

Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Qi Guo

Date

Modeling Rich Interactions for Web Search Intent Inference, Ranking and Evaluation

By

Qi Guo

Doctor of Philosophy

Computer Science and Informatics

Eugene Agichtein, Ph.D.

Advisor

James J. Lu, Ph.D.

Committee Member

Ryen W. White, Ph.D.

Committee Member

Li Xiong, Ph.D.

Committee Member

Hongyuan Zha, Ph.D.

Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.

Dean of the James T. Laney School of Graduate Studies

Date

Modeling Rich Interactions for Web Search Intent Inference, Ranking and Evaluation

By

Qi Guo

M.S., Emory University, 2010

Advisor: Eugene Agichtein, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Computer Science and Informatics

2012

Abstract

Modeling Rich Interactions for Web Search Intent Inference, Ranking and Evaluation

By Qi Guo

Billions of people interact with Web search engines daily and their interactions provide valuable clues about their interests and preferences. While modeling search behavior, such as queries and clicks on results, has been found to be effective for various Web search applications, the effectiveness of the existing approaches are limited by ignoring what the searcher sees (*examination*) and does (*context*) before clicking a result. This thesis aims to address these limitations by modeling and interpreting a wider range of searcher interactions, including mouse cursor movement and scrolling behavior (or pinching, zooming and sliding with a touch screen), that could be served as a proxy of searcher examination, contextualized in a search session. The thesis focuses on improving three fundamental and interrelated areas of Web search, namely, *intent inference*, *ranking* and *evaluation*. To improve the first area, the thesis developed techniques to infer the immediate search goals in a search session, along multiple dimensions, including top-level general intent (e.g., navigational vs. informational), commercial intent (e.g., research vs. purchase) and advertising receptiveness (i.e., interest in search ads). To improve the second area, the thesis developed the *Post-Click Behavior (PCB)* relevance prediction model for estimating the “intrinsic” document relevance from the examination and interaction patterns on the viewed result documents. To improve the third area, the thesis developed techniques for predicting search success, which include a principled framework to study Web search success, and fine-grained interaction models that improve prediction accuracy for both desktop and mobile settings. As demonstrated with extensive empirical evaluation, the developed techniques outperform the state-of-the-art methods that only use query, click and time signals, enabling more intelligent Web search systems.

Modeling Rich Interactions for Web Search Intent Inference, Ranking and Evaluation

By

Qi Guo

M.S., Emory University, 2010

Advisor: Eugene Agichtein, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2012

Acknowledgments

First and foremost, I would like to thank my advisor, Professor Eugene Agichtein, for helping me grow as a researcher in so many different ways. I started my graduate study with the interest in building intelligent information systems that can automatically adapt to user interests and preferences and was intrigued by Eugene's earlier work in this area. Over the years, my ideas were further developed and realized with Eugene's guidance and encouragements, turning into this thesis. Eugene spared no effort in making insightful suggestions on my research no matter how busy he was and taught me many different things, ranging from seeing the big picture, to finding the right problems to work on, to improving the writing and presentation skills. Eugene also encouraged me to explore opportunities outside Emory, including collaborating with other universities, interning with different companies and attending conferences, which helps me broaden my horizon, and connect and learn from other researchers. Eugene is an amazing mentor and I am extremely fortunate to have had him as my PhD advisor.

I would also like to thank my thesis committee members: Professor James J. Lu, Dr. Ryen W. White, Professor Li Xiong, and Professor Hongyuan Zha. Their insightful comments and suggestions help me better formalize the research problem (e.g., Chapter 4), strengthen the technical content, and connect the ideas in different chapters and turn the thesis into a more coherent piece. In addition to being a member in my thesis committee, Ryen was my mentor during my two summer internships with Microsoft. During the time, I was able to learn greatly from Ryen to further develop myself as a researcher from a different perspective, and get a broader view of my research and a deeper understanding of the real-world challenges.

A large fraction in the thesis benefits from a variety of collaborations and academia activities. Some of the early work of Chapter 3 was resulted from the collaboration with Charles L.A. Clarke and Azin Ashkan from the University of Waterloo. The work in Chapter 3 was also recognized and supported by a Yahoo! Key Scientific Challenge (KSC) award, through which I was able to interact with outstanding senior researchers and fellow graduate students and had valuable discussion and feedback. Some of the initial ideas of

post-click behavior modeling in Chapter 4 were formed through the discussion during the Yahoo! KSC summit. Special thanks go to Henry Feild from University of Massachusetts Amherst, who I also met during the Yahoo! KSC summit – two studies in Chapter 4 and Chapter 5 used the dataset generously made publicly available by him. The work in Chapter 5 was resulted from collaborations with my colleagues Mikhail Ageev, Dmitry Lagun, and Shuai Yuan at Emory. In particular, Dmitry and I, being office mates, had numerous stimulating and enjoyable discussions from time to time, which helped me greatly in shaping and improving some of the ideas in the thesis.

