the following five feature categories:

- **Transition Features**

In this work, we use the behavior log data of the current search in order to predict a QRT between the current search and subsequent one. In addition to this data, the data about past searches can also be useful for the QRT prediction in the current search. The previous research [8] stated that Specialization often arises after Generalization and vice versa, which suggests that knowing a QRT regarding the last search may help us predict the next query reformulation. In line with this thinking, we use the QRT of the last search as one of the features of machine learning.

- **Query Features**

It is reasonable to assume that search queries issued by a user represent his/her search intent to some extent. For example, if a current query is given, we can predict the size of the current search topic. Other features such as the hit count for the current query and the number of suggested queries can be extracted from search log data easily. These features may trigger a change of search target just as a short search query drives users to narrow down the search target in the next search. Therefore, we add this kind of feature to the candidate ones for classifier construction. Query-pair features such as the Jaccard coefficient between the current query and the following one used in [8] could be useful for predicting QRTs. However, the goal of this work is the QRT prediction from the current search behavior, which is eventually aimed at supporting the subsequent searches of users. Thus, we do not take into account features regarding the subsequent search this time.

- **Click Features**

As well as query-related features, ones related to users' clicks on hyperlinks can also be used for the learning process. Which search results are clicked by a user is one example of this feature category. As stated in the previous work about click-through [48], this information can be used as a metric of users' satisfaction with the current search. For example, suppose that a user modifies the current search query without clicking any search results. In this case, it may be because he/she has found a spelling error in the current query or he/she is not satisfied with the search result against the query. This situation may cause the query reformulation of the Error Correction, Generalization or Specialization types with higher probability than other types.

- **Span Features**

Features in this category express to what extent users engage in search or browsing ac-

tivities. Specifically, how many SERPs a user investigates or how long the user browses Web pages during the current search task belong to this feature category. In particular, the latter feature, which is called dwell time, has been used for predicting the relevance of Web pages [31]. From these features, we can predict the difficulty of the current search task. If the current search task is very hard, a user needs to investigate many search results. Since it takes a user a long time to find relevant information in this case, he/she may give up the current task and change the search topic to a completely different one in the next search. In this way, this kind of features can enable us to predict QRTs such as Parallel Move and Mission Change in particular, where a user's intent transits to another topic.

- **Gaze Features**

Even if a user browses the same Web page, areas that attract his/her attention vary depending on many factors such as the type of the search task or his/her search intent. In fact, the existing research [30] reported that in a transactional task [9], where a searcher performed a certain transaction such as shopping or downloading something, the click-through rate for advertisements in SERPs was higher than for other task types. In this case, the following search may be affected by Web pages browsed after clicking the advertisement [17]. With this background, we expect that information on which area a searcher sees can be a key indicator to understand his/her next search intent. Therefore, we use features derived from users' eye movements for QRT prediction.

- **Mouse Features**

If mouse position data are available, one can obtain various types of information, including scroll speed and browsed areas in Web pages. These data may partly imply users' interests for Web pages as shown in [30]. We therefore use those features in this work. In addition, previous work [45] reported that the correlation between mouse cursor movements and eye movements tended to be high in SERPs, which indicates that search engines could use mouse-related features instead of eye-related features for predicting users' search intent and supporting their Web searches in practice.

We adopt, on the basis of the above discussion, the features shown in Table 4.1. The next section describes the performance of our classifier constructed with these features.

² When clustering a gaze position (x, y) in a page whose width is w and height is h , we assign it into the left area if $x < 250$, otherwise the right area if $x > w - 250$, otherwise the top area if $y < 250$, otherwise the bottom area if $y > h - 250$, otherwise the center area. We empirically use 250 px as the clustering threshold.

Table 4.1: Features used for constructing our QRT classifier.

Transition Features	
$[f_1]$	whether the last QRT is Generalization
$[f_2]$	whether the last QRT is Specialization
$[f_3]$	whether the last QRT is Parallel Move
$[f_4]$	whether the last QRT is Mission Change
$[f_5]$	whether the last QRT is Error Correction
Query Features	
$[f_6]$	the number of terms in query
$[f_7]$	the number of characters in query
$[f_8]$	hit count of search results
$[f_9]$	the number of suggested queries
$[f_{10}]$	the number of advertisements
Gaze Features	
$[f_{11,12,13}]$	Avg., Max., Min. ratio of gaze for search results
$[f_{14,15,16}]$	Avg., Max., Min. ratio of gaze for suggested queries
$[f_{17,18,19}]$	Avg., Max., Min. ratio of gaze for advertisements
$[f_{20,21,22}]$	Avg., Max., Min. eye movement speeds in SERPs
$[f_{23,24,25}]$	Avg., Max., Min. eye movement speeds in Web pages
$[f_{26,27,28}]$	Avg., Max., Min. ratio of gaze for left area of Web page ²
$[f_{29,30,31}]$	Avg., Max., Min. ratio of gaze for right area of Web page ²
$[f_{32,33,34}]$	Avg., Max., Min. ratio of gaze for top area of Web page ²
$[f_{35,36,37}]$	Avg., Max., Min. ratio of gaze for bottom area of Web page ²
$[f_{38,39,40}]$	Avg., Max., Min. ratio of gaze for center area of Web page ²
Click Features	
$[f_{41}]$	the number of clicks on search results
$[f_{42}]$	the number of clicks on suggested queries
$[f_{43}]$	the number of any clicks on advertisements
$[f_{44}]$	the number of clicks on any areas (<i>i.e.</i> , $f_{41} + f_{42} + f_{43}$)
$[f_{45}]$	the number of switching between search and browsing tabs
$[f_{46}]$	Avg. number of switching per click (<i>i.e.</i> , f_{45} / f_{44})
$[f_{47}]$	whether any search results is clicked
$[f_{48}]$	the rank of firstly clicked search result
$[f_{49}]$	Avg. ranks of clicked search results
$[f_{50}]$	whether any suggested query is clicked
$[f_{51}]$	the rank of clicked suggested query
$[f_{52}]$	whether any advertisement is clicked
$[f_{53}]$	the rank of firstly clicked advertisement
$[f_{54}]$	Avg. ranks of clicked advertisements
Span Features	
$[f_{55}]$	max offset of displayed search results
$[f_{56}]$	the number of browsed SERPs
$[f_{57}]$	the total time spent in search process
$[f_{58}]$	Avg. time spent in each SERP
$[f_{59}]$	the number of browsed Web pages
$[f_{60}]$	the total time spent in browsing process
$[f_{61}]$	Avg. time spent in each Web page
Mouse Features	
$[f_{62,63,64}]$	Avg., Max., Min. mouse cursor movement speeds in SERPs
$[f_{65,66,67}]$	Avg., Max., Min. mouse cursor movement speeds in Web pages

Table 4.2: Actual examples of search tasks and query reformulations included in the collected behavior log data. A symbol on an arrow means a QRT between two queries before and after the edge.

Search Task	Example of Query Reformulations
Job Search	“NTT DoCoMo group discussion” $\xrightarrow{E}$ “NTT DoCoMo group interview” $\xrightarrow{P}$ “NTT Data second interview”
Doctoral Course	“doctor course continuance” $\xrightarrow{S}$ “doctor course job” $\xrightarrow{P}$ “doctor course decision” $\xrightarrow{E}$ “doctor course determination”
Trip to Europe	“Italy Capri” $\xrightarrow{S}$ “Italy Capri Blue Grotto” $\xrightarrow{G}$ “Italy Capri sightseeing destination”
PC Trouble	“PC heat dissipation” $\xrightarrow{S}$ “PC high temperature” $\xrightarrow{P}$ “PC temperature speecy” $\xrightarrow{E}$ “PC temperature speccy”
Programming	“typing” $\xrightarrow{S}$ “dynamic typing strong weak” $\xrightarrow{P}$ “Python object-oriented” $\xrightarrow{S}$ “Python object-oriented inheritance”

4.4 EVALUATION

We constructed a classifier using the collected behavior log data and evaluated the classifier in order to determine whether our method could predict the query reformulation type accurately. In the remainder of this section, we first give an overview of the collected behavior log data. Then, we describe the process of classifier construction, followed by the evaluation results.

4.4.1 Overview of Collected Behavior Log Data

We collected behavior log data in a Web search with QRTs as labels through the procedure described in Section 4.3.1. We used the data of the author of this thesis, collected for five days (2012/4/4, 4/25, 5/4, 5/6 and 5/7) in order to construct a classifier. Instead of specifying some search context to the subject during the process of collecting logs, we allowed him to search and browse Web pages in his own way.

As a result of the above process, we obtained behavior log data during information seeking consisting of 183 search queries. Table 4.2 shows the search tasks and query reformulations in those tasks included in the log data. In addition, the distribution ratio of QRTs is listed in Table 4.3 together with that presented in [8]. In the table, we calculate the QRT distributions

Table 4.3: Distributions of QRTs in the collected behavior log data. The left side is about two kinds of distributions of our collected data: all the five QRTs (all-types) and the four QRTs except Mission Change (non-MC). The QRT distributions reported in [8] are shown on the right side. UK means the QRT distributions in Yahoo! UK, and US means those in Yahoo! US.

QRT	Log Distributions in this Work		Log Distributions in Existing Work [8]	
	all-types	non-MC	UK	US
Generalization	13.1%	15.1%	4.4%	9.5%
Specialization	36.6%	42.1%	37.5%	30.1%
Parallel Move	29.5%	34.0%	47.7%	55.5%
Error Correction	7.7%	8.8%	10.4%	5.0%
Mission Change	13.1%	_____	_____	_____
Dataset Size	0.2k		6.5M	10.5M

in two ways: *all-types* is the result for all five QRTs; and *non-MC* is about four of these types, excluding Mission Change. The non-MC result indicates to what extent each QRT arises on average within one search session.

In the all-types result, Specialization and Parallel Move are dominant QRTs. Behavior logs that belong to one of these two types account for 66.1% of all logs. Moreover, this ratio rises to 76.1% in the non-MC result. These results are similar to that of the previous work [8], where the authors analyzed the QRT distributions using large-scale query log data of two search engines; Yahoo! UK and Yahoo! US, denoted as *UK* and *US* in the table, respectively. There is a little difference between the two results. Specialization arises more often compared to Parallel Move in the distribution of our dataset. On the other hand, it is inverted in the result of the previous work. This may be because our subject is a careful searcher who tends to spend a longer time on one topic or aspect in the search process than the average user. The actual number of Mission Change logs in the collected data is 24, which means that the average number of queries per session is about 7. However, the difference is not so great, so we think the collected data is sufficiently coherent with natural user search behavior.

4.4.2 Classifier Setup

Our addressed problem concerns multi-class classification (*i.e.*, output class belongs to the QRT set T described in Section 4.2). We adopted the existing SVM library LIBSVM³. As for SVM setting, we used RBF kernel as the kernel function, and default values as its parameters. In order to investigate both classification accuracy for the entire QRT set and that for each QRT, we constructed two types of QRT classifiers using the following ways [26].

- *One-versus-one*: This approach constructs a sub-classifier for each combination of two classes, and each classifier predicts which one of the two classes the input data belong to. Thus, the number of sub-classifiers is $nC_2 = n(n-1)/2$, where n is the number of classes. The output class is determined on the basis of the majority voting of the output of each sub-classifier.
- *One-versus-rest*: This approach constructs a sub-classifier for each class. This means there are n sub-classifiers in this approach. The output class is determined on the basis of the decision value or the belonging probability of each sub-classifier.

4.4.3 Classification by One-Versus-One Approach

We first report that the evaluation result for clustering the entire QRT set using the classifier constructed through the one-versus-one approach.

Classification Accuracy

To calculate the classification accuracy, we applied k -fold cross-validation to the behavior log data, where k was set to the sampled dataset size ($= 183$) given the scale of our collected dataset. We prepared two baselines for comparing the results: Baseline_{rand} and Baseline_{dom}. Given a behavior data as input, the former randomly outputs a QRT from all the five types, and the latter always outputs the dominant QRT (Specialization in our case) as output. Table 4.4 shows the classification accuracy of each method, where accuracy was calculated by dividing the number of true-positive instances by the whole data size.

As shown in the table, the accuracy of our method is 0.41, which is better than the other two baselines. However, the difference in accuracy between our method and the baselines' is not so large. We discuss the effect of the training dataset size on prediction performance in Section 4.5.

³ <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

Table 4.4: Classification accuracy for each method.

Proposed	Baseline _{rand}	Baseline _{dom}
0.410	0.200	0.366

Table 4.5: Useful/Useless features for QRT prediction with accuracy after ablation. A feature that has low/high accuracy makes a great/small contribution to the prediction.

Top Three Features Contributing to Classification		
1: f_{47} (0.393)	2: f_{27} (0.399)	3: $f_{3,6,61}$ (0.404)
Bottom Three Features Contributing to Classification		
1: f_1 (0.432)	2: $f_{9,39}$ (0.421)	3: $f_{5,10,62,63,64,23,38}$ (0.415)

Useful Features

We tried to find features that contributed to predicting the entire QRT set. For each feature listed in Table 4.1, we compared the accuracy before and after ablating it from the feature set. If the accuracy decreases after the feature ablation, it can be said that the feature makes a contribution to the classification. We show the features that make a high/low contribution in Table 4.5. Regarding features that are not shown in the table, their accuracy did not change.

From this table, one can find that for some features such as whether any search result is clicked (f_{47}), attention degree for left area of Web pages (f_{27}) and the number of terms that compose the query (f_6), the accuracy after ablation decreases compared to the original one. This result shows that these features are more useful than other features for predicting users' QRTs.

In many Web pages, the advertisement information tends to be shown on the left or right side. The result implies that the feature concerning the interest in such an area can be effective for predicting some kinds of QRTs (*e.g.*, Mission Change). The result also implies that one can predict a QRT from the feature concerning the number of terms in the query. This can be because a user who starts his/her search tasks with a short and broad query tends to perform the next search with a longer and more specific query. In the collected behavior log actually, about 60% of QRTs are Specialization if the number of query terms is less than three, while this drops to about 18% if the query contains more than two terms.

On the other hand, ablating some features including attention degree for center area of Web

pages ($f_{38,39}$) and whether the last QRT is Generalization (f_1) resulted in the improvement of classification accuracy. This means that these features are harmful for QRT prediction. While the attention on the left area is useful for classification, attention on the center area is not useful. This may be because, whatever search intent users have, they usually browse the center area of a Web page first since it often contain main topic of the page. Therefore, it is hard to predict the next QRT from such information. Regarding the last transition feature, we believe that it improves the classification accuracy due to the alternating Generalization and Specialization, which is reported in [8]. However, the evaluation result implies the transition feature does not make any contributions to predicting QRTs. Though we could not find any specific reason, it might be due to the sample size of the users' behavior log data. However, this can be interpreted as a positive result that implies that our methodology can be applied to the prediction of the first QRT in a search session as well.

4.4.4 Classification by One-Versus-Rest Approach

By using the one-versus-rest approach, we can get the ranking of QRTs for given input data. Using the other classifier constructed through this approach, we also evaluated the classification result for each QRT.

Classification Accuracy

We first divided a series of collected behavior log data sets into two groups using timestamp information. Then we repeated the following steps by interchanging training data and test data:

1. Construct a classifier using one of these groups as training data through the one-versus-rest approach;
2. For each test data in another group, rank each QRT on the basis of the belonging probability of the data for its class and evaluate the ranking.

As the metrics for evaluating the obtained ranking, we adopted Mean Reciprocal Rank (MRR), as there is only one correct answer in the ranking. MRR is defined as follows:

$$\text{MRR} = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{1}{\text{rank}_i},$$

where D denotes test data, and rank_i denotes the rank of the first correct answer (QRT in this case) in the ranking for a data instance $d_i \in D$.

Table 4.6 shows the scores of MRR for the first half group and the last half group, and the mean of those scores. When we used the first half behavior log data as test data, the Generalization classifier had the highest prediction accuracy. On the other hand, the Specialization classifier

Table 4.6: MRR for the ranking of the predicted QRTs.