I would like to thank all my collaborators and co-authors in various projects: Mikhail Ageev, Eugene Agichtein, Blake Anderson, Azin Ashkan, Daniel Billsus, Wei Chai, Charles L.A. Clarke, Selden Deemer, Fernando Diaz, Susan T. Dumais, Haojian Jin, Ryan P. Kelly, Julia Kiseleva, Dmitry Lagun, Qiaoling Liu, Arthur Murphy, Denis Savenkov, Jue Wang, Ryen W. White, Elad Yom-tov, Shuai Yuan, Yunqiao Zhang. While some of the collaborations ended up not being part of the thesis, they helped me develop as a researcher and/or shape some of the ideas – this thesis would not have been possible without you.

I would also like to thank all my friends at Emory University, in Atlanta, and those who I made during my internships with Microsoft, Yahoo! and Shopping.com, for making my PhD journey full of joy. Special thanks go to the former and current members of Emory Information Intelligent Lab (IRLab): Mikhail Ageev, Ablimit Aji, Alvin Grissom II, Tianyong Hao, Haojian Jin, Ryan P. Kelly, Julia Kiseleva, Alexander Kotov, Dmitry Lagun, Baoli Li, Qiaoling Liu, Yandong Liu, Akshatha K. Pai, Denis Savenkov, JongHo Shin, Yu Wang, and Nikita Zhiltsov.

My family gives me motivations and tremendous support I cannot describe in words. The thesis would not be possible without the unconditional love from my mom, my dad, my twin brother, and my grandma and grandpa. I thank my brother Lin for taking care of my parents and grandparents when I am away. I am sorry that I could not be with you when grandpa passed away during my preparation of this thesis. Last but not the least, I would like to thank my dearest wife Jingying for her love, understanding and support, especially when we were separated by the Pacific Ocean in my first couple of years of graduate school.

This research was funded by the National Science Foundation under Grant IIS-1018321, by the Microsoft Research “Beyond Search” Award, a Yahoo! Key Scientific Challenges

Award, the Yahoo! Labs Faculty Research Engagement Program, and by travel support from the Association for Computing Machinery Special Interest Group on Information Retrieval.

*To my family,
and in the loving memory of my grandpa,
who passed away a few weeks before my thesis defense.*

Contents

1	Introduction	1
1.1	Contributions	8
1.2	Organization	9
2	Background and Related Work	10
2.1	Inferring Search Intent	10
2.1.1	General Search Intent	10
2.1.2	Topical Search Intent	12
2.1.3	Commercial Search Intent	12
2.2	Estimating Document Relevance	14
2.2.1	Click-through	15
2.2.2	Dwell Time	17
2.2.3	Examination	18
2.2.4	Personalization	20
2.2.5	Search Context	21
2.3	Evaluating Search Experience	22
2.3.1	Query-level	23
2.3.2	Session-level	23
2.3.3	Multi-engine level	24
3	Inferring Search Intent	27
3.1	Motivation	29

3.1.1	Inferring General Search Intent	29
3.1.2	Inferring Commercial Search Intent	30
3.1.3	Application: Search Advertising	31
3.2	Search and User Model	33
3.2.1	Search Model: Tasks and Goals	33
3.2.2	User Model: Goal-driven Search	34
3.3	Infrastructure, Features and Algorithms	36
3.3.1	Infrastructure	36
3.3.2	Features	37
3.3.3	Classifier Implementation	40
3.4	Inferring General Search Intent	42
3.4.1	Data Collection	42
3.4.2	Metrics	43
3.4.3	Methods Compared	44
3.4.4	Results and Discussion	45
3.5	Inferring Commercial Search Intent	51
3.5.1	Data Collection	51
3.5.2	Methods Compared	52
3.5.3	Results and Discussion	52
3.5.4	Ad Click-through on Real Search Data	54
3.6	Inferring Advertising Receptiveness	55
3.6.1	Methods Compared	56
3.6.2	Data and Evaluation Metrics	57
3.6.3	Results and Discussion	57
3.7	Summary	60
4	Estimating Document Relevance	61
4.1	Landing Page Examination	64
4.2	Post-Click Behavior (PCB) Features	68
4.2.1	Dwell Time	68

4.2.2	Result Rank	68
4.2.3	Cursor Movements	70
4.2.4	Vertical Scrolling	70
4.2.5	Interactions in the Areas of Interest (AOI)	70
4.2.6	Task/Session-level Context	71
4.2.7	User Normalization	71
4.3	Relevance Estimation Models	72
4.3.1	Ridge Regression (RR)	72
4.3.2	Bagging with Regression Trees (BRT)	72
4.4	Experimental Setup	73
4.4.1	Data	73
4.4.2	Evaluation Metrics	74
4.4.3	Methods Compared	75
4.5	Results and Discussion	76
4.5.1	Feature Association with Relevance	77
4.5.2	Predicting Document Relevance	80
4.5.3	Re-ranking	83
4.6	Summary	86
5	Evaluating Search Experience	87
5.1	Predicting Search Success with UFindIt	90
5.1.1	Search Success Model	91
5.1.2	Acquiring Search Behavior Data	93
5.1.3	Predicting Search Success	94
5.1.4	Results and Discussion	98
5.1.5	Real World Success Prediction:A Log-based Study	100
5.2	Predicting Search Success with FSB	106
5.2.1	Fine-grained Session Behavior (FSB) Features	106
5.2.2	Data	110
5.2.3	Evaluation Metrics	111

5.2.4	Results and Discussion	111
5.3	Predicting Success in Mobile Search	118
5.3.1	Methodology	118
5.3.2	Results and Discussion	120
5.4	Summary	124
6	Conclusions and Future Work	126
6.1	Summary of Findings	126
6.2	Integrating Intent Inference, Ranking and Evaluation	128
6.2.1	Integrating Intent Inference	129
6.2.2	Integrating Relevance Estimation	129
6.2.3	Integrating Automatic Evaluation	130
6.2.4	Pre-click and Post-click Instrumentation	130
6.2.5	Infrastructures for Offline and Online Deployment	131
6.2.6	Evaluating the Deployed System	132
6.3	Limitations and Future Work	134
	Bibliography	137

List of Figures

1.1	(a) A user of an eye-tracker with camera integrated into a computer monitor; (b) eye gaze trajectories captured overlaid on a search engine result page	2
1.2	Density of distance between mouse cursor and eye gaze positions on SERPs and non-SERPs (landing pages): Euclidean distance and distances in X and Y directions.	3
1.3	Top search results from two commercial search engines for query “jaguar”	5
1.4	A refined classic Web IR model	7
3.1	Searcher mouse trajectories on the search engine result pages for query with navigational intent: “facebook” (a) and query with informational intent: “spanish wine” (b)	30
3.2	Searcher gaze position (a) and corresponding mouse trajectory (b) for query with research intent	31
3.3	Mouse trajectory on a SERP for query “green coffee maker” with an ad click on the <i>next</i> search result page	32
3.4	Relationship between a search task, immediate goals and specific searches to accomplish each goal.	33
3.5	An example user session, consisting of two consecutive disjoint search tasks.	34
3.6	Sample states and observations for a single search within a task.	35
3.7	Session aggregation of search-level predictions	40

3.8	Mouse trajectories for searches with navigational/re-finding intent: query “rpi rankings” (a) and query “emory financial aid” (b)	48
3.9	Mouse trajectories for abandoned searches: query “unpaid parking ticket” followed by a click on “did you mean” (a) and query “the things they carried” followed by a reformulation (b)	49
3.10	CRF model configuration with two hidden states, A+ (receptive), and A- (non-receptive), with labels assigned according to the observed <i>future</i> ad click-through - here on the third search result page within the session.	56
3.11	Mouse trajectory and gaze heatmap on a search engine result page when mouse is not used to mark or focus user interest	59
4.1	Cursor-based “Reading” examination heatmap of a relevant document (a) compared to “Scanning” of a non-relevant document (b), both with equal dwell time (30 seconds).	62
4.2	An example of “Reading” a relevant long document (a) vs. “Scanning” a non-relevant long document (b).	65
4.3	NDCG at K for the <i>DTR</i> baseline and the full models with (<i>PCB_User</i>) and without (<i>PCB</i>) user normalization features in re-ranking all the pages.	85
4.4	NDCG at K for the <i>DTR</i> baseline and the full model (<i>PCB</i>) in re-ranking only the landing pages.	85
5.1	Example mouse cursor heat maps: (a) - search result page in a successful search session; (b) - search result page in a unsuccessful search session.	89
5.2	An example of zooming interaction with an Android smart phone with a touch screen: (a) the user was viewing the original picture; (b) the user zoomed in using two fingers on the touch screen.	90
5.3	Possible state transitions in QRAV model.	92