QRT	The First Half Log Data	The Last Half Log Data	Avg.
Generalization	0.75	0.46	0.60
Specialization	0.40	0.67	0.54
Parallel Move	0.61	0.55	0.58
Error Correction	0.59	0.64	0.61
Mission Change	0.55	0.48	0.51
Macro Average	0.58	0.56	0.57
Micro Average	0.53	0.54	0.54

was the worst prediction model. However, when the last half behavior log data was used as test data, predication accuracy of these two classifiers was reversed. Although the reason for this result is not clear, we think it is partly because there is a bias in QRT distribution between the above two groups. In fact, the ratio of log data whose label is Error Correction is about 13% in the first half group and about 2% in the last half group.

In spite of these differences, there were some tendencies that were common in both data group. MRR scores of the ranking for Error Correction was always higher than the values of Macro and Micro Average. In contrast, MRR scores for Mission Change was lower than Macro Average in both data groups. These results may be due to the difference between the scope of each QRT; the scope of Error Correction is narrow while that of Mission Change is too wide. Therefore, we need to investigate features that contribute to predicting Mission Change in particular for the future.

Useful Features

We also evaluated which features contribute to classification for each QRT. We conducted the feature ablation procedure *recursively*. In other words, we continued to ablate the most effective feature until the feature set became empty. More specifically, we sampled test data D_t from all logs for each QRT t and obtained the ranking of ablated features by applying the procedure $\text{ITERATEABLATION}(F, |F|, D_t)$ shown in Algorithm 4.1, where F represents a set of features listed in Table 4.1.

The result of recursive feature ablation is shown in Table 4.7. For each QRT, this table shows the best 20 ablation features (in column f) with ablation order (in r), and the difference of the

Algorithm 4.1 ITERATEABLATION(F, n_{abl}, D)**Input:** a set of features F , number of features to be ablated n_{abl} ($\leq |F|$), test data D **Output:** an ordered set of ablated features F_{abl}

```

1:  $F_{abl} \leftarrow \emptyset$ 
2: while  $|F_{abl}| < n_{abl}$  do
3:   //  $\text{Accuracy}_{\text{SVM}}(\cdot, \cdot)$  represents the accuracy of SVM classifier
4:   // built with given features and tested with given test data.
5:    $f_{abl} \leftarrow \arg \min_{f \in F} \text{Accuracy}_{\text{SVM}}(F \setminus \{f\}, D)$ 
6:    $F_{abl} \leftarrow F_{abl} \cup \{f_{abl}\}$ 
7:    $F \leftarrow F \setminus \{f_{abl}\}$ 
8: end while
9: return  $F_{abl}$ 

```

order from the mean value (in d). Features are sorted in ascending order of the sum of r and d (in $r + d$). The lower values a feature has for $r + d$ in QRT t , the greater the contribution to classification it makes and the more specific it is compared to other features for t .

From this result, we can find some characteristics of effective features in each QRT. As for the Generalization classifier, the hit count feature (f_8) is included in the column. This implies that when the hit count is too small for a user, he/she tends to reformulate the current query to a broader one. The feature regarding the number of tokens in the search query (f_6) is ablated much earlier in the Specialization classifier. It is likely that this feature made a contribution to this classifier because a short query was inappropriate for the user to find the relevant information, and he/she narrowed down the search with a Specialization query. Regarding the Parallel Move and the Mission Change classifiers, many advertisement-related features are present (e.g., $f_{10,43,54}$). Also, features about eye-gaze to boundary areas in Web pages (e.g., f_{31}) seem to be useful for classifying these QRTs. As these features attract users in terms of content and visual effect, their intent transits from the current search to the advertisement information. We found an interesting result for Error Correction. Its column in the table includes the features of the number of query tokens and query characters ($f_{6,7}$). Long queries would increase the probability of typing error. Therefore, these features are ranked higher for Error Correction.

Table 4.7: Ablated features for each QRT classifier. In this table, f denotes the feature name shown in Table 4.1, r denotes the ablation order, d denotes the difference between r and the mean ablation order. A feature having a low value for $r + d$ in QRT t makes a great contribution to classification of t . As for the features indicated in boldface, the reason why these worked well could be inferred.

Generalization				Specialization				Parallel Move				Mission Change				Error Correction			
f	r	d	$r + d$	f	r	d	$r + d$	f	r	d	$r + d$	f	r	d	$r + d$	f	r	d	$r + d$
f_{27}	28	-33	-5	f_6	4	-23	-19	f_{67}	20	-25	-5	f_{47}	6	-14	-8	f_4	5	-19	-14
f_{54}	10	-12	-2	f_{22}	28	-30	-2	f_{35}	19	-22	-3	f_{10}	24	-29	-5	f_9	14	-18	-4
f_{42}	1	-1	0	f_{15}	19	-19	0	f_{63}	12	-14	-2	f_{49}	13	-17	-4	f_8	3	-6	-3
f_{51}	4	-2	2	f_{65}	14	-14	0	f_{55}	3	-4	-1	f_7	8	-9	-1	f_{44}	1	-3	-2
f_{44}	3	-1	2	f_{52}	2	-1	1	f_{54}	11	-11	0	f_{46}	5	-6	-1	f_5	16	-17	-1
f_{43}	2	1	3	f_{43}	1	0	1	f_{62}	8	-7	1	f_{42}	1	-1	0	f_{45}	12	-11	1
f_{50}	7	-3	4	f_{61}	23	-21	2	f_{52}	2	-1	1	f_{44}	3	-1	2	f_{46}	6	-5	1
f_{56}	12	-7	5	f_{20}	10	-8	2	f_{43}	1	0	1	f_{43}	2	1	3	f_{42}	2	0	2
f_8	8	-1	7	f_{42}	3	1	4	f_{14}	5	-3	2	f_{50}	7	-3	4	f_2	17	-14	3
f_{53}	6	1	7	f_{53}	5	0	5	f_{53}	4	-1	3	f_{52}	4	1	5	f_{50}	7	-3	4
f_{52}	5	2	7	f_{47}	13	-7	6	f_{11}	9	-5	4	f_{60}	19	-10	9	f_{41}	9	-4	5
f_{66}	22	-12	10	f_{57}	9	-3	6	f_{17}	14	-7	7	f_{41}	11	-2	9	f_{43}	4	3	7
f_{16}	23	-12	11	f_{51}	6	0	6	f_{51}	7	1	8	f_8	9	0	9	f_7	13	-4	9
f_{46}	11	0	11	f_{11}	11	-3	8	f_{18}	24	-14	10	f_4	18	-6	12	f_{56}	15	-4	11
f_{55}	9	2	11	f_{14}	8	0	8	f_{37}	23	-13	10	f_{55}	10	3	13	f_{52}	8	5	13
f_{33}	31	-19	12	f_{62}	12	-3	9	f_{42}	6	4	10	f_{34}	28	-14	14	f_{54}	18	-4	14
f_{20}	16	-2	14	f_{44}	7	3	10	f_{26}	29	-17	12	f_{31}	35	-20	15	f_{16}	25	-10	15
f_{11}	14	0	14	f_{45}	17	-6	11	f_{58}	25	-12	13	f_{14}	12	4	16	f_6	21	-6	15
f_{48}	33	-18	15	f_{17}	16	-5	11	f_{23}	27	-13	14	f_{15}	17	1	18	f_{53}	10	5	15
f_{59}	29	-14	15	f_{38}	31	-18	13	f_{65}	21	-7	14	f_9	26	-6	20	f_{51}	11	5	16

4.5 DISCUSSION

4.5.1 Sufficient Number of Training Data

In the evaluation of classification result for the entire QRT set, there is not much difference in accuracy between our method and the baseline method $\text{Baseline}_{\text{dom}}$. One possible reason may be the small size (= 183) of the sample data. To observe the relationship between the training data size and classification accuracy, we calculated the accuracy while changing the training size from 30 to 150. Since the classification accuracy may be affected by sampled data, we performed the above calculation five times with different sample sets and averaged the resulting accuracy scores. The result is given in Table 4.8.

Table 4.8: Classification accuracy among different sampling sizes.

Iteration	Sampling size of training data				
	30	60	90	120	150
1st	0.33	0.39	0.39	0.44	0.44
2nd	0.33	0.33	0.33	0.33	0.44
3rd	0.33	0.33	0.44	0.33	0.44
4th	0.28	0.33	0.33	0.39	0.38
5th	0.33	0.33	0.44	0.44	0.44

From this table, we can see that the bigger the sampling size, the more accurate the prediction of QRTs by the classifier becomes. This implies that an increase in behavior log data can lead to more accurate prediction. Therefore, it is expected that the classification performance shown in Tables 4.4 and 4.6 will improve by continuing the search behavior log data collection in the future.

4.5.2 Influence of Preceding Search Behavior on Next Search

The classifier constructed using our method takes the current search behavior data as input, and outputs the QRT from the current search to the following search. In other words, we assume that user behaviors for QRTs are based on the first order Markov model [19] (*i.e.*, their next search intent is determined only by the current search state). However, this was too strict to model such behaviors, as is clear from the low accuracy of evaluation. For example, suppose a user finds an unknown term during a certain search task. In this case, he/she is likely to search for the meaning of the term, and then to return to the former search task. Our method could not deal with such search processes appropriately. Conditional Random Field (CRF) may offer a solution since this model takes into account contextual information (*e.g.*, QRT sequence) as hidden states. Though CRF cannot be applied into our task this time mainly due to the small data size, we believe that classification accuracy will improve by using this model.

4.5.3 Predictability of Subsequent Query from Search Behavior

In this work, we mainly addressed the task of predicting QRTs from search behaviors. However, this is a first-step work to support users' subsequent searches. If search engines can predict what query a user will issue on the following search from his/her current behavior, they can help the user in various ways including query suggestions.

Using the collected log data, we briefly see if it is possible to know the subsequent query from

Table 4.9: MAP for the ranking of predicted subsequent queries.

Method	Generalization	Specialization	Parallel Move	Error Correction	Mission Change
TF	0.09	0.04	0.03	0.09	0.02
GAZE	0.09	0.10	0.04	0.19	0.02

eye-gazing behavior in the following manner:

1. For each log group $BL(q_i^u)$, extract terms gazed at by the subject u and rank them in descending order of gazing frequency (referred to as GAZE method);
2. Evaluate the rankings using Mean Average Precision (MAP).

When calculating precision, we treat a term included not in the current query but in the following one as relevant. We also rank terms that appeared in browsed pages in descending order of term frequency, and evaluate the ranking for comparison (referred to as TF method). The result is shown in Table 4.9.

Though all QRTs have low precision scores due to the difficulty of this task, the GAZE method outperforms the TF method for Specialization and Error Correction. This result implies that when a user reformulates a query to either equal to or more than specific one, terms which draw user’s attention in current search tend to be used as the next query. On the other hand, in both method the precision is very low for Parallel Move and Mission Change. Prediction improvement for such QRTs is an issue in the future.

4.6 SUMMARY

In this chapter, we proposed a machine learning based approach to predict five types of users’ query reformulations during Web searches. We constructed a classifier using the collected log data about search behaviors and evaluated the classification accuracy. As a result, we found that our method could predict the query reformulation types with about 41% accuracy, despite the small size of training data. In addition, further analysis implied that some features such as query term length, attention area in Web pages, and clicks on search results were vital for capturing the dynamics of searchers’ topical intents.

As future work, we need to confirm to what extent the classification accuracy improves and how much training data is required for much better prediction by using a larger quantity of training data. In addition, many features described in this work were easy to calculate but superficial.

Therefore, we need to develop more semantic and meaningful features for constructing a classifier, such as the relationship between attention terms and issued query, page reading speed, and the patterns of eye movements (*e.g.*, comparing and skipping) in order to improve the classification accuracy. Moreover, to understand user behavior it is important to know which kinds of features are user-independent or user-dependent. Therefore, we will construct a classifier using one person's behavior data and then evaluate the classifier with another person's behavior data.

ANALYZING HOW INCONSISTENT INFORMATION AFFECTS SEARCH BEHAVIOR

5.1 INTRODUCTION

People sometimes conduct complex search tasks [87] where their information needs are not satisfied by a single search query. A task can be complex even if it comprises a single search goal [49]. Take, as one such example, a user who queries a search engine about the worst year for droughts in US history, expecting a single answer. If the engine returned different search results giving different years as the answer to the query, this user would be unsure about what the correct answer was and might finish the task dissatisfied. To fulfill the goal of this task, more queries would need to be issued and more documents assessed so that he/she could collect possible years with supporting evidence and judge the most likely one as the task answer.

The issue in the above example is not limited to *relevance* [12], which has been studied for decades, but more complex notions, like *credibility* [28], might also play important roles in such intellectual search activities. As explained in the literature [94], the credibility of information is a subjective quality whose interpretation depends on its receivers. If users do not understand the underlying algorithms of search engines, they might trust answers described in top-ranked documents as credible and finish their task without in-depth exploration. While specialized knowledge has been shown to affect search strategies and outcome evaluations of users [41, 83, 85, 94], little is still known about how answer inconsistency affects them.

We investigated how users handle inconsistent information found in search tasks, focusing on *search expertise* (defined as specialized knowledge on the internal workings of search engines) as a possible factor affecting user behavior and satisfaction. The research questions addressed in this work are: (1) how do found answers relate to task satisfaction? (2) when do users finish searching for answers? and (3) how do users' targeted answers change as time passes? To answer these questions, we selected four tasks comprising factoid questions [23] and having inconsistent answers, extracted answers that 30 study participants had found in these tasks, and analyzed answer-finding behavior of the participants in terms of the presence or absence of search expertise.

The main findings from our analysis can be summarized as follows: (1) finding inconsistent answers caused dissatisfaction for users with search expertise (search experts), while effort in searching for answers was the dominant factor in task satisfaction for those without search expertise (search non-experts); (2) search experts tended to spend longer completing tasks than search non-experts even after finding possible answers; and (3) search experts narrowed down the targeted information to promising answers as time passed as opposed to search non-experts, whose found answers were still diverse even in the closing stage of their task sessions. On the basis of these findings, which suggest search non-experts care little about the consistency in their found answers, we discuss the design implications for making search non-experts aware of the existence of inconsistent answers and helping them to search for supporting evidence for answers.

5.2 DATASET

We adopted as a source of our analysis the publicly available search logs¹ of a user study conducted by Feild *et al.* [27]. Four tasks were selected from this source, in which inconsistent information existed as the task answers. We manually checked Web pages that study participants had browsed in these tasks to identify their found answers.

5.2.1 User Study

Feild *et al.* [27] conducted their study in the fall of 2009, aiming at predicting frustration of users from their search behavior. Given the purpose of frustration prediction, all the 12 search tasks in this study were designed to be difficult to achieve (*i.e.*, they could not be completed just by browsing a single Web page). As study participants, 30 people (23 males and 7 females) were recruited from diverse departments (computer science, English, kinesiology, physics, *etc.*) of a university. Each participant was asked to conduct seven search tasks one by one, whose ordering was determined by the Latin square design so as to remove task ordering effects. The

¹ <http://hank.feild.net/downloads.html>

Table 5.1: Task descriptions given to participants of user study.

Task	Description
Drought	In what year did the USA experience its worst drought? What was the average precipitation in the country that year?
Pixels	How many pixels must be dead on a MacBook before Apple will replace the laptop? Assume the laptop is still under warranty.
TV	What was the best selling television (brand & model) of 2008?
Verizon	What's the helpline phone number for Verizon Wireless in MA?

participant's activities (including search queries and browsed Web pages) were logged with the timestamps during each task session. A post-questionnaire was administered after each task was executed, in which participants reported the following items on the basis of their search outcomes: (1) the degree to which their information needs were satisfied (on a five-point scale: 1 = not satisfied at all; 5 = completely satisfied) and (2) answers that they found and judged as likely. Further information on this user study can be found elsewhere [27].

5.2.2 Task Selection

Three search tasks of the above-mentioned user study (*e.g.*, "Name three bridges that collapsed in the USA since 2007.") could be categorized into *list* questions [23], which asked for distinct, multiple instances satisfying the information need. These were excluded from our analysis as users would anticipate the existence of multiple answers in such tasks. This left nine tasks corresponding to *factoid* questions [23], which asked for a single answer. These tasks were relevant to our interests as different documents might support contradictory answers as being correct. We decided, however, to exclude four of them comprising question templates (*e.g.*, "Find the hours of the PetsMart nearest [Wichita, KS; Thorndale, TX; Nitro, WV]."), because different variants for an identical task had different answers, which would make our analysis more complicated. One task whose correct answer was time-dependent (*i.e.*, "How much did the Dow Jones Industrial Average increase/decrease at the end of yesterday?") was also excluded for similar reasons. In this way, the remaining four tasks were targeted for our analysis: *Drought*, *Pixels*, *TV*, and *Verizon*. Table 5.1 summarizes the descriptions of these tasks given to study participants. As shown at the top of Table 5.2, participants did find inconsistent answers in these tasks.