5.4	An example search game interface, which has the question, the search query window, and a dropdown box for choosing a search engine to use (Google, Yahoo! Search, or Bing). When the answer is found, the participant submits it together with the supporting URL. The query result page is opened in new tab, allowing natural querying and browsing.	94
5.5	CRF implementation of session-level model. The labels represent overall session success; the observations at each step in the sequence are the features in Table.	97
5.6	Recall-precision curves for compared algorithms, for different definitions of session success in QRAV model.	101
5.7	Success (a) and Satisfaction (b) by Task ID.	121
5.8	(a) An example page that leads to a successful task (“social networks”): the movie show times were presented on the top of the search engine result page; (b) an example page that leads to an unsuccessful task (“marta schedule”): no instant answer was presented on SERP and the official site was not optimized for the mobile setting (e.g., fonts too small)	123

List of Tables

3.1	Dataset statistics	43
3.2	Distribution of labeled general intents in the 300 search sample . . .	44
3.3	Summary of the features used for general intent detection	45
3.4	Accuracy and F1 for different methods (Problem 1)	46
3.5	Accuracy and F1 for different methods (Problem 2)	47
3.6	Distribution of the re-labeled ambiguous searches	47
3.7	Accuracy and F1 for different methods (Problem 3)	48
3.8	Accuracy and F1 for different methods (Problem 4)	50
3.9	Most important features for general intent detection (ranked by In-formation Gain)	51
3.10	Summary of the features used for representing searcher context and interactions in inferring search intent	53
3.11	Classification performance for research vs. purchase.	54
3.12	Feature ablation results for intent classification.	54
3.13	Search ad click-through statistics on all search pages (All), and for searches classified as “Research” and “Purchase”.	55
3.14	Precision, Recall, and F1 for predicting ad receptiveness within a search task	58
4.1	Feature descriptions and Pearson’s correlations with relevance Levels (** indicates statistical significance at $p < .01$ level; * indicates statistical significance at $p < .05$ level).	69

4.2	Pearson’s correlation between the predicted and actual document relevance for the single feature groups. The groups are listed in descending order of the BRT performance. (* indicates a significant improvement over all the worse-performing groups in the same column at $p < .05$ level, + indicates a significant improvement over the <i>DTR baseline</i> in the same column at $p < .05$ level)	81
4.3	Pearson’s correlation between the predicted and actual document relevance for the combined feature groups. The groups are listed in ascending order of the BRT performance. (* indicates a significant decrease in performance from <i>PCB</i> in the same column when removing the feature group at $p < .05$ level, + indicates a significant improvement over the <i>DTR baseline</i> in the same column at $p < .05$ level)	82
4.4	Pearson’s correlation between the predicted and actual document relevance when adding user normalization features (* indicates a significant improvement over the <i>DTR baseline</i> in the same column at $p < .05$ level)	83
4.5	NDCG at K for the combined feature groups with one feature group removed at a time, the groups are listed in ascending order of NDCG@10.	84
5.1	Behavior features used for CRF. “Q*” features are defined only for SERPs (if the state=Q), “R*” features are defined only for non-SERP pages. Discretization thresholds are shown in the “Bins” column. For the features used in the search behavior analysis, the aggregation function is shown in the last column.	103
5.2	Descriptive statistics for search sessions collected with the UFindIt game. The corresponding statistics from reference [10] are shown in parentheses.	104
5.3	Prediction of search session success for different levels of success in QRAV model. Relative improvement against MML+Time model is shown in parenthesis.	104