Table 5.2: Statistics on task answers found by study participants with answer samples.

	Drought	Pixels	TV	Verizon
Answer samples	1930 to 1931	Any pixels	Samsung	800-922-0204
	1950s	5 pixels	LN32B460	800-899-4249
	1988 to 1989	Case by case	LN52A650	800-256-4646
	2001 to 2003	No public policy	Sony	1-800-VERIZON
# of answers	33	17	5	7
Cohen's κ	0.75	0.52	0.78	1.00

5.2.3 Answer Extraction

While our interests lie in the effect of inconsistent answers on answer-finding behavior of users, the original logs contained only the answers that participants believed to be most likely. To obtain an entire set of answers that they found through their sessions of the four targeted tasks, two of the authors independently extracted answer entities (simply referred to as answers hereinafter) from each page they browsed. Note that some pages could have more than one answer. We excluded SERPs from extraction sources for two reasons: (1) users do not always scan all titles and snippets in SERPs; and (2) these descriptions are sometimes insufficient for answer evidence due to being short and incomplete. To make the extracted answers as consistent as possible, the assessors shared a clear criterion for answer extraction; descriptions in non-SERPs should be extracted if and only if they, at least partly, answer the task questions listed in Table 5.1. Table 5.2 summarizes the inter-assessor agreement scores of extracted answers in terms of Cohen's κ . When calculating these scores, we regarded pages from which both (or neither) of the assessors extracted answers as accordantly judged. The score across all four tasks was 0.78, which could be regarded as substantial agreement according to the literature [56]. A similar tendency was observed for each task except for the Pixels task, where the score corresponded to moderate agreement ($\kappa = 0.52$). As can be seen from the samples of extracted answers in Table 5.2, the Pixels task did not have common notation patterns as its answers unlike other tasks, which could be a possible reason for its comparatively low score. In addition, the assessors might miss some answers to be extracted as they were not native speakers of English, which was the language used in the target pages. To deal with this problem, we decided to use in our analysis a union of answers extracted by the assessors as the ground truth of answers that participants found.

5.3 ANALYSIS

This section describes our analysis on how searchers handle inconsistent information. In light of the argument described in Section 5.1, we focused on search expertise as a possible factor affecting their search behavior and outcomes, where search expertise was defined as specialized knowledge on the internal workings of search engines. Given the information on participants' major fields logged in the original dataset, we regarded those who studied information retrieval as having search expertise. Hereinafter, we will refer to users with search expertise as *search experts* and those without it as *search non-experts*². There were 66 task sessions in total, of which 17 ($\simeq 26\%$) were done by 7 search experts and 49 ($\simeq 74\%$) by 23 non-experts. Participants performed at least two tasks on average.

When testing significant effects in our analysis, we used the significance level of $\alpha = .05$. On ground of the small sample size of our dataset, however, the phrase “a trend toward significance” is also used when the p -value of a significance test falls within the range from .05 to .10. Overall, the means of task satisfaction reported by search experts and non-experts were 3.65 and 3.47, respectively. The difference was shown to be insignificant by Welch's t -test ($t(41) = .704$, $p = .485$), which suggests that both experts and non-experts felt, on average, almost the same degree of satisfaction. To explore in detail the effect of search expertise on answer-finding behavior, we analyzed the relationship between found answers and task satisfaction (Section 5.3.1), the distribution of time spent finding answers and completing tasks (Section 5.3.2), and the change in participants' targeted answers over time (Section 5.3.3).

5.3.1 How do Answers Relate to Satisfaction?

We began by analyzing the effect of search expertise on task satisfaction. As effort expended to complete tasks has shown to be negatively correlated with search experience [3, 32], we formulated the following hypothesis on the relationship between effort and satisfaction:

Hypothesis 1 (H1): *Increased search effort leads to dissatisfaction with the task regardless of the presence or absence of search expertise.*

Given the characteristics of our targeted tasks, where inconsistent information was found from different documents, users might feel dissatisfied when finding different answers. We hypothesized that search expertise played an important role in the relationship between answer consistency and task satisfaction. Our second hypothesis is summarized as follows:

² The term “search” will be omitted if it is clear from the context.

Table 5.3: Features related to search effort and found answers. Value on the left of each cell denotes mean (with standard deviation) for feature values, and that on the right denotes Pearson’s r (with p -value) between the feature and task satisfaction.

Feature	Experts				Non-Experts			
	mean		r with sat.		mean		r with sat.	
# of queries	3.06	(1.39)	−.266	(.303)	2.53	(1.23)	−.546	(.000)
# of pages	4.88	(2.23)	−.631	(.007)	4.02	(2.38)	−.072	(.625)
# of total answers	4.18	(3.13)	−.380	(.132)	3.53	(2.34)	.114	(.435)
# of unique answers	3.71	(2.76)	−.339	(.183)	3.02	(1.84)	.181	(.213)
Answer entropy	1.69	(0.85)	−.458	(.074)	1.59	(0.73)	−.021	(.896)

Hypothesis 2 (H2): *Only search experts show dissatisfaction with the task when finding many inconsistent answers.*

Procedure

We considered two effort features for the hypothesis **H1**: (1) the number of issued queries and (2) the number of browsed pages. As for the hypothesis **H2**, we designed three features to measure the answer consistency: (1) the total number of found answers, (2) the number of unique found answers, and (3) the entropy of found answers. These three features are calculated as follows: (1) $\sum_{a \in A} m(a)$, (2) $|A|$, and (3) $-\sum_{a \in A} \frac{m(a)}{\sum_{a' \in A} m(a')} \log_2 \frac{m(a)}{\sum_{a' \in A} m(a')}$, where A is a set of answers found by a user in a session of a certain task, and $m(a)$ denotes how many times an answer $a \in A$ was found by the user in the session. We calculated these features for each session in our dataset and used Pearson’s r between each feature and task satisfaction to test the two hypotheses.

Results

Table 5.3 lists the statistics for these features for task sessions grouped by search expertise. The value on the left of each cell in the table is the mean (with standard deviation) for the feature values, and that on the right is Pearson’s r (with p -value) between the feature and satisfaction. We will first describe the feature trend common in both user groups. We will then discuss the effect of search expertise on task satisfaction.

A general trend we found from Table 5.3 was that the effort features tended to be negatively correlated with task satisfaction irrespective of search expertise, which confirms our hypothesis **H1**. This result is consistent with the previous work that showed the negative effect of user effort on search performance [3, 32]. However, Table 5.3 also indicates another finding that the most

dominant effort feature depends on the presence or absence of search expertise; the number of pages significantly correlated with task satisfaction only for search experts ($r = -.631$, $p = .007$), while the number of queries was significant only for non-experts ($r = -.546$, $p < .001$). Search experts are, by our definition, familiar with the internal functions of search engines, such as query processing and document ranking. We conjecture that having such technical knowledge helps them to formulate effective search queries with less effort than others (as discussed by Hölscher and Strube [41]), which may reduce the load of querying and result in the low effect of the query effort feature.

In contrast to the effort features, which tended to be similar among the participants, we can observe from Table 5.3 that the answer features seem to affect task satisfaction differently depending on search expertise. The correlation between these features and task satisfaction tends to be negative for search experts. In particular, answer entropy showed a trend toward significance for them ($r = -.458$, $p = .074$). For non-experts, however, none of these features significantly correlates with task satisfaction. This result supports our hypothesis **H2** on the difference in satisfaction criteria between search experts and non-experts. A possible reason for the difference is their degree of trust in search engines. According to an online survey conducted by Nakamura *et al.* [65], many general users trust the ranking of search results to some extent despite insufficiently understanding the underlying mechanism of Web search engines, which suggests that search non-experts may have an over-reliance on top documents in the ranking. We infer that specialized knowledge on search engines prevents search experts from relying excessively on top documents and encourages them to examine the consistency of obtained answers by expending additional effort.

5.3.2 When do Users Finish Answer Searches?

The analysis in Section 5.3.1 suggested that search experts might expend more effort than non-experts to examine the answer consistency. Aiming at better understanding on how users expend effort to complete tasks, we formulated and tested the following hypothesis on time spent completing tasks:

Hypothesis 3 (H3): *Search experts take longer to complete tasks than non-experts.*

Procedure

To test the hypothesis above, we first analyzed how *completion time*, which is time spent completing individual tasks, differed between search experts and non-experts. In addition to completion time, the following two time-related measures were taken into account in our analysis to reveal in which phases users spent much time during their sessions: *discovery time*, which is time spent

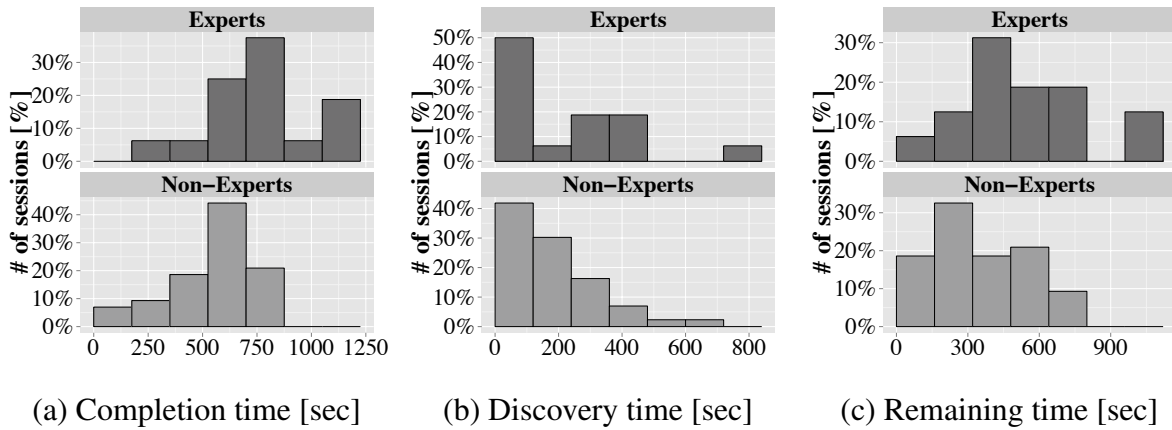


Figure 5.1: Histograms for number of sessions with respect to time spent finding answers.

finding the first answer; and *remaining time*, which is time elapsing from the first answer discovery to the task completion. Note that remaining time can be obtained by subtracting discovery time from completion time. Completion time has been reported to be negatively correlated with task satisfaction [93]. Shorter discovery time may indicate better search skills, while remaining time may reflect the degree to which users care about their found answers. We calculated these three measures for each session and compared the difference in the time distributions of each measure between search experts and non-experts. Welch's *t*-test was applied for testing whether two distributions were significantly different.

Results.

Figures 5.1a, 5.1b, and 5.1c show the histograms of task sessions with respect to completion time, discovery time, and remaining time, respectively. The top and bottom histograms in each figure represents the session distribution for search experts and search non-experts, respectively.

We can find from Figure 5.1a that search experts tended to spend longer completing the tasks than non-experts. More specifically, about half of the sessions by search experts lasted for more than 10 minutes, while less than one quarter of the sessions by non-experts lasted as long. Welch's *t*-test revealed that the completion time was significantly longer for search experts than non-experts ($t(24) = 3.29$, $p = .003$), which supported our hypothesis **H3**. We plot the histograms of discovery time in Figure 5.1b to better understand the reason for the search experts' long sessions. While search experts, on average, found the first answers faster than non-experts, Welch's *t*-test did not reveal a significant difference in the time distributions between these two groups ($t(20) = -.985$, $p = .337$). As shown in Figure 5.1c, our analysis on remaining time revealed that search experts tended to spend longer completing tasks than non-experts after finding the first answers, which was shown to be significant by Welch's *t*-test ($t(21) = 2.24$, $p = .036$).

Our main finding from the analysis above is that search experts tended to spend longer completing tasks than non-experts primarily due to the relative increase of remaining time. Cautious attitudes by search experts can be inferred from this result. That is, they did not seem to finish search tasks just by finding possible answers. Taking into account the expert-specific characteristic reported in Section 5.3.1 (*i.e.*, the negative correlation between answer entropy and task satisfaction), we conjectured that search experts spared more time and effort to find evidence that supported the answers they had found.

5.3.3 How do Users Change Targeted Answers?

The analysis results in Section 5.3.2 suggested the possibility that search experts spent a long time finding evidence that supported their found answers. To confirm whether this conjecture was correct, we investigated the change in answers targeted by users over time. The following hypothesis was tested in this investigation:

Hypothesis 4 (H4): *Search experts narrow down the search scope of answers as time passes.*

Procedure.

To grasp the temporal trends in found answers, we categorized them into the following two classes: *submitted answers*, which contained answers that users found and submitted as their responses for the post-task questionnaire; and *ignored answers*, which contained answers they found but did not submit. The ignored answers for a session were extracted by eliminating the submitted answers from all the answers found in the session. For each of these classes, we measured how the corresponding answers were found in the *early* and *closing* stages of sessions in the following manner. First, we treated a sequence of browsed pages, except SERPs, in each session in chronological order as the early stage and those in reverse chronological order as the closing stage³. Then, we calculated normalized Discounted Cumulated Gain (nDCG) [47] for the sequence of each stage by regarding pages with answers belonging to the class as relevant. Note that a session receives a high nDCG score when answers of the targeted class are frequently found in the targeted stage. We applied paired *t*-test to test whether nDCG scores significantly differed between the early and closing stages.

³ We also tried dividing browsed pages into the two stages with some empirically defined cutoffs, but these alternatives yielded similar results, which are omitted from this thesis.

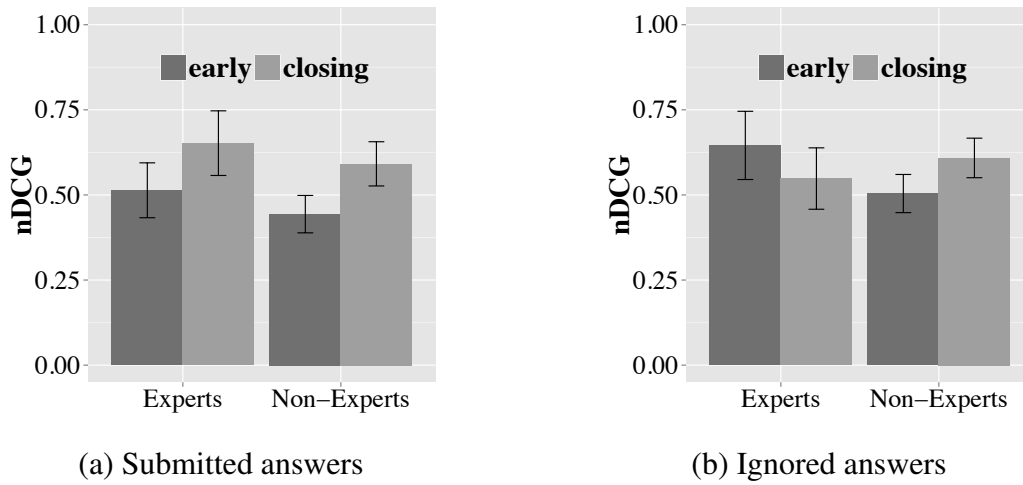


Figure 5.2: Changes in targeted answers over time.

Results.

We show the mean nDCG scores (with SEM bars) for the early and closing stages in Figure 5.2. The results for submitted answers are shown in Figure 5.2a while those for ignored answers in Figure 5.2b. The error bar on each bar chart represents the standard error of the mean.

As shown in Figure 5.2a, we observed a common trend between search experts and non-experts for submitted answers; the mean nDCG score tended to increase as time passed for both user groups (from .514 to .652 for search experts and from .444 to .591 for non-experts). These changes were shown by paired t -test to be an increasing trend toward significance for search experts ($t(16) = 1.75$, $p = .099$) and to be a significant increase for non-experts ($t(48) = 3.50$, $p = .001$). As for ignored answers, on the other hand, we observed different trends between search experts and non-experts as shown in Figure 5.2b; the mean nDCG score tended to decrease from .646 to .548 for search experts while tended to increase from .504 to .609 for non-experts. Applying paired t -test revealed the decrease for search experts as insignificant ($t(16) = -1.26$, $p = .225$) and the increase for non-experts as significant ($t(48) = 2.07$, $p = .044$). These results indicated that only search experts tended to search for documents with submitted answers more intensively than those with ignored answers at the closing stage of their sessions, which supports our hypothesis **H4**.