5.4	Prediction of search success by the CRF model, when adding one best-performing feature at a time.	104
5.5	Prediction of search success for real-world log using CRF trained on contest data, for success definitions in [68] and [47] respectively.	105
5.6	Coarse-grained behavior feature descriptions and Pearson's correlations with success ratings (** indicates statistical significance at $p < .01$ level; * indicates statistical significance at $p < .05$ level). . .	107
5.7	Sample fine-grained cursor feature descriptions and Pearson's correlations with success ratings (** indicates statistical significance at $p < .01$ level; * indicates statistical significance at $p < .05$ level).	108
5.8	Sample fine-grained scroll feature descriptions and Pearson's correlations with success ratings (** indicates statistical significance at $p < .01$ level; * indicates statistical significance at $p < .05$ level).	109
5.9	Accuracy, Precision, Recall and F1-measure for the full FSB model, the single feature groups, and the QCT, MB baselines. The percentage of improvement over the QCT baseline is reported for the F1-measure.	116
5.10	Accuracy, Precision, Recall and F1-measure for feature group ablation. The difference compared to the FSB full model is reported for the F1-measure.	117
5.11	Accuracy, Precision, Recall and F1-measure for individual aggregation functions. The difference compared to the FSB full model is reported for the F1-measure.	117
5.12	Task descriptions (initial queries).	119
5.13	Results of predicting success and satisfaction. Significance of differences is indicated between models and: Baseline: $\Delta p < .05$, $\blacktriangle p < .01$; Server: $\circ p < .05$, $\bullet p < .01$	122

Chapter 1

Introduction

Web search has transformed the society and now is the primary method of accessing and discovering information. Billions of people interact with Web search engines daily. These interactions with search engines convey valuable information about the needs and preferences of the users. For example, a click on a result document may indicate that the searcher was interested in the topic of the document, and the amount of the time spent on the document may further indicate how much the searcher found that the document was relevant. Furthermore, if a searcher issued multiple search queries with very few clicks, then she may be struggling in finding the needed information; in contrary, if the searcher clicked on multiple search results and spent considerable amount of time on each, she may be exploring and was successful.

Most of the existing work on modeling searcher interactions focuses on the query, result click-through and the time spent on visiting a page [82, 88, 111, 3, 144, 33, 68], and tends to ignore how searchers actually view and interact with the visited pages. However, the same search query, click and time spending patterns may be actually associated with very different examination patterns that could be indicative of different search goals or different levels of searcher satisfaction. For example, extensive examination before clicking may indicate that the search goal was more exploratory or that the searcher was less confident in the clicked result and careful reading after a click may indicate higher satisfaction with the visited document than extensive quick skimming. Being agnostic about how users actually examine the search result pages and clicked documents, the interpretation of the search

behavior tends to be inaccurate, largely limiting the capabilities of utilizing the implicit feedback provided by the users. To address this limitation, the thesis aims to model the searcher examination and interaction patterns when visiting a search engine result page (SERP) or a result document, in addition to the limited set of behavioral signals that were studied in the literature.

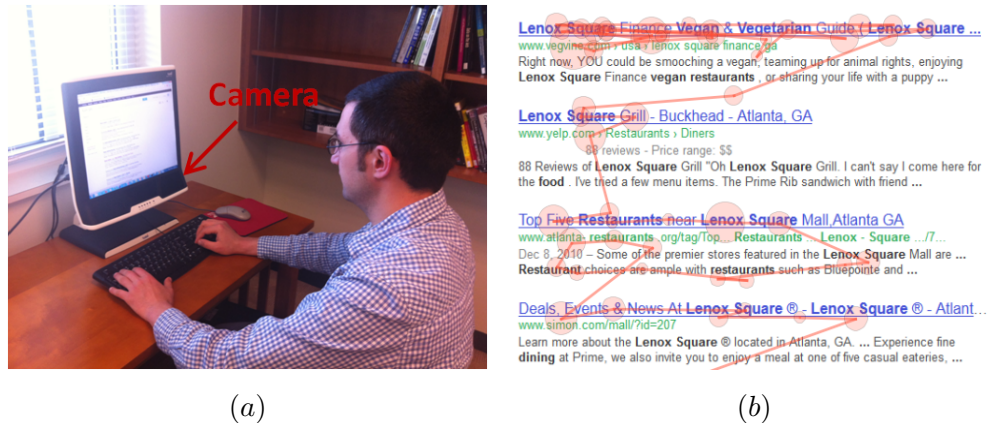


Figure 1.1: (a) A user of an eye-tracker with camera integrated into a computer monitor; (b) eye gaze trajectories captured overlaid on a search engine result page

One possible approach of modeling examination involves the use of eye-tracking devices. Figure 1.1 (a) shows a picture of using an eye-tracker with the camera integrated at the lower part of a computer monitor. Figure 1.1 (b) shows the eye gaze trajectories on a SERP captured by the eye tracker, where the pink circles represent the gaze *fixations* and the lines connecting the circles represent the gaze *saccades*. Fixations refer to spatially stable gaze periods for approximately 300 milliseconds while the saccades refer to rapid gaze movements within approximately 50 milliseconds between fixations. The larger the pink circle is, the longer the time the user spends on the particular region of the page. As we can see, the user spent most time carefully reading the first result (suggested by the horizontal gaze trajectories) while skimming through the lower-ranked results on the SERP, indicating his primary interest in the first result (even without a click). However, eye-tracking is not yet a scalable solution for modeling examination. For one, eye-tracking devices

require extra efforts from the users – before using the eye-tracker, the user needs to calibrate; while using the tracker, the user needs to stay still to prevent the eye-tracker from losing focus (otherwise re-calibration may be needed). For another, eye-tracking devices are expensive – the price of a competent eye-tracker at present is as expensive as a luxury car. As these two limitations are not likely to be addressed in the foreseeable future, more scalable alternative is needed to benefit the billions of search engine users.