We inferred from these findings that search experts may attempt to find other sources that supported answers that they had found to date by narrowing down the scope of searches to these answers. This behavior could also be interpreted as a sign of their cautious attitudes to found answers, as was described in Section 5.3.2. In contrast, our results also suggested that search non-experts seemed to search for any kind of answer, irrespective of its class, even in the closing

stage of their sessions. The difference between search experts and non-experts may imply that search expertise affected whether and how users validated their found answers, which is discussed in more detail in Section 5.4.

5.4 DISCUSSION

We have shown that the absence of search expertise leads users (1) to judge task satisfaction mainly by search effort that they expend, (2) to finish search tasks without in-depth exploration, and (3) not to narrow down the targeted information to promising answers even in the closing stage of their task sessions.

5.4.1 Implications

Our findings suggest that search non-experts tended to be less concerned about the consistency in their found answers. Possible explanations for this trend are: (1) they were unaware of the existence of inconsistent answers; (2) they were unwilling to conduct in-depth exploration as additional effort would have been required; and (3) they intended to conduct in-depth exploration but failed due to their lack of knowledge or skills. To help search non-experts facing the situations above to properly handle inconsistent information, search engines first need to notify them of the existence of multiple instances as possible answers. However, just presenting all the instances may cause them to feel apprehensive about the real answer, similar to the way that search experts felt dissatisfied when finding diverse answers (Table 5.3). Such apprehension can be reduced by providing complementary information (*e.g.*, majority and typicality) for each answer. Ranking documents containing these answers is also important as many users have been shown to interpret search results as an ordering of likelihood [86]. In addition, search engines need to provide functions for in-depth exploration of each answer. It would be worth establishing techniques to retrieve evidence supporting a given answer and measure its reliability, which would help search non-experts to find the most credible answer with less apprehension.

5.4.2 Limitations

There are, at least, four limitations of our methodology that we should acknowledge. First, we had to use a small dataset comprising 66 task sessions for our analysis, given the effort required for answer extraction. As this dataset was sampled from search logs in a user study [27] designed for a different purpose, search experts and non-experts were not spread evenly across tasks. A purpose-built study with large, balanced samples would be needed to confirm our findings and gain more insight into answer-finding behavior. Second, the present work focused only on factoid questions to simplify our analysis. There are, however, other types of questions (*e.g.*, list and

definition [23]). It is also worth investigating the effect of search expertise on answer-finding behavior for such question types and identifying the differences from those for the factoid type. Third, two assessors were not native speakers of English, which was the language of Web pages from which they extracted answers that study participants had found. Although we used, given the inter-assessor agreement, a union of their extracted answers as the ground truth to avoid answers being missed, this issue may have affected our analysis. The last limitation is about our definition of search expertise. The present work defined search expertise as specialized knowledge on the internal workings of search engines. This definition may be too strict because users who have been using search engines for years are likely to have experienced search skills even if they do not have the specialized knowledge. Furthermore, while we analyzed the effect of search expertise on search behavior and outcomes, our analysis results might also be affected by domain expertise [85].

5.5 SUMMARY

This chapter investigated the effect of search expertise on search behavior and satisfaction criteria of users. Targeting four tasks comprising factoid questions with inconsistent answers, we analyzed answer-finding behavior of 30 study participants in terms of the presence or absence of search expertise. Our findings suggested that search non-experts tended to be less concerned about the consistency in their found answers. To help search non-experts to properly handle inconsistent information, search engines may need to notify them of the existence of multiple answer instances and establish techniques for retrieving evidence that supports a given answer.

Future work includes further investigating the effect of search expertise by conducting a purpose-built study with large, balanced samples. While we focused only on factoid questions in the present work, we are also interested in answer-finding behavior for other question types (*e.g.*, list and definition). It would also be worth exploring not only search expertise but also other user factors, like domain expertise, as possible factors affecting answer-finding behavior and search outcomes of users.

VISUALIZING THE AMOUNT OF MISSED INFORMATION IN SEARCH RESULTS

6.1 INTRODUCTION

Searchers often issue more than one query and browse multiple documents in exploratory search tasks [87]. These tasks can be characterized as open-ended and multi-faceted, requiring searchers to collect multiple relevant documents [88]. Many people conduct exploratory searches in the medical and health domains [14]. Take, as an example, a searcher who investigates the effect of smoking through Web searches in order to decide whether to continue smoking or not. This search topic contains various aspects such as diseases caused by smoking, increases in the price of cigarettes, and how smoking affects mental health. To fully understand this topic and make appropriate decisions, the searcher needs to collect extensive information covering different aspects. Such tasks are called *intrinsically diverse* search tasks in the literature [69].

There are several problems common to intrinsically diverse tasks. First, searchers often cannot easily come up with effective queries for collecting documents that cover diverse aspects. They have to issue more queries to complete the tasks if search engines return few documents relevant to unexplored aspects. Second, deciding when to stop searching is also difficult for searchers. Quitting the tasks too early without in-depth exploration prevents searchers from finding essential information; on the other hand, if they have already obtained extensive information, continuing these tasks for too long wastes time and effort, as they will not be able to acquire much more gain. The above problems are primarily due to the fact that (1) searchers may not know what aspects exist in the search topic and how important they are and (2) they cannot guess

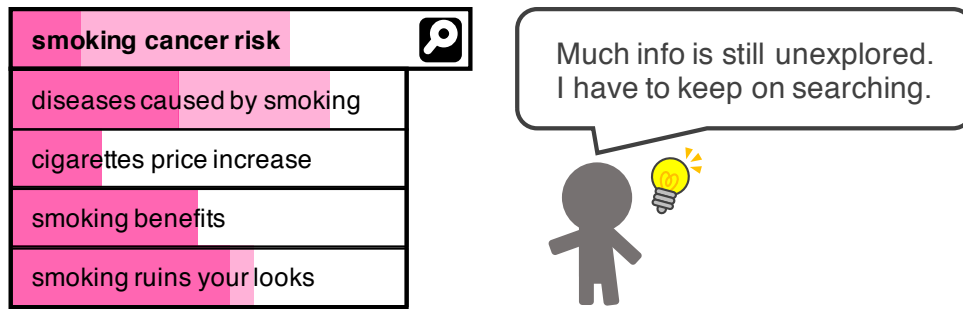


Figure 6.1: Our idea of visualizing missed information.

how much important information on each aspect is available on the Web and how much of the information remains unexplored.

In this work, we propose a query suggestion interface called ScentBar, with the goal of helping searchers discover unexplored aspects and find the appropriate timing for search stopping in intrinsically diverse tasks. As shown in Figure 6.1, ScentBar visualizes, for both the search query and suggestion queries, the amount of missed information¹ important to the current search topic in the form of a stacked bar chart so that users can grasp their search progress visually. We define the amount of missed information for a query as the additional gain that can be obtained from unclicked search results of the query. Gain is formalized as a set-wise metric based on aspect importance, aspect novelty, and per-aspect document relevance and is estimated by using a state-of-the-art algorithm [79] for subtopic mining and search result diversification.

We conducted a user study involving 24 participants to investigate how ScentBar affected their search strategies and search outcomes. Post-hoc analyses revealed that ScentBar had the following advantages when the gain estimation algorithm worked reasonably: (1) ScentBar users stopped examining search results after collecting a greater amount of relevant information; (2) they issued queries whose search results contained more missed information; (3) they obtained higher gain, particularly at the late stage of their sessions; and (4) they obtained higher gain per unit time. These results suggest that the simple query visualization helps make the search process of intrinsically diverse tasks more efficient, unless inaccurate estimates of missed information are visualized.

The main contributions of this work are as follows.

- We formalized an intent-aware metric for evaluating the gain that searchers acquire from a set of their collected documents and provided the explanation of its relation to other metrics for search result diversification.

¹ We use the phrase “missed information” to refer to information that has been retrieved by the system but that the searcher misses collecting, similar to the definition by Mansourian and Ford [61].

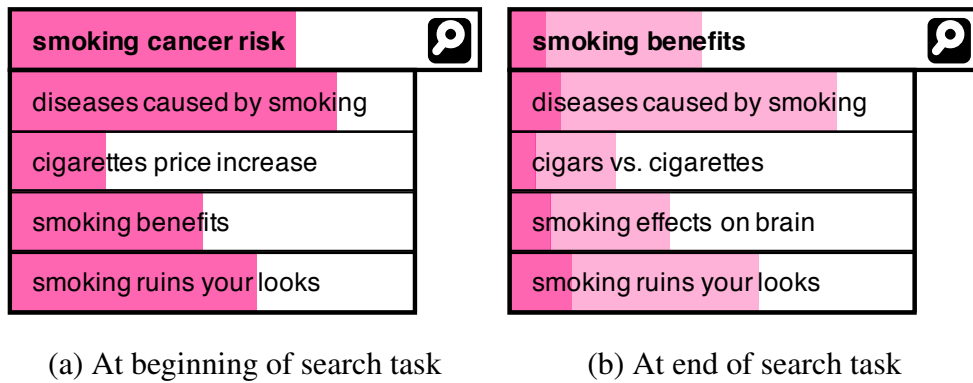


Figure 6.2: How ScentBar works. The state of missed information after the initial search is shown in Figure 6.1.

- We proposed ScentBar, which visualizes the amount of missed information estimated as the additional gain searchers can obtain, and conducted a user study involving 24 participants to investigate the effect of ScentBar on their search strategies and outcomes (*e.g.*, query formulation and search stopping).

6.2 PROPOSED INTERFACE

In this section, we illustrate how ScentBar works through an example and describe the research questions addressed in this work.

6.2.1 How It Works

When a user is typing a search query, ScentBar visualizes the amount of *missed information* for both the search query and suggestion queries in the form of a stacked bar chart. To be precise, missed information for a query represents information that (1) can be obtained from the search results of the query, (2) is important to the search topic, and (3) the user has not yet obtained.

Take the search topic described in Section 6.1 as an example again. Figure 6.2a illustrates the state of ScentBar visualization when a user starts his/her search task with the query “smoking cancer risk” to learn the effect of smoking. At the beginning of this task, the amount of missed information for a query can be interpreted as the total amount of important information that the user can obtain from the search results of this query. For example, the user can infer from Figure 6.2a that the queries “smoking cancer risk” and “diseases caused by smoking” return search results containing more important information than those such as “cigarettes price increase”.

When the user has obtained sufficient important information from the returned documents, the amount of missed information for the search query decreases, as shown in Figure 6.1. The

length of the bar for the search query (indicated in the dark color) is much shorter than its initial length (indicated in the light color), which suggests this query contains only a small amount of missed information. Note that the bar for the suggestion query “diseases caused by smoking” also has a short length compared to the initial state. This is because the suggestion query shares with the search query a certain amount of important information. In contrast, the visualized bars for other suggestion queries (e.g., “smoking ruins your looks”) remain almost unchanged from the initial ones, which suggests great amounts of important information are still unexplored for these queries.

When the user has exhaustively collected important information on the search topic from a variety of angles through iterative searches, the amount of missed information considerably decreases for any query related to the topic. Figure 6.2b illustrates the state of ScentBar visualization at the end of the search task. As suggested in the figure, there is little important information left for either the search query or the suggestion queries.

6.2.2 Research Questions

In this work, we investigate the effect of displaying missed information on searchers’ strategies and outcomes. More specifically, we address the five research questions listed below.

In intrinsically diverse tasks, searchers are more likely to issue multiple queries to exhaustively collect relevant information from different angles. Thus, we frame the following two research questions on individual searches (*i.e.*, after a query is issued before another one is issued or the session ends):

RQ1 *How does ScentBar affect users’ decisions on when to stop the current search?*

RQ2 *How does ScentBar affect users’ decisions on which query to use for the next search?*

We expect our visualization enables searchers to utilize more rational strategies in individual searches. For example, ScentBar users may be able to stop the current search after collecting a sufficient amount of important information by monitoring the bar visualization for the search query. Missed information may also affect searchers’ query formulation strategies; ScentBar users may use a more effective query for the subsequent search by comparing the visualized bars for suggestion queries.

The remaining three research questions relate to the overall search sessions of intrinsically diverse tasks:

RQ3 *How does ScentBar affect the temporal change in gain that users acquire through their search process?*

RQ4 *How does ScentBar affect users' decisions on when to stop their task sessions?*

RQ5 *How does ScentBar affect the relationship between the effort that users expend and the gain that they obtain?*

Missed information may affect searchers' strategies and outcomes at the session level as well as at the query level. For example, ScentBar users may acquire high gain at any point in their sessions. Analogous to **RQ1**, ScentBar may also enable users to rationally decide when to stop the task sessions². It can also be hypothesized that ScentBar makes users' search processes more cost-effective: they may be able to collect a sufficient amount of important information from a variety of angles without expending much effort.

6.3 MISSED INFORMATION ESTIMATION

In this section, we first introduce *gain*, a metric for evaluating search outcomes in intrinsically diverse tasks, and then define the amount of *missed information* by using the gain metric. We then describe our algorithm of estimating the gain components.

6.3.1 Gain

Let us consider how searchers can obtain gain in intrinsically diverse tasks. As described in Section 6.1, these tasks require searchers to collect extensive information covering a variety of different aspects. Considering this characteristic, it would be natural to assume that the gain they obtain through their searches is independent of their browsing order of documents. Thus, we formalize gain as a set-wise metric, whose relation to other metrics is discussed in Section 6.3.2. We also derive from this characteristic the following requirements that the gain metric should satisfy:

Importance

Documents relevant to a central aspect of the search topic produce higher gain than those relevant to a peripheral one.

Relevance

Highly relevant documents produce higher gain than partially relevant ones.

Novelty

Documents relevant to an unexplored aspect produce higher gain than those relevant to a fully explored aspect.

² As in the literature [63], we refer to stopping behavior regarding **RQ1** as *query stopping* and that regarding **RQ4** as *session stopping*. *Search stopping* is used as the generic phrase for these two.

First, we decompose the gain metric from the topic level to the aspect level in an intent-aware manner [2]. More specifically, given a topic t , we formulate the topic-level gain $\text{Gain-IA}_t(D)$ that can be obtained from a set of documents $D = \{d_1, d_2, \dots\}$ as

$$\text{Gain-IA}_t(D) = \sum_{a \in A_t} \Pr(a | t) \cdot \text{Gain}_a(D),$$

where A_t is a set of aspects for t , $\Pr(a | t)$ is a probability mass function representing the importance of an aspect a to t , and $\text{Gain}_a(D)$ is the per-aspect gain that can be obtained from the documents D with respect to a . To satisfy the first requirement, the above formula puts greater value on the per-aspect gain for highly important aspects when calculating the topic-level gain.

Next, the per-aspect gain is calculated by

$$\text{Gain}_a(D) = \sum_{i=1}^{|D|} \text{Rel}_a(d_i) \cdot \text{Disc}_a(\{d_1, \dots, d_{i-1}\}),$$

where $\text{Rel}_a(d) \in [0, 1]$ denotes the degree of relevance of a document d to the aspect a . To satisfy the second requirement, this formula sums up the per-aspect relevance degrees of individual documents and returns a high value if D contains many documents that are highly relevant to a . Note that $\text{Gain}_a(\emptyset) = 0$.

The other term $\text{Disc}_a(\cdot)$ in the above formula is a function designed to satisfy the third requirement; it discounts the degree of per-aspect document relevance $\text{Rel}_a(d_i)$ when the aspect a is well covered by a set of documents $\{d_1, \dots, d_{i-1}\}$ that have already been browsed. We define the discount function as

$$\text{Disc}_a(D') = \prod_{d'_j \in D'} (1 - \text{Rel}_a(d'_j)).$$

When D' contains many documents that are highly relevant to the aspect a , the above formula returns a value of nearly zero. In this case, newly browsed documents relevant to a receive high discounts from this formula and therefore contribute little to the per-aspect gain for a . Note that $\text{Disc}_a(\emptyset) = 1$.

6.3.2 Relation of Gain to Others

Gain-IA, the intent-aware gain metric defined above, apparently looks similar to the intent-aware version of Expected Reciprocal Rank (ERR-IA) [15, 16], which is an evaluation metric for diversified search results. ERR-IA evaluates $R_q = (d_1, d_2, \dots)$, a list of ranked documents for a query q , with the following formula:

$$\sum_{a \in A_q} \Pr(a | q) \sum_{i=1}^{|R_q|} \frac{\text{Rel}_a(d_i)}{i} \prod_{j=1}^{i-1} (1 - \text{Rel}_a(d_j)).$$

The only difference between ERR-IA and Gain-IA is that the latter does not have the reciprocal rank factor (*i.e.*, i^{-1}). As ERR-IA evaluates the effectiveness of *ordered* search results, it discounts the value of lower-ranked documents. Unlike ERR-IA, our objective is to evaluate the gain that searchers can obtain from *unordered* documents, separately from their expended effort. Thus, we designed Gain-IA to be a set-wise metric that is affected by neither the rank of nor the browsing order of documents. This is obvious from the fact that Gain-IA has the following equivalent form:

$$\text{Gain-IA}_t(D) = \sum_{a \in A_t} \Pr(a | t) \cdot \left[1 - \prod_{d \in D} (1 - \text{Rel}_a(d)) \right],$$

which is derived from the following lemma:

LEMMA 1. $\text{Gain}_a(D) = 1 - \text{Disc}_a(D)$.

PROOF. Let $D^{(n)}$ be a set of n documents (*i.e.*, $\{d_1, \dots, d_n\}$) and $P(n)$ be the statement of Lemma 1 for $D^{(n)}$ (*i.e.*, $\text{Gain}_a(D^{(n)}) = 1 - \text{Disc}_a(D^{(n)})$). We prove by mathematical induction that $P(n)$ holds for every non-negative integer n .

When $n = 0$, $P(0)$ clearly holds from the definitions of $\text{Gain}_a(\cdot)$ and $\text{Disc}_a(\cdot)$ for $D^{(0)} = \emptyset$.

Assuming $P(k)$ holds for some non-negative integer k , $P(k+1)$ also holds for $k+1$ from the following deformation:

$$\begin{aligned} \text{Gain}_a(D^{(k+1)}) &= \text{Gain}_a(D^{(k)}) + \text{Rel}_a(d_{k+1}) \cdot \text{Disc}_a(D^{(k)}) \\ &= 1 - \text{Disc}_a(D^{(k)}) + \text{Rel}_a(d_{k+1}) \cdot \text{Disc}_a(D^{(k)}) \\ &= 1 - \text{Disc}_a(D^{(k)}) \cdot (1 - \text{Rel}_a(d_{k+1})) \\ &= 1 - \text{Disc}_a(D^{(k+1)}). \end{aligned}$$

Therefore, $P(n)$ is true for every non-negative integer n . □

Note that the rewritten Gain-IA is exactly the same as the objective function formulated by Agrawal *et al.* [2] for search result diversification. Assuming that $\text{Rel}_a(d)$ represents the probability that a document d satisfies a searcher who believes an aspect a is as important to a topic t as $\Pr(a | t)$, this function can be viewed as the probability that the searcher gains satisfaction with the topic t by browsing a set of documents D .

6.3.3 Missed Information

Using the above-mentioned gain metric, we define the amount of *missed information* for a query as the additional gain that can be obtained from unclicked search results of the query. Let D_u be

a set of documents that a user u has already browsed in the current task on a topic t and D_q^K be a set of top- K documents returned for a query q . Then, $MI_{u,t}(q)$, the amount of missed information for q , is defined as

$$MI_{u,t}(q) = \text{Gain-IA}_t(D_u \cup D_q^K) - \text{Gain-IA}_t(D_u).$$

Under situations where the above formula behaves ideally, a high value of $MI_{u,t}(q)$ indicates that unclicked documents $D_q^K \setminus D_u$ contains a large amount of information important to the topic t that the user u misses collecting from the search results of the query q .

6.3.4 Gain Components

The following three components are required to calculate the gain and the amount of missed information on a topic t : (1) A_t , a set of aspects for t ; (2) $\Pr(a | t)$, a probability representing the importance of an aspect a to t ; and (3) $\text{Rel}_a(d)$, the degree of relevance of a document d to a . As for the parameter K (*i.e.*, the number of search results to be fetched), which is also required to calculate the amount of missed information, we report the value used in our experiment in Section 6.4. We describe below our algorithm of estimating these three components. Note that, if the topic is known beforehand, these components can be estimated before ScentBar users conduct their searches.

As overviewed in Section 2.1.2, existing approaches to explicitly mining subtopics have the same task of estimating the above components. Thus, we decided to use one of the state-of-the-art explicit subtopic mining algorithms developed by Tsukuda *et al.* [79] to perform this estimation task. Their algorithm, which is easy to implement, achieved the second-best performance among 14 submissions in the Subtopic Mining subtask of the NTCIR-10 INTENT-2 task [70]. In what follows, we describe the outline of the algorithm and clarify some differences from theirs³. As described hereinafter, this algorithm relies on a search engine to fetch search results for given queries. The search engine for the algorithm is identical to the one for our search interface (see Section 6.4 for more details).

Topic Aspects

To estimate a set of aspects A_t for a given topic t , we first mine a set of subtopics S_t for t using three resources, following Tsukuda *et al.* [79]. The first resource consists of query suggestions (*i.e.*, related queries and auto-completion queries) returned by Web search engines in response to the topic query t . Each suggested query is regarded as a subtopic of t . The second resource is query logs, from which queries starting from t are extracted as subtopics. The last resource

³ We use parameter values reported in the original paper [79].

is the search results for the topic query. The clustering algorithm proposed by Zeng *et al.* [98] is applied to the search results for extracting key phrases as subtopics from individual clusters. Once S_t is mined in this way, Ward's method [82] is applied to S_t for obtaining a set of subtopic clusters $\mathcal{C}_t$, each of which is regarded as belonging to an aspect of the topic t .

As described in Section 6.4, the search topics used in our experiment were selected from tasks in past NTCIR workshops. Thus, we utilized the resources distributed to the task participants to mine subtopics for these topics.

Aspect Importance

To estimate the importance probability of an aspect of a topic t , we first estimate $\text{Imp}_t(s)$, the importance of a subtopic s to t , by using $\text{Imp}_t(s) = \sum_{d \in D_s^N \cap D_t^N} 1/\text{Rank}_t(d)$, as with Tsukuda *et al.* [79], where $\text{Rank}_t(d)$ denotes the rank of a document d returned for the topic query t . Next, we select a representative subtopic a from each cluster $C \in \mathcal{C}_t$ by using $a = \arg \max_{s \in C} \text{Imp}_t(s)$. The selected subtopic a is regarded as a member of the aspect set A_t for the topic t . Finally, we estimate the importance probability $\Pr(a | t)$ of each aspect $a \in A_t$ to t by

$$\Pr(a | t) = \frac{\text{Imp}_t(a)}{\sum_{a' \in A_t} \text{Imp}_t(a')}.$$

Per-Aspect Document Relevance

The algorithm proposed by Tsukuda *et al.* [79] performs the relevance estimation only for top- N documents returned for each aspect query $a \in A_t$. This is probably because their objective is to obtain diversified search results comprising about ten documents. However, searchers are expected to browse many more documents in intrinsically diverse tasks. To obtain more accurate estimates of missed information, the relevance should be estimated for as many related documents as possible. Thus, we expand the target of the relevance estimation with respect to an aspect a to $D_a = \bigcup_{s \in C_a} D_s^N$, where $C_a \in \mathcal{C}_t$ is a cluster of subtopics belonging to a . In light of the subtopic importance, we estimate $\text{Rel}_a(d)$, the relevance degree of a document $d \in D_a$ to an aspect $a \in A_t$, by

$$\text{Rel}_a(d) = \frac{\sum_{s \in C_a} \text{Imp}_t(s) \cdot \text{Rel}_s(d)}{\sum_{s \in C_a} \text{Imp}_t(s)},$$

where $\text{Rel}_s(d)$ is d 's relevance degree to a subtopic s and is estimated by $\text{Rel}_s(d) = 1/\sqrt{\text{Rank}_s(d)}$, as with Tsukuda *et al.* [79].

6.4 EXPERIMENTAL DESIGN

We conducted a user study in a laboratory setting to investigate the research questions listed in Section 6.2.2. The details of the experimental design are described below.



Figure 6.3: Screenshot of ScentBar (topic: “global warming”).

6.4.1 Interfaces

We designed the type of interfaces was to be a within-subjects factor in our user study, where we compared two search interfaces: (1) *w/ scent*, *i.e.*, the proposed interface ScentBar, which visualizes missed information for individual queries; and (2) *w/o scent*, *i.e.*, a baseline interface without missed information visualization.

Figure 6.3 shows the screenshot of ScentBar for the sample search topic “global warming”. Note that the baseline interface has the same appearance as ScentBar except for the query visualization. The brief description of the current search topic is shown at the top of these interfaces. While a user is typing a query into the search box (at the upper left), the interfaces obtain auto-completion queries through the Google Suggest API⁴ and show them at the bottom of the search box. ScentBar also visualizes missed information for these queries in the form of a stacked bar chart. The suggested auto-completion queries remain at the same place even after the user issues the current query to our search system so that he/she can easily issue them for future searches. The search results returned for the query are shown on the right side of the interfaces. When

⁴ <http://www.google.com/complete/search>

the user clicks the title of a search result, a document viewer is displayed within the interfaces and the user can browse the landing page on the viewer. The interfaces also have a feature for bookmarking Web pages displayed on the document viewer and present the information on which search results have already been clicked and bookmarked for the current query (on the right side of each search result) and on how many documents have already been clicked and bookmarked through the current search task (at the bottom of the interfaces).

6.4.2 Search System

Both interfaces described above used the same internal search system to fetch search results in response to queries. We built the search system on Apache Solr and indexed documents in ClueWeb09-JA, the Japanese portion of the ClueWeb09 collection⁵, into this system. We utilized the Okapi BM25 algorithm with default Solr parameters (*i.e.*, $k_1 = 1.2$ and $b = 0.75$) to rank documents.

When testing our implemented interfaces, however, we found that near-duplicate documents sometimes occupied high positions in the search results, which would prevent users from exhaustively collecting information from multiple aspects. To deal with this problem, we decided to diversify the search results by applying the MMR algorithm [13] to R_q^K , the top- K documents retrieved by BM25 for a query q . More precisely, MMR selects d_k , a document ranked at the k -th position, by using

$$d_k = \arg \max_{d \in R_q^K \setminus S^{k-1}} \left[\lambda \cdot \text{Rel}_q^{\text{BM25}}(d) - (1 - \lambda) \max_{d' \in S^{k-1}} \text{Sim}(d, d') \right],$$

where S^{k-1} denotes a set of $k - 1$ documents that MMR has already selected and $\text{Rel}_q^{\text{BM25}}(d)$ returns a document d 's relevance score⁶ that BM25 calculates for a query q . As for $\text{Sim}(\cdot, \cdot)$, we used the cosine similarity between tf-idf vectors, each of which was constructed from a set of terms appearing in the title and snippet areas.

In this way, the resulting list of top- K diversified documents D_q^K was returned to the experimental interfaces in response to a query q . The control parameter was set to $\lambda = 0.3$, following Tsukuda *et al.* [79] and Dou *et al.* [25]. The number of search results to be fetched was set to $K = 100$, following Qvarfordt *et al.* [68].

6.4.3 Search Topics

As the topic source, we decided to use the NTCIR INTENT-1 [76], INTENT-2 [70], and IMine-1 [59] tasks, where the aspects of each topic and thier importance are provided for evaluating

⁵ <http://www.lemurproject.org/clueweb09.php>

⁶ We normalized $\text{Rel}_q^{\text{BM25}}(\cdot)$ to $[0, 1]$ by using the BM25 score of a document ranked at the first position for a query q .

Table 6.1: Search topics (with oracle aspects) used in our user study (translated from Japanese).

ID	Topic	Aspects
T1	symptoms of diabetes	causes, early symptoms, terminal illness, subjective symptoms, prevention, complicating diseases, drowsiness, . . .
T2	clinical depression	symptoms, work, society, communication, treatment
T3	dress codes for wedding ceremony	males, females, families and relatives, hairstyles, after-party, invited guests, wedding reception, . . .
T4	dinosaurs	fossils, extinction, pictorial books, species, museums, Jurassic period, images, . . .

search result diversification. We selected four topics that satisfied the characteristics of intrinsically diverse tasks (*i.e.*, having diverse documents relevant to different aspects). These topics are listed in Table 6.1. Two of them were selected from the medical domain, in which people often conduct exploratory searches to collect diverse information from multiple sources [14], while the remaining two were from non-medical domains. Table 6.1 also shows a subset of *oracle* aspects for these topics, which were provided by the NTCIR task organizers. As shown in this table, there are a number of different aspects in these topics (the mean number of oracle aspects per topic was 10.50), and multiple relevant documents may be needed to cover each aspect.

We used these topics to design four types of intrinsically diverse search tasks. More specifically, the following description was given to the participants of our user study for the search task on a topic t :

You were given the assignment of submitting a thorough report on the topic t . To fully understand t , collect relevant information on this topic from a number of different aspects that you think is important. You may end this search task when you feel there is little important information left.

6.4.4 Procedure

We recruited study participants via the recruiting Web site of our university. As a result, we had 24 participants (4 females and 20 males) in our user study: 16 undergraduate students, 6 graduate students, and 2 researchers. Their mean age was 23.29 (SD = 3.67) and their fields of study included informatics (six participants), engineering (four), law (four), science (three), *etc.* At

the end of the experiment, they received bookstore gift cards (equivalent to \$16 per person) as a reward for participating. Prior to the actual tasks, participants were assigned a training task (on the topic of global warming) to get familiar with the tasks and interfaces. During the training task, we provided participants with detailed instructions on how to use the interfaces and on what they were expected to do in our experiment. The training task took approximately 15 minutes.

A demographic questionnaire was administered at the beginning of the experiment. After filling out the questionnaire, participants worked on the four tasks one by one (two tasks with ScentBar and two with the baseline interface). The order of interfaces and tasks was determined with a Graeco-Latin square to remove the ordering effect of these variables. All tasks followed the same procedure. First, participants read the task description and completed a pre-task questionnaire asking about their knowledge on the topic and their estimated difficulty of the task. Next, they started the task with the assigned interface. To make the search process as realistic as possible, they were allowed to freely formulate search queries and were asked to use the bookmark feature provided by both interfaces when they found documents useful for completing the task. We did not allow them to follow links in the landing page because we could not guarantee that the linked pages were stored in ClueWeb09-JA, the document corpus used by our search system. After completing the task, participants filled out a post-task questionnaire asking about task-level feedback (*e.g.*, their experienced difficulty). Finally, we administered an exit questionnaire at the end of the experiment to gather feedback on both interfaces.

As mentioned above, we asked the participants to finish each task when they felt they had exhaustively collected relevant information from various aspects of the search topic. We did not tell them the total number of tasks so as to avoid its effect on their stopping decisions, since they knew in advance from the recruitment information that the maximum amount of time required for our experiment was two hours. Instead, we instructed them that there were a certain number of tasks and that the experiment would finish when they had completed all tasks or two hours had passed from the beginning of the experiment. Note that, in a pilot study, every task was completed within 20 minutes. In fact, all participants completed the whole experiment within two hours.

6.5 RESULTS

This section reports the results of our user study and attempts to answer the research questions listed in Section 6.2.2. As normality was not guaranteed for the experimental data, we used non-parametric significance tests in our post-hoc analyses. Significant effects are reported on the significance level $\alpha = 0.05$.

6.5.1 Ground-Truth Data

To evaluate gain that participants obtained, we prepared the ground-truth data for each search topic used in our user study. As described in Section 6.3.4, the following three components are required to calculate the gain and the amount of missed information: (1) a set of aspects for the search topic, (2) the importance of each aspect, and (3) per-aspect document relevance. As these topics were selected from the NTCIR tasks, we used oracle data provided by the task organizers as for the first two components.