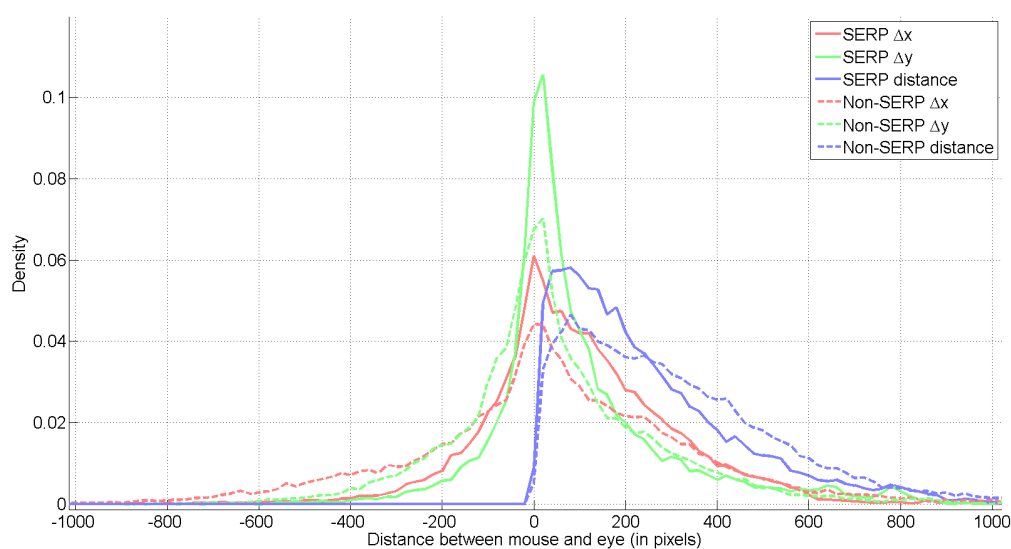


Figure 1.2: Density of distance between mouse cursor and eye gaze positions on SERPs and non-SERPs (landing pages): Euclidean distance and distances in X and Y directions.

Instead of relying on eye-tracking devices, the thesis explored the modeling of fine-grained interactions such as mouse cursor movements and scrolling behavior (or pinching, zooming, sliding behavior with a touch screen), which can be served as a proxy of eye-tracking to capture searcher’s attention. These interactions can be captured with Javascript code that could be returned as part of a SERP or with an instrumented browser plugin. Unlike eye-tracking, collecting and modeling the fine-grained interactions is scalable to billions of search engine users as it does not

require any changes in their usage (e.g., calibration, sitting still) of the search engine or purchasing additional expensive equipments. This exploration was inspired by the recent research by Rodden et al. [117, 118], where the the coordination between mouse cursor and eye gaze positions was discovered. Figure 1.2 shows the density of distance between mouse and eye on the SERPs and non-SERPs (i.e., landing pages) computed based on data collected from a user study of hundreds of search tasks performed with general Web search engines (e.g., Bing ¹, Google ²), broken down in X and Y directions. As we can see, the distribution of distance peaks around zero for both directions and both types of pages, confirming the previous findings [117, 118] that mouse and eye sometimes are coordinated (e.g., when users use mouse cursor to help reading or mark promising results) and mouse cursor movements may be a reasonable proxy of eye gaze movements.

Another limitation addressed in the thesis comes from the “one size fits all” paradigm of the existing methods – in response to the same search query, a search engine typically returns the same set of results. However, Web search queries are often ambiguous and a same query may carry different meanings for different users. For example, the query “jaguar” may refer to the third-largest feline after the tiger and the lion that could be found in Americas or the British luxury sports car brand or even the Jacksonville Jaguars American football team from Florida (if the searcher missed the last “s” in the query). In response to the diverse underlying search goals of the same query, existing methods and systems typically focus on returning the search results for the most likely intent and sometimes blend in results for other intents. Figure 1.3 shows the returned SERPs from two commercial search engines in response to the submitted query “jaguar”. As we can see, most of the top search results returned by the two search engines were about the luxury car brand, including its official website, models, pricing, and links to the local car dealers. Only a few results were about the feline and almost none was about the football team. Such search results may be satisfactory for a car shopper but are far from optimal

¹www.bing.com

²www.google.com

for people who were interested in learning about the big cat or football fans who wanted to read about the recent news of the team.