Relevance Assessment

To prepare the remaining ground-truth data (*i.e.*, per-aspect document relevance), three assessors judged the relevance of each document that had been browsed by participants. A four-grade scale was used to assess the relevance of a document for each oracle aspect: irrelevant (= 0), partially relevant (= 1), highly relevant (= 2), and perfect (= 3). The following criteria were shared between assessors to make the assessed relevance as consistent and reliable as possible: documents covering (1) less than 30%, (2) at least 30% but less than 60%, (3) at least 60% but less than 90%, and (4) more than 90% of the information on a certain aspect should be respectively labeled as (1) irrelevant, (2) partially relevant, (3) highly relevant, and (4) perfect for that aspect. The resulting relevance grade $g_{a,d} \in \{0, 1, 2, 3\}$ of a document d for an aspect a was converted into the oracle relevance degree by

$$\text{Rel}_a^*(d) = \frac{2^{g_{a,d}} - 1}{2^3}.$$

This conversion is often used when calculating evaluation metrics, such as ERR [16], and guarantees that the resulting relevance degree is any of $\{0, 0.125, 0.375, 0.875\}$.

Accuracy of Gain Estimation

The amount of missed information is defined as the additional gain, which is estimated by the algorithm mentioned in Section 6.3.4. If the estimated gain is far different from the oracle one, our interface ends up showing searchers missed information that is unrealistic and unreliable, which may have a negative effect on their searches. Thus, we decided to investigate the gap between the gain estimated by our algorithm and that calculated using the ground-truth data, before conducting post-hoc analyses. The former is referred to as *estimated gain* and the latter as *oracle gain* hereinafter.

To this end, we calculated the estimated and oracle gain scores at each point where participants browsed documents and measured correlations for pairs of gain scores using Pearson's r , Spearman's ρ , and Kendall's τ . Note that strong correlation for a topic indicates high consistency

Table 6.2: Correlation coefficients between oracle and estimated gains. The top three values for each measure are shown in bold face.

Topic	Pearson's r	Spearman's ρ	Kendall's τ
T1	0.834	0.851	0.683
T2	0.845	0.860	0.678
T3	0.824	0.824	0.713
T4	0.710	0.702	0.529

between the change in estimated gain and that in oracle gain, which suggests the gain algorithm works reasonably for that topic. Table 6.2 shows the values of these three correlation measures calculated for each topic.

As shown in this table, quite high correlation was observed for all topics except T4 (“dinosaurs”), which suggests that ScentBar users who conducted searches on the topic T4 might have experienced unexpected changes in visualized missed information. When analyzing the possible causes of the inaccurate gain estimation for T4, we found that aspects mined by our algorithm as important to this topic included “merchandising sales” and “skeleton models selling”. As can be seen in Table 6.1, none of these were contained in the oracle aspects for this topic. We also found that many documents retrieved by the query “dinosaurs” at high positions tended to be less relevant to the oracle aspects. As described in Section 6.3.4, the gain estimation algorithm relies on aspect mining and document ranking. Thus, the aspect mismatch and poor document ranking would cause the estimated gain to be inaccurate for the topic T4.

As noted above, inaccurate gain estimation could have an undesirable effect on participants’ fulfilling search tasks; they might get confused by seeing unreliable missed information and quit searching at an inappropriate point. To investigate the effect of the accuracy of the estimated gain, the analyses reported in the remainder of this section were conducted at the following three topic levels: (1) all topics; (2) *HC topics*, which comprise T1, T2, and T3, where the estimated gain had high correlation with the oracle one; and (3) *LC topic*, which comprises T4, where low correlation was observed between the estimated gain and the oracle one.

6.5.2 Query Stopping

To answer **RQ1**, we investigated the effect of missed information on query stopping (*i.e.*, the point at which a user stops examining search results for the current query). We hypothesized that ScentBar users would stop the current search once they had exhaustively collected relevant documents from the search results. If this is the case, query stopping should occur when the

Table 6.3: Mean change (with SD) in oracle missed information at the query level. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.

	All topics		HC topics		LC topic	
	w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
ΔMI	0.109 (0.110)	0.138 (0.134)	0.128 (0.120)	0.186 (0.147)	0.069 (0.073)	0.071 (0.072)

amount of missed information for the query greatly decreases from its initial value. Thus, for each query issued by participants, we calculated the change in the amount of oracle missed information between the beginning of the task and the query stopping point. A large change indicates that the searcher obtained a great amount of relevant information from the current search.

Table 6.3 shows the mean change (with SD) in the amount of oracle missed information at the query level. As for all topics, the amount of oracle missed information decreased greatly when participants used ScentBar. The difference from the baseline interface was shown to be significant by the Mann-Whitney U test ($p = 0.001$). The greater difference was observed for the HC topics ($p < 0.001$). In contrast, the change in oracle missed information was quite small for the LC topic ($p = 0.683$). These results indicate that when gain was estimated reasonably accurately, ScentBar users collected a greater amount of relevant information through individual searches than the baseline interface users, which supports our hypothesis on query stopping. When inaccurate estimates were presented to participants, however, ScentBar had no effect on the amount of collected information at the query level.

6.5.3 Query Selection

To answer **RQ2**, we investigated the effect of missed information on users' query selection decisions in the subsequent search. We hypothesized that ScentBar would help users issue queries whose search results contained much missed relevant information. To test this hypothesis, we calculated, for each query issued by participants, the amount of oracle missed information on query issuing.

Table 6.4 shows the mean (with SD) of the amount of oracle missed information on query issuing. On average, queries with large amounts of oracle missed information were issued through ScentBar rather than through the baseline interface for all topics. The Mann-Whitney U test revealed a significant difference between these interfaces ($p = 0.002$). While a clearer trend was observed for the HC topics ($p < 0.001$), the difference was insignificant for the LC topic

Table 6.4: Mean (with SD) of oracle missed information on query issuing. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.

	All topics		HC topics		LC topic	
	w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
MI	0.211 (0.194)	0.241 (0.191)	0.238 (0.214)	0.298 (0.210)	0.155 (0.126)	0.162 (0.123)

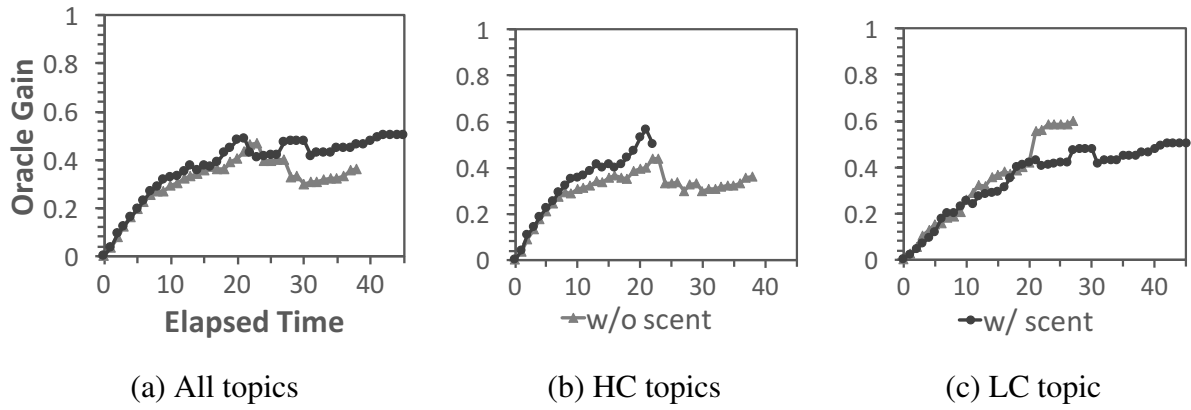


Figure 6.4: Mean change in oracle gain over time.

($p = 0.521$). These results indicate that when the gain was estimated reasonably, ScentBar users issued queries more effective for exhaustively collecting relevant information than the baseline interface users. This supports our hypothesis on query selection. When the gain estimation was inaccurate, however, queries issued through the two interfaces differed little from one another in terms of the amount of oracle missed information on query issuing.

6.5.4 Gain Change

To answer **RQ3**, we investigated the effect of missed information on the change in gain over time. More specifically, we split participants' task sessions into one-minute segments and calculated the (cumulative) oracle gain they had obtained at each cutoff time point. Note that when calculating the gain at a time point, we excluded participants who had completed the task before that time.

Figure 6.4 shows the gain curves averaged over the individual interfaces. The two gain curves for all topics (Figure 6.4a) behaved similarly until seven minutes had passed from the task initiation. After that, the oracle gain for ScentBar outperformed that for the baseline interface at nearly every time point. The difference between the interfaces widened for the HC topics (Figure 6.4b) while the gain curve for ScentBar underperformed that for the baseline interface at most time

Table 6.5: Top: mean change (with SD) in oracle missed information at the session level. Bottom: mean (with SD) of oracle gain obtained through task sessions. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.

	All topics		HC topics		LC topic	
	w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
ΔMI	0.162 (0.132)	0.195 (0.151)	0.188 (0.143)	0.256 (0.158)	0.106 (0.080)	0.110 (0.086)
Gain	0.358 (0.132)	0.404 (0.147)	0.370 (0.133)	0.415 (0.150)	0.322 (0.129)	0.371 (0.138)

points (Figure 6.4c). These results indicate that when the gain estimation was accurate enough, ScentBar users obtained as much gain at an early time of their task sessions as, and higher gain after that time than, the baseline interface users. In contrast, when ScentBar presented inaccurate estimates to participants, they obtained less gain than the baseline interface users who spent the same amount of time or more searching.

6.5.5 Session Stopping

To answer **RQ4**, we investigated the effect of missed information on session stopping (*i.e.*, the point at which a user completed the current search task). As in the case of our analysis on query stopping (reported in Section 6.5.2), we hypothesized that ScentBar users would complete the current task after exhaustively collecting information relevant to the search topic. If this is the case, session stopping should occur when the amount of oracle missed information for most queries related the task greatly decreases from the initial values. Thus, for each query issued by participants, we calculated the change in the amount of oracle missed information between the beginning and end of the task. In addition, we also calculated the (total) oracle gain obtained from the documents that individual participants had collected through their task sessions.

In Table 6.5, the mean change (with SD) in the amount of oracle missed information at the session level is shown at the top, and the mean (with SD) of the oracle gain obtained through task sessions is at the bottom. As for the former measure, the amount of oracle missed information decreased greatly for all topics when participants used ScentBar. The Mann-Whitney U test revealed a significant difference from the baseline interface ($p = 0.002$). The larger difference was observed for the HC topics ($p < 0.001$), while there was no significant difference for the LC topic ($p = 0.845$). As for the latter measure, the mean oracle gain for ScentBar was higher for all

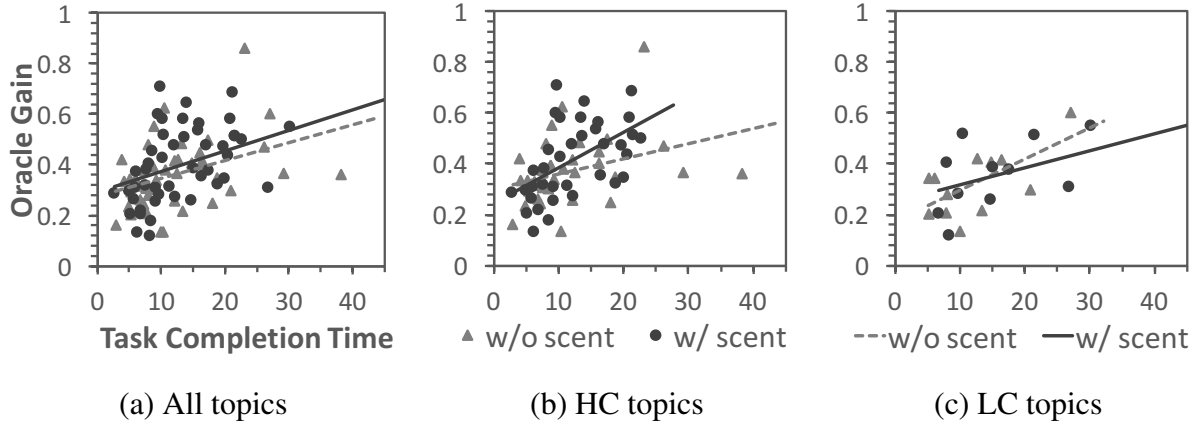


Figure 6.5: Relationship between effort and gain.

topics than that for the baseline interface. While the same trend was observed for both the HC and LC topics, we could not find any significant differences between the two interfaces ($p = 0.097$ for all topics, $p = 0.191$ for the HC topics, and $p = 0.410$ for the LC topic). These results partially support our hypothesis on session stopping: when the gain estimation algorithm worked reasonably, ScentBar users completed their tasks after collecting greater amounts of relevant information from the SERPs of their issued queries. However, the resulting total gain obtained through ScentBar did not show statistically significant improvement compared to the baseline interface, which implies some users of ScentBar might complete their tasks earlier than others who used the baseline interface. The analysis related to this issue is reported in Section 6.5.6.

6.5.6 Relationship Between Effort and Gain

To answer **RQ5**, we investigated the effect of missed information on the relationship between expended effort and acquired gain. In this work, the time that participants spent completing the tasks was regarded as the effort that they expended. We plotted scatter charts in which the x -axis represented the task completion time (in minutes) and the y -axis represented the oracle gain that participants obtained through their task sessions. We also performed a linear regression analysis to better understand the relationship between the effort (modeled as the explanatory variable) and the gain (as the response variable).

Figure 6.5 shows the scatter charts and regression lines of the effort-gain data. As can be seen in Figure 6.5a, little difference existed between the two interfaces for all topics. In fact, the slope β of both regression lines was almost the same: $\beta = 0.008$ ($R^2 = 0.187$) for ScentBar and $\beta = 0.007$ ($R^2 = 0.158$) for the baseline interface. As for the HC topics (Figure 6.5b), the great improvement was observed for ScentBar ($\beta = 0.014$, $R^2 = 0.281$), while the slope was nearly unchanged for the baseline interface ($\beta = 0.006$, $R^2 = 0.120$). In contrast, for the LC topic

(Figure 6.5c), the slope of the regression line for ScentBar ($\beta = 0.007$, $R^2 = 0.310$) was lower than that for the baseline interface ($\beta = 0.012$, $R^2 = 0.416$). These results indicate that when the gain estimation was accurate enough, ScentBar users obtained higher gain per unit time than the baseline users. However, the former users could not obtain gain as effective as the latter ones when ScentBar displayed inaccurate estimates. Figure 6.4 also demonstrates that some ScentBar users completed tasks earlier without expending much effort while others spent longer to obtain much more gain.

6.5.7 Search Interaction

We analyzed participants' interaction data to investigate how ScentBar affects their search behavior in intrinsically diverse tasks. To investigate its effect on query selection, we calculated the number of queries issued by participants (#Queries) and the fraction of suggestion queries selected by them (%SuggQueries). As document-related measures, we calculated the number of, precision of, and recall of documents browsed by participants (#Docs, Precision, and Recall, respectively) and the fraction of aspects covered by their browsed documents (AspectCoverage). We calculated the number of clicks after the last relevant document both at query and session levels (#ClicksLastRel-Q and -S) for analyzing search stopping. While Precision, Recall, and AspectCoverage are interaction metrics related to search outcomes, the other metrics are related to search behavior.

Table 6.6 shows the results of our analysis based on these measures. On average, more than ten queries were issued through both interfaces for all topics. While the difference in #Queries between the interfaces was not significant according to the Mann-Whitney U test, ScentBar users issued significantly more suggestion queries in the HC topics ($p = 0.017$). We also calculated some query complexity measures (*e.g.*, query length), but did not find any clear trends. As for the document-related measures, ScentBar users, on average, browsed more documents covering more aspects, although the difference was not statistically significant. Remarkably, ScentBar users obtained significantly higher recall than baseline interface users ($p = 0.048$), which demonstrates the advantage of ScentBar because recall is one of the most desired properties in intrinsically diverse tasks. Last but not least, no significant difference was observed for the number of irrelevant document clicks just before search stopping. This result was contrary to our expectation that ScentBar could prevent searchers from continuing tasks when no additional gain would be obtained. We discuss this issue in more detail in Section 6.6.

Table 6.6: Results of search interaction analysis. The mean (with SD) values are reported. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.

		All topics		HC topics		LC topic	
		w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
search behavior	#Queries	10.417 (6.523)	11.563 (10.886)	9.444 (6.381)	8.944 (7.131)	13.333 (6.315)	19.417 (15.963)
	%Sugg	0.376	0.480	0.355	0.505	0.437	0.405
	Queries	(0.230)	(0.275)	(0.238)	(0.279)	(0.201)	(0.259)
	#Clicks	0.156	0.164	0.165	0.175	0.133	0.142
	LastRel-Q	(0.494)	(0.511)	(0.543)	(0.465)	(0.342)	(0.596)
	#Clicks	0.208	0.563	0.278	0.389	0.000	1.083
	LastRel-S	(0.617)	(1.610)	(0.701)	(0.871)	(0.000)	(2.875)
	#Docs	22.333 (14.961)	25.729 (15.284)	22.667 (13.980)	24.972 (15.242)	21.333 (18.252)	28.000 (15.857)
search outcomes	Precision	0.714 (0.126)	0.723 (0.106)	0.685 (0.114)	0.729 (0.111)	0.801 (0.124)	0.705 (0.091)
	Recall	0.078 (0.050)	0.093 (0.057)	0.078 (0.047)	0.095 (0.061)	0.076 (0.060)	0.088 (0.043)
	Aspect	0.718	0.757	0.731	0.756	0.677	0.760
	Coverage	(0.154)	(0.157)	(0.162)	(0.162)	(0.125)	(0.146)

6.5.8 Questionnaires

Finally, we report the results of the post-task and exit questionnaires, which are summarized in Table 6.7. Note that the per-interface results of the exit questionnaire for the HC and LC topics in the table were calculated from the ratings given by participants who addressed those topics with that interface.

Table 6.7: Mean rating scores (with SD) of post-task and exit questionnaires in which five-point scale was used for each question. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.

Question	All topics		HC topics		LC topic	
	w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
post-task						
Q1 Was it difficult to exhaustively collect relevant information?	3.10 (1.32)	3.13 (1.31)	2.72 (1.23)	2.69 (1.19)	4.25 (0.87)	4.42 (0.67)
Q2 Did you exhaustively collect relevant information in an efficient manner?	3.25 (1.21)	3.02 (1.34)	3.56 (1.16)	3.50 (1.18)	2.33 (0.89)	1.58 (0.51)
Q3 Were you satisfied with the quality of the search results?	2.92 (1.18)	2.88 (1.35)	3.17 (1.08)	3.31 (1.19)	2.17 (1.19)	1.58 (0.90)
Q4 Were you satisfied with the quality of the suggestion queries?	3.04 (1.15)	3.33 (1.06)	3.33 (1.04)	3.47 (1.00)	2.17 (1.03)	2.92 (1.16)
Q5 Were you satisfied with the information you collected?	3.43 (1.15)	3.29 (1.30)	3.78 (0.99)	3.75 (1.11)	2.42 (1.00)	1.92 (0.79)
Q6 How much relevant information do you think was left unexplored?	3.38 (1.38)	3.25 (1.18)	2.97 (1.28)	2.89 (1.09)	4.58 (0.90)	4.33 (0.65)
Q7 Was missed information useful for your decision on query stopping?	—	3.04 (1.35)	—	3.28 (1.32)	—	2.33 (1.23)
Q8 Was missed information useful for your decision on query selection?	—	3.33 (1.24)	—	3.39 (1.27)	—	3.17 (1.19)
Q9 Was missed information useful for your decision on session stopping?	—	3.15 (1.32)	—	3.44 (1.32)	—	2.25 (0.87)
exit						
Q10 Was the amount of missed information convincing?	—	2.96 (0.91)	—	3.08 (1.00)	—	2.83 (0.83)
Q11 Did missed information change as you expected?	—	2.50 (0.83)	—	2.58 (0.90)	—	2.42 (0.79)
Q12 Was this interface useful for exhaustively collecting relevant information?	2.96 (0.86)	3.79 (1.06)	2.75 (0.87)	3.92 (0.90)	3.17 (0.83)	3.67 (1.23)
Q13 Was this interface easy to use?	2.96 (0.95)	3.88 (0.95)	3.00 (0.85)	3.92 (0.90)	2.92 (1.08)	3.83 (1.03)
Q14 Do you want to use this interface again?	2.50 (1.14)	3.50 (1.06)	2.67 (1.30)	3.42 (1.24)	2.33 (0.98)	3.58 (1.44)

The results for the questions Q1 and Q2 indicate that most participants faced difficulty exhaustively collecting relevant information in the LC topic. The question Q3 suggests that the low quality of search results presented by the search system would be a possible cause of this difficulty. As unreliable missed information was displayed in this topic (Q10 and Q11), ScentBar did not guide participants in effective decision-making on search stopping (Q7 and Q9). In fact, The Mann-Whitney U test revealed that the baseline interface was rated significantly better than ScentBar for the question Q2 ($p = 0.039$), which indicates that the inaccurate estimation had harmful effects on the subjective efficiency for ScentBar, as well as the objective one (Section 6.5.6). The results for the questions Q7–Q11 indicate that the ratings for missed information improved in the HC topics. As for usability, ScentBar was much more preferred by participants than the baseline interface (Q12–Q14). The differences between the two interfaces for these questions were shown to be significant by the Wilcoxon signed-rank test ($p < 0.001$ for Q12, $p < 0.001$ for Q13, and $p = 0.004$ for Q14).

Participants were also asked to answer (optional) open questions about the two interfaces in the exit-questionnaire. Some of them made favorable response to ScentBar, including “the bar visualization provided a good indication of the task progress”, “I gained satisfaction when missed information changed as expected”, and “this interface was useful for finding potential queries that I could not come up with”. On the other hand, several participants remarked that they was surprised or disappointed at the unexpected change of missed information, which suggests that the effectiveness of ScentBar depended crucially on whether the visualization behaved in a manner consistent with their expectations.

6.6 DISCUSSION

Our user study uncovered how ScentBar affected searchers’ strategies and outcomes. This section discusses the implications of our experimental results and the limitations of this work.

6.6.1 Implications

Our results demonstrated that ScentBar had the following advantages when the gain estimation algorithm worked reasonably: (1) ScentBar users stopped individual searches after collecting much relevant information from the SERPs (Section 6.5.2); (2) they issued search queries whose search results contained much missed information (Section 6.5.3); (3) they obtained high gain, particularly at the late stage of their sessions (Section 6.5.4); and (4) they obtained high gain per unit time (Section 6.5.6). Our post-hoc analysis of search interaction data revealed that query suggestion was more frequently used by ScentBar users (Section 6.5.7), which implies that suggesting queries with missed information affected searchers’ query formulation strategies and thus

enabled them to conduct intrinsically diverse tasks more efficiently.

The search interaction analysis also revealed that displaying missed information did not affect the number of irrelevant document clicks just before search stopping. A possible explanation for this phenomenon is that seeing visualized missed information made participants more cautious about making decisions on search stopping. In fact, ScentBar users spent, on average, longer completing the tasks (13.6 minutes) than the baseline interface users (11.8 minutes), although the difference was not statistically significant. These results suggest that additional features would be needed for supporting searchers' decision-making on search stopping. While ScentBar suggests which queries have much missed information, it does not provide any clues about which search results users should assess to obtain additional gain. Possible solutions to this issue include visualizing the per-aspect relevance for each search result, like *AspecTiles* [46], and/or re-ranking search results containing much missed information at high positions, both of which can be incorporated with ScentBar. Another approach could be notifying searchers of the effort they have expended so far as well as their missed information. This could help searchers more rationally decide whether they should conduct additional searches to obtain more gain or stop searching to avoid wasting time. Techniques for quantifying search cost (*e.g.*, [6]) would be useful for implementing this idea.

Another finding from our user study is that ScentBar worsened rather than improved search performance when missed information was estimated inaccurately. More specifically, the gain per unit time obtained by ScentBar users was lower than that by the baseline interface users. This is probably because unconvincing visualization of missed information and its unexpected change prevented searchers from making rational decisions on query selection and search stopping. Our results suggest that the accuracy of gain estimation is critical for ScentBar to produce intended effects.

6.6.2 Limitations

This work has several limitations that we should acknowledge. We discuss each of them in turn below.

First, while we formalized gain as an intent-aware metric, as with other evaluation metrics like *ERR-IA* [16], this approach has a drawback in that the content overlap among documents is not taken into account when the gain is evaluated. Thus, this metric has the potential to give a high score to documents with high overlap. While we tried to prevent participants from collecting overlapped documents by presenting diversified search results to them, this is an ad-hoc solution, not an optimal one. Ideally speaking, gain should be modeled on the basis of more fine-grained units, such as nuggets [21]. Thus, for more accurate evaluation, our gain formalization and

estimation algorithm would require further improvement by, for example, modeling and mining aspects hierarchically [43].

Second, we used only four search topics to evaluate the effectiveness of ScentBar in intrinsically diverse tasks. Limiting the number of topics was necessary given that the interfaces were designed to be the within-subject factor in our user study (*i.e.*, every participant used both interfaces) and that we could not use a predefined length of task execution time as our research questions included the effect of ScentBar on session stopping. To validate the generality of our claim about the effectiveness of ScentBar, we would need to conduct additional experiments with more search topics and more diverse participants in the future.

Third, we selected a simple interface as the baseline in our user study while various interfaces showing information scent have been proposed. As outlined in Section 2.3.1, the query preview interface proposed by Qvarfordt *et al.* [68] provides a similar functionality as ours: their interface visualizes the number of (un)clicked search results for the current query. To estimate its effectiveness in intrinsically diverse tasks, we measured correlations between the number of unclicked search results and the oracle missed information scent in a manner similar to the gain accuracy estimation (Section 6.5.1). The correlation coefficient was 0.334 even in the best topic. Given the fact that ScentBar worsened the search performance for the LC topic, this suggests that such simple visualization could be less effective for intrinsically diverse tasks, while direct comparison in a user study setting would be needed to gain more insights.

6.7 SUMMARY

In this chapter, we proposed ScentBar, a query suggestion interface visualizing the amount of missed information for individual queries. We defined the amount of missed information for a query as the additional gain that can be obtained from unclicked search results of the query. Results of our user study involving 24 participants showed the following advantages of our interface for search topics where gain was estimated reasonably accurately: (1) ScentBar users stopped SERP examination after collecting more relevant information; (2) they issued queries whose search results contained more missed information; (3) they obtained higher gain, particularly at the late stage of their sessions; and (4) they obtained higher gain per unit time. These results suggest that unless inaccurate estimates of missed information were visualized, ScentBar enabled users to utilize effective query formulation strategies, while it had little noticeable effect on their search stopping behavior.

Future directions of this work include exploring ways for utilizing missed information. While ScentBar visualizes what is left for browsing (*i.e.*, the visualization decreases as searchers obtain gain), showing how much searchers have browsed might have a different effect on their search

behavior. Providing the visualization in different places and/or at different timing would also be worth exploring. Missed information could also be incorporated with query suggestion algorithms to rank suggestion queries from which searchers would be able to obtain much additional gain at high positions.

CONCLUSIONS

7.1 SUMMARY

This thesis discussed techniques and methodologies for estimating search intents and for evaluating search outcomes in the context of complex search tasks. Our key idea for search intent estimation and for search outcome evaluation is to analyze fine-grained search behavior of users who conduct complex tasks. Four research topics that we addressed in this thesis are summarized as follows:

- **Estimating Attractive Terms from Gazing Behavior**

We tackled the problem of estimating a set of terms that clarify a searcher's ambiguous topical search intent from their observable search behavior. We focused on searchers' gazing behavior on their browsed Web pages and assumed that terms drawing a high degree of a searcher's attention are likely to be relevant to his/her search intent. On the basis of the assumption, we compared four methods, each of which leveraged gazing data differently from each other. We conducted a laboratory experiment with eight participants. Experimental results showed that the nMGT model, which assumes positive correlation between term frequency and gaze frequency, outperformed other models and baseline methods in terms of estimation accuracy. In addition, our qualitative analysis revealed that relevant terms estimated by the nMGT model differed from participant to participant, which suggests that eye-gazing data can capture the individual difference in user interests even when they conduct a search on the same topic.

- **Predicting Query Reformulation from Current Search Behavior**

We proposed a method to detect the change of search scope in advance of the following search. Our addressed task was a classification problem with five query reformulation types (*i.e.*, Generalization, Specialization, Parallel Move, Mission Change, and Error Correction). The proposed method constructs a SVM classifier using various kinds of user interactions in the current search and predicts the query reformulation type of the following search. As a result of evaluation with search log data, our SVM-based method was shown to be able to predict the query reformulation types with about 41% accuracy, despite the small size of training data. We also found that the following behavioral features were most effective for the subsequent query reformulation type prediction: query term length, attention area in Web pages, and clicks on search results.

- **Analyzing How Inconsistent Information Affects Search Behavior**

We investigated how users handled inconsistent information by focusing on the effect of search expertise on their search behavior and satisfaction criteria. We defined search expertise as specialized knowledge on the internal workings of search engines. The research questions we addressed were: (1) how does the inconsistency of answers that users find relate to the degree of task satisfaction that they gain? (2) when do users complete search tasks where inconsistent answers are found? and (3) how do answers targeted by users change as time passes? The main findings from our analysis using publicly available search log data were summarized as follows: (1) finding inconsistent answers caused dissatisfaction for users with search expertise (search experts), while effort in searching for answers was the dominant factor in task satisfaction for those without search expertise (search non-experts); (2) search experts tended to spend longer completing tasks than search non-experts even after finding possible answers; and (3) search experts narrowed down the targeted information to promising answers as time passed as opposed to search non-experts, whose found answers were still diverse even in the closing stage of their task sessions. Our findings suggested that search non-experts tended to be less concerned about the consistency in their found answers and raised important challenges to be tackled to help search non-experts to properly handle inconsistent information (*e.g.*, search engines would need to notify them of the existence of multiple answer instances and to establish techniques for retrieving evidence that supports a given answer).

- **Visualizing the Amount of Missed Information in Search Results**

We proposed ScentBar, a query suggestion interface visualizing the amount of missed information for individual queries. We defined the amount of missed information for a query as the additional gain that could be obtained from unclicked search results of the query.

Results of our user study involving 24 participants showed the following advantages of our interface for search topics where gain was estimated reasonably accurately: (1) ScentBar users stopped SERP examination after collecting more relevant information; (2) they issued queries whose search results contained more missed information; (3) they obtained higher gain, particularly at the late stage of their sessions; and (4) they obtained higher gain per unit time. These results suggest that unless inaccurate estimates of missed information were visualized, ScentBar enabled users to utilize effective query formulation strategies, while it had little noticeable effect on their search stopping behavior.

The main contributions of this work are as follows.

- Technical contributions
 - We proposed analytical methods for better understanding the ambiguity of, and the transitivity of, search intents in complex tasks.
 - We proposed analytical methods for more accurately evaluating search outcomes, such as search satisfaction and search stopping, for complex tasks.
- Social contributions
 - We uncovered the characteristics of search behavior observed in complex tasks and those individual differences.
 - We enabled searchers conducting intrinsically diverse tasks to make more rational judgments, particularly about query selection, and to obtain relevant information more efficiently.

7.2 FUTURE DIRECTIONS

There still remains several research topics that need to be explored in future work. First, on the basis of the results of our analyses on search intent estimation and on search outcome evaluation, we would like to establish methodologies for supporting users' searches in complex tasks. Possible user-support methodologies include: re-ranking search results containing more attractive terms to higher positions; proactively suggesting related queries on the basis of the prediction of the change in search scope; and assisting searchers who will be likely to feel dissatisfaction. Second, it is worth investigating how such methodologies affect user behavior to gain more insights into complex tasks. Third, while we focused mainly on topical search intents in Chapters 3 and 4, how to estimate other types of search intents (*e.g.*, cognitive search intents [51]) from user behavior would be a challenging research topic. Finally, we are also interested in uncovering how user

satisfaction (Chapter 5) and search stopping (Chapter 6) relate to other search outcomes, such as user frustration [27] and search success [34]. We believe these research directions will lead to the development of intent- and outcome-aware Web search engines, which can make users' search process in complex tasks more efficient and effective.

BIBLIOGRAPHY

- [1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 345–354, New York, NY, USA, 2011. ACM.
- [2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Jeong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.
- [3] A. Aula, R. M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 35–44. ACM, 2010.
- [4] L. 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.
- [5] L. Azzopardi, D. Kelly, and K. Brennan. How query cost affects search behavior. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 23–32, New York, NY, USA, 2013. ACM.
- [6] L. Azzopardi and G. Zuccon. An analysis of theories of search and search behavior. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 81–90, New York, NY, USA, 2015. ACM.
- [7] F. Baskaya, H. Keskustalo, and K. Järvelin. Modeling behavioral factors in interactive information retrieval. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 2297–2302, New York, NY, USA, 2013. ACM.