Jaguar 2012 Official Site - Compare & Configure Your Jaguar Ad

Locate a Dealer Today.
www.JaguarUSA.com


Locate A Dealer Request A Quote
Book A Test Drive Build Your Jaguar

Jaguar: Luxury Cars & Sports Cars | Jaguar USA
www.jaguarusa.com

The official home of **Jaguar** USA. Our luxury cars feature innovative designs along with legendary performance to deliver one of the top sports cars in the industry.

Models & Pricing Jaguar C-X16
Certified Pre-Owned Locate a Dealer
Jaguar XF – Sports Sedan F-Type
Jaguar XJ – Luxury Sedans Build Your Own Jaguar

Images of jaguar
bing.com/images



Jaguar International - Market selector page
www.jaguar.com

Jaguar Cars Limited: Registered Office: Abbey Road, Whitley, Coventry CV3 4LF. Registered in England No. 1672070. You need ...

Jaguar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Panthera_onca

Etiymology Taxonomy and evolution Biology and behavior Ecology
The **jaguar** is a big cat, a feline in the Panthera genus, and is the only Panthera species found in the Americas. The **jaguar** is the third-largest feline after the ...

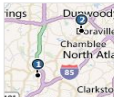
Jaguar International - Home
www.jaguar.com/g/en

Our mission at **Jaguar** has been to create and build beautiful fast cars. The XK, XF, and XJ bring the exhilaration of driving to life.

Jaguars, Jaguar Pictures, Jaguar Facts - National Geographic
animals.nationalgeographic.com/animals/mammals/jaguar

Learn all you wanted to know about **jaguars** with pictures, videos, photos, facts, and news from National Geographic.

Jaguar near Georgia 30032
bing.com/local



1. Jaguar Hennessy - Website - (404) 261-5700
3040 Piedmont Rd NE - Atlanta - Directions
2. Jaguar East - (770) 451-4921
3701 Longview Dr - Chamblee - Directions

Ads related to jaguar

Jaguar® Official Site | jaguarusa.com
www.jaguarusa.com/
Explore The Luxurious **Jaguar** Models On Our Official Site.
351,835 people +1'd or follow Jaguar USA

Locate a Dealer Build & Price
Request a Quote View Offers
Schedule a Test Drive Jaguar F-TYPE

Consider a Mercedes-Benz® | MBUSA.com
www.mbusa.com/Atlanta
Discover the Safety & Performance Innovations That Set the Benchmark.
489 people +1'd this page
Coupes - Build Your Own - Compare Vehicles - Sedans

Jaguar: Luxury Cars & Sports Cars | Jaguar USA
www.jaguarusa.com/
The official home of **Jaguar** USA. Our luxury cars feature innovative designs along with legendary performance to deliver one of the top sports cars in the ...

Models & Pricing
Available in four extraordinary models
(XF, XF Portfolio, XF ...
More results from jaguarusa.com »

Luxury Coupes
The Jaguar XK, in coupe or convertible, combines Jaguar's ...

Jaguar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Jaguar

The **jaguar** is a big cat, a feline in the Panthera genus, and is the only Panthera species found in the Americas. The **jaguar** is the third-largest feline after the tiger ...
Jaguar Cars - Jaguar (disambiguation) - Jacksonville Jaguars - Jaguarundi

Jaguar Cars - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Jaguar_Cars

Jaguar Cars Ltd, known simply as **Jaguar** is a British luxury and sports car manufacturer, headquartered in Whitley, Coventry, England. It is part of the **Jaguar** ...

Hennessy Jaguar Atlanta
www.hennessyjaguaratlanta.com/
3 Google reviews

3040 Piedmont Road
Northeast
Atlanta
(404) 261-5700

Hennessy Jaguar Gwinnett
www.hennessyjaguarwinnett.com/
Google+ page

3393 Old Norcross Road
Duluth
(770) 680-3000

Marietta Sportscar
www.mariettasportscar.com/
Score: 26 / 30 - 17 Google reviews

1616 Roswell Road
Marietta
(770) 420-8787

Euro Performance, Inc.
europerformanceinc.com/
1 Google review

4489 Tilly Mill Road
Atlanta
(770) 451-6969

DENEZ
plus.google.com
Google+ page

5098 Miller Rd
Decatur

(a)

(b)

Figure 1.3: Top search results from two commercial search engines for query “jaguar”

To address this “one size does not fit all” problem, search engines such as Bing and Google have started to *personalize* the search results in recent years by profiling users based on their long-term search history (e.g., previous queries and result click-through), promoting result documents that were frequently visited (or that are similar to those frequently visited documents) for each of the sign-on users or users with their browser cookies enabled. Search *personalization* certainly alleviates the problem of diversity in information needs but suffers from two limitations. First, the same user may use the same query to reflect different information needs – for ex-

ample, a user who is interested in both basketball and machine learning may query “michael jordan” for either the former Chicago Bulls star or the Berkeley professor at different times. Second, storing the long-term search history for personalization may raise the privacy concerns for the search engine users [126, 148, 49]. In order to address these limitations, the thesis focuses on the modeling of search context in a session (or, in other words, the short-term search history). For example, if the user searched for “scottie pippen” prior to “michael jordan”, then she is more likely to be interested in the basketball player in the current session; in contrast, if the previous query was “graphical models” then the current search query probably refers to the machine learning professor. Modeling the session context also limits the storage of search history, substantially reducing the concern of privacy.

In summary, the thesis focuses on modeling a rich set of searcher interactions and the session-level search context to improve the understanding of immediate searcher goals and preferences for enabling more intelligent information retrieval (IR) systems and better search experience. Specifically, the thesis explored the opportunities of improvements in the three fundamental areas of Web IR, namely, intent inference, ranking and evaluation.

The connections of these three areas can be better viewed through the Web IR model in Figure 1.4: a user comes to the search engine with a task in mind, which is associated with a information need (or a few information needs), and then verbalize each information need (usually mentally, not loud) and translate it into a search query in some languages. Given the submitted query, the search engine attempts to infer the underlying information need through the *intent inference* module and returns a ranked list of results selected from the *Corpus* (i.e., a collection of Web documents, crawled and indexed by the back-end modules of the search engine) by certain matching mechanisms specified in the *Ranking* module. Then the user interacts with the returned results by examining and possibly clicking and viewing one or more of them (*interaction*), and a query refinement process may be needed to create new queries to refine the results. To gauge and improve the performance of the search engine, the *evaluation* module records the queries and corresponding

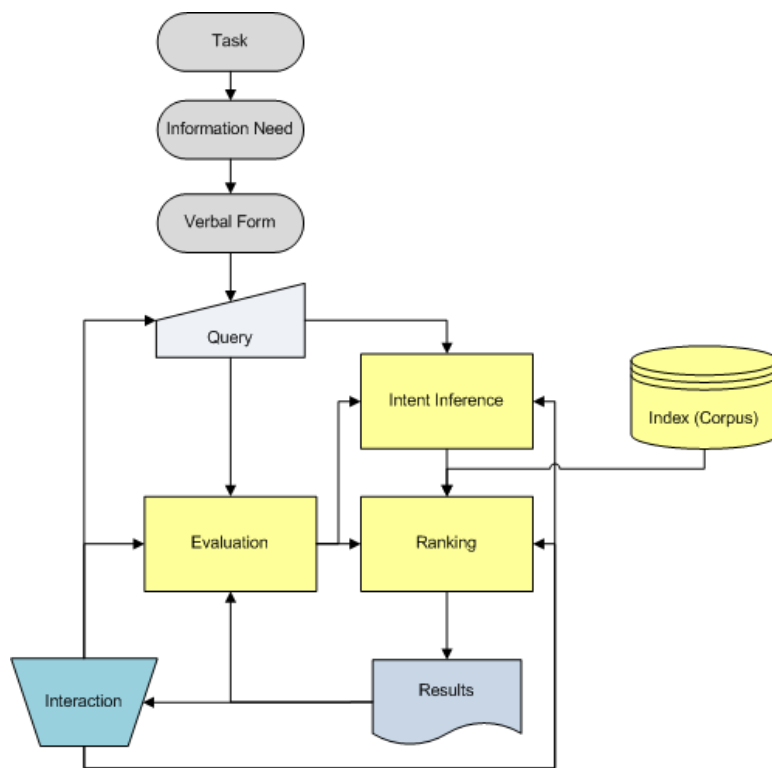


Figure 1.4: A refined classic Web