- [8] P. Boldi, F. Bonchi, C. Castillo, and S. Vigna. From "dango" to "japanese cakes": Query reformulation models and patterns. In *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '09, pages 183–190. IEEE Computer Society, 2009.
- [9] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, Sept. 2002.
- [10] G. Buscher, A. Dengel, and L. van Elst. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 387–394. ACM, 2008.
- [11] G. Buscher, L. van Elst, and A. Dengel. Segment-level display time as implicit feedback: a comparison to eye tracking. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 67–74. ACM, 2009.
- [12] S. Büttcher, C. L. A. Clarke, and G. V. Cormack. *Information retrieval: implementing and evaluating search engines*. MIT Press, Cambridge, Mass., 2010.
- [13] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM.
- [14] M.-A. Cartright, R. W. White, and E. Horvitz. Intentions and attention in exploratory health search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 65–74, New York, NY, USA, 2011. ACM.
- [15] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011.
- [16] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.
- [17] Z. Cheng, B. Gao, and T.-Y. Liu. Actively predicting diverse search intent from user browsing behaviors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 221–230. ACM, 2010.

- [18] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 490–497, New York, NY, USA, 2001. ACM.
- [19] F. Chierichetti, R. Kumar, P. Raghavan, and T. Sarlos. Are web users really markovian? In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 609–618. ACM, 2012.
- [20] L. B. Chilton and J. Teevan. Addressing people's information needs directly in a web search result page. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 27–36, New York, NY, USA, 2011. ACM.
- [21] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM.
- [22] C. L. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Proceedings of the 2Nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, ICTIR '09, pages 188–199, Berlin, Heidelberg, 2009. Springer-Verlag.
- [23] H. T. Dang, D. Kelly, and J. Lin. Overview of the TREC 2007 question answering track. In *Proceedings of the 16th Text REtrieval Conference*, TREC '07, 2007.
- [24] M. Dostert and D. Kelly. Users' stopping behaviors and estimates of recall. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 820–821, New York, NY, USA, 2009. ACM.
- [25] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 475–484, New York, NY, USA, 2011. ACM.
- [26] K.-B. Duan and S. S. Keerthi. Which is the best multiclass svm method? an empirical study. In *Multiple Classifier Systems*, volume 3541 of *Lecture Notes in Computer Science*, pages 732–760. Springer Berlin / Heidelberg, 2005.
- [27] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 34–41, New York, NY, USA, 2010. ACM.

- [28] B. J. Fogg and H. Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 80–87, New York, NY, USA, 1999. ACM.
- [29] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, Apr. 2005.
- [30] Q. Guo and E. Agichtein. Ready to buy or just browsing?: detecting web searcher goals from interaction data. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 130–137. ACM, 2010.
- [31] Q. Guo and E. 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*, WWW '12, pages 569–578. ACM, 2012.
- [32] Q. Guo, D. Lagun, and E. Agichtein. Predicting web search success with fine-grained interaction data. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2050–2054, New York, NY, USA, 2012. ACM.
- [33] Q. Guo, R. W. White, Y. Zhang, B. Anderson, and S. T. Dumais. Why searchers switch: Understanding and predicting engine switching rationales. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 335–344, New York, NY, USA, 2011. ACM.
- [34] A. Hassan, R. Jones, and K. L. 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*, WSDM '10, pages 221–230, New York, NY, USA, 2010. ACM.
- [35] A. Hassan and R. W. White. Personalized models of search satisfaction. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 2009–2018, New York, NY, USA, 2013. ACM.
- [36] A. Hassan, R. W. White, S. T. Dumais, and Y.-M. Wang. Struggling or exploring?: Disambiguating long search sessions. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 53–62, New York, NY, USA, 2014. ACM.
- [37] J. He, V. Hollink, and A. de Vries. Combining implicit and explicit topic representations for result diversification. In *Proceedings of the 35th International ACM SIGIR Conference*

- on *Research and Development in Information Retrieval*, SIGIR '12, pages 851–860, New York, NY, USA, 2012. ACM.
- [38] M. A. Hearst. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 59–66, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.
- [39] M. A. Hearst. *Search User Interfaces*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.
- [40] O. Hoeber and X. D. Yang. A comparative user study of web search interfaces: Hotmap, concept highlighter, and google. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 866–874. IEEE Computer Society, 2006.
- [41] C. Hölscher and G. Strube. Web search behavior of internet experts and newbies. *Computer Networks*, 33(1-6):337–346, 2000.
- [42] B. Hu, Y. Zhang, W. Chen, G. Wang, and Q. Yang. Characterizing search intent diversity into click models. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 17–26. ACM, 2011.
- [43] S. Hu, Z. Dou, X. Wang, T. Sakai, and J.-R. Wen. Search result diversification based on hierarchical intents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 63–72, New York, NY, USA, 2015. ACM.
- [44] J. Huang and E. N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 77–86. ACM, 2009.
- [45] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 1225–1234. ACM, 2011.
- [46] M. Iwata, T. Sakai, T. Yamamoto, Y. Chen, Y. Liu, J.-R. Wen, and S. Nishio. Aspectiles: Tile-based visualization of diversified web search results. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 85–94, New York, NY, USA, 2012. ACM.
- [47] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

- [48] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142. ACM, 2002.
- [49] R. Jones and K. L. Klinkner. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 699–708, New York, NY, USA, 2008. ACM.
- [50] M. P. Kato, T. Sakai, and K. Tanaka. Structured query suggestion for specialization and parallel movement: Effect on search behaviors. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 389–398, New York, NY, USA, 2012. ACM.
- [51] M. P. Kato, T. Yamamoto, H. Ohshima, and K. Tanaka. Cognitive search intents hidden behind queries: A user study on query formulations. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 313–314, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.
- [52] M. P. Kato, T. Yamamoto, H. Ohshima, and K. Tanaka. Investigating users' query formulations for cognitive search intents. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 577–586, New York, NY, USA, 2014. ACM.
- [53] D. Kelly, A. Cushing, M. Dostert, X. Niu, and K. Gyllstrom. Effects of popularity and quality on the usage of query suggestions during information search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 45–54, New York, NY, USA, 2010. ACM.
- [54] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 193–202, New York, NY, USA, 2014. ACM.
- [55] D. Kraft and T. Lee. Stopping rules and their effect on expected search length. *Information Processing & Management*, 15(1):47–58, 1979.
- [56] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

- [57] J. Li, S. Huffman, and A. 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. ACM, 2009.
- [58] Q. Liu, E. Agichtein, G. Dror, Y. Maarek, and I. Szpektor. When web search fails, searchers become askers: Understanding the transition. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 801–810, New York, NY, USA, 2012. ACM.
- [59] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. Kato, H. Ohshima, and K. Zhou. Overview of the ntcir-11 imine task. In *Proceedings of the 11th NTCIR Conference*, NTCIR-11, 2014.
- [60] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
- [61] Y. Mansourian and N. Ford. Search persistence and failure on the web: a “bounded rationality” and “satisficing” analysis. *Journal of Documentation*, 63(5):680–701, 2007.
- [62] G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, New York, NY, USA, 1995.
- [63] D. Maxwell, L. Azzopardi, K. Järvelin, and H. Keskustalo. Searching and stopping: An analysis of stopping rules and strategies. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 313–322, New York, NY, USA, 2015. ACM.
- [64] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: A survey study of status message q&a behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1739–1748, New York, NY, USA, 2010. ACM.
- [65] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama, and K. Tanaka. Trustworthiness analysis of web search results. In *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'07, pages 38–49, Berlin, Heidelberg, 2007. Springer.
- [66] P. Pirolli. *Information Foraging Theory: Adaptive Interaction with Information*. Human Technology Interaction Series. Oxford University Press, USA, 2007.
- [67] C. Prabha, L. S. Connaway, L. Olszewski, and L. R. Jenkins. What is enough? satisficing information needs. *Journal of Documentation*, 63(1):74–89, 2007.

- [68] P. Qvarfordt, G. Golovchinsky, T. Dunnigan, and E. Agapie. Looking ahead: Query preview in exploratory search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 243–252, New York, NY, USA, 2013. ACM.
- [69] K. Raman, P. N. Bennett, and K. Collins-Thompson. Toward whole-session relevance: Exploring intrinsic diversity in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 463–472, New York, NY, USA, 2013. ACM.
- [70] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the ntcir-10 intent-2 task. In *Proceedings of the 10th NTCIR Conference*, NTCIR-10, 2013.
- [71] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 881–890, New York, NY, USA, 2010. ACM.
- [72] T. Saracevic. Relevance reconsidered. In *Proceedings of the 2nd Conference on Conceptions of Library and Information Science*, CoLIS 2, pages 201–218, 1996.
- [73] G. Shani and N. Tractinsky. Displaying relevance scores for search results. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 901–904, New York, NY, USA, 2013. ACM.
- [74] B. Shneiderman, D. Byrd, and W. B. Croft. Clarifying search: A user-interface framework for text searches. *D-Lib Magazine*, 3(1):18–20, 1997.
- [75] S. Shogen, T. Yamamoto, and K. Tanaka. A query suggestion interface with features of queries and search results. In *Proceedings of the 15th International Conference on Web-Age Information Management*, WAIM '14, pages 615–619, 2014.
- [76] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the ntcir-9 intent task. In *Proceedings of the 9th NTCIR Workshop Meeting*, NTCIR-9, 2011.
- [77] P. Thomas, A. Moffat, P. Bailey, and F. Scholer. Modeling decision points in user search behavior. In *Proceedings of the 5th Information Interaction in Context Symposium*, IiX '14, pages 239–242, New York, NY, USA, 2014. ACM.
- [78] E. G. Toms and L. Freund. Predicting stopping behaviour: A preliminary analysis. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 750–751, New York, NY, USA, 2009. ACM.

- [79] K. Tsukuda, T. Sakai, Z. Dou, and K. Tanaka. Estimating intent types for search result diversification. In *Proceedings of the 9th Asia Information Retrieval Societies Conference, AIRS '13*, pages 25–37, Berlin, Heidelberg, 2013. Springer.
- [80] A. H. Turpin and W. 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.
- [81] H. Wang, Y. Song, M.-W. Chang, X. He, A. Hassan, and R. W. White. Modeling action-level satisfaction for search task satisfaction prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14*, pages 123–132, New York, NY, USA, 2014. ACM.
- [82] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.
- [83] R. White. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13*, pages 3–12, New York, NY, USA, 2013. ACM.
- [84] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09*, pages 87–96, New York, NY, USA, 2009. ACM.
- [85] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09*, pages 132–141, New York, NY, USA, 2009. ACM.
- [86] R. W. White and E. Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Trans. Inf. Syst.*, 27(4):23:1–23:37, Nov. 2009.
- [87] R. W. White and R. A. Roth. Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.
- [88] B. M. Wildemuth and L. Freund. Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, HCIR '12*, pages 4:1–4:10, New York, NY, USA, 2012. ACM.

- [89] W.-C. Wu and D. Kelly. Online search stopping behaviors: An investigation of query abandonment and task stopping. *Proceedings of the American Society for Information Science and Technology*, 51(1):1–10, 2014.
- [90] W.-C. Wu, D. Kelly, and A. Sud. Using information scent and need for cognition to understand online search behavior. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 557–566, New York, NY, USA, 2014. ACM.
- [91] S. Xu, H. Jiang, and F. C. Lau. Personalized online document, image and video recommendation via commodity eye-tracking. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, pages 83–90. ACM, 2008.
- [92] S. Xu, H. Jiang, and F. C. Lau. User-oriented document summarization through vision-based eye-tracking. In *Proceedings of the 14th international conference on Intelligent user interfaces*, IUI '09, pages 7–16. ACM, 2009.
- [93] Y. Xu and D. Mease. Evaluating web search using task completion time. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 676–677, New York, NY, USA, 2009. ACM.
- [94] Y. Yamamoto and K. Tanaka. Enhancing credibility judgment of web search results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1235–1244, New York, NY, USA, 2011. ACM.
- [95] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1561–1564, New York, NY, USA, 2010. ACM.
- [96] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. 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, pages 91–100, New York, NY, USA, 2014. ACM.
- [97] L. Zach. When is “enough” enough? modeling the information-seeking and stopping behavior of senior arts administrators: Research articles. *Journal of the American Society for Information Science and Technology*, 56(1):23–35, 2005.
- [98] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and*

- Development in Information Retrieval*, SIGIR '04, pages 210–217, New York, NY, USA, 2004. ACM.
- [99] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 15–24, New York, NY, USA, 2009. ACM.

PUBLICATIONS

JOURNAL PAPERS

- [1] 梅本和俊, 山本岳洋, 田中克己. 検索専門性と事前知識に着目した検索行動とタスク満足度の関係性分析. 情報処理学会論文誌データベース, Vol. 7, No. 4, pp. 13–28, 2014.
- [2] 梅本和俊, 中村聡史, 山本岳洋, 田中克己. Web 検索時の行動情報を用いたクエリ修正タイプの予測. 情報処理学会論文誌データベース, Vol. 6, No. 3, pp. 132–147, 2013.
- [3] 梅本和俊, 山本岳洋, 中村聡史, 田中克己. 視線情報からの注目語抽出に基づく検索意図のリアルタイム推定. 情報処理学会論文誌データベース, Vol. 6, No. 3, pp. 120–131, 2013.
- [4] 梅本和俊, 山本岳洋, 中村聡史, 田中克己. ユーザの視線を利用した検索意図推定とそれに基づく情報探索支援. 日本データベース学会論文誌, Vol. 10, No. 1, pp. 61–66, 2011.

INTERNATIONAL CONFERENCE PAPERS

- [1] Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. How do Users Handle Inconsistent Information? The Effect of Search Expertise. In Proceedings of the 31st ACM Symposium on Applied Computing (SAC 2016), to appear, 2016.
- [2] Kazutoshi Umemoto, Satoshi Nakamura, Takehiro Yamamoto, and Katsumi Tanaka. Predicting Query Reformulation Type from User Behavior. In Proceedings of the 28th ACM Symposium on Applied Computing (SAC 2013), pp. 894–901, 2013.
- [3] Kazutoshi Umemoto, Takehiro Yamamoto, Satoshi Nakamura, and Katsumi Tanaka. Search Intent Estimation from User’s Eye Movements for Supporting Information Seeking. In Proceedings of the 11th International Working Conference on Advanced Visual Interfaces (AVI 2012), pp. 349–356, 2012.

INTERNATIONAL WORKSHOP PAPERS

- [1] Tomohiro Manabe, Kosetsu Tsukuda, Kazutoshi Umemoto, Yoshiyuki Shoji, Makoto P. Kato, Takehiro Yamamoto, Meng Zhao, Soungwoong Yoon, Hiroaki Ohshima, and Katsumi Tanaka. Information Extraction based Approach for the NTCIR-10 1CLICK-2 Task. The 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-10), 2013.
- [2] Kazutoshi Umemoto, Satoshi Nakamura, Takehiro Yamamoto, and Katsumi Tanaka. Estimating Search Intent Transition from User's Behavior Log Data. The 4th International Workshop with Mentors on Databases, Web and Information Management for Young Researchers (iDB 2012), 2012.

DOMESTIC SYMPOSIUM AND WORKSHOP PAPERS

- [1] 梅本和俊, 山本岳洋, 田中克己. 網羅的情報収集タスクにおける検索結果の未閲覧情報量の提示. 第8回データ工学と情報マネジメントに関するフォーラム (DEIM 2016), 2016 (発表予定).
- [2] 梅本和俊, 中村聡史, 山本岳洋, 田中克己. 検索意図の遷移検出に基づく動的なクエリ推薦に向けた行動ログデータの分析. 第148回ヒューマンコンピュータインタラクション研究会, 2012.
- [3] 梅本和俊, 山本岳洋, 中村聡史, 田中克己. 視線情報を用いたユーザの検索意図推定手法の検討. 第20回Webインテリジェンスとインタラクション研究会, 2011.
- [4] 梅本和俊, 山本岳洋, 中村聡史, 田中克己. 視線情報を利用したユーザの意図推定とそれに基づく情報探索支援. 第3回データ工学と情報マネジメントに関するフォーラム (DEIM 2011), 2011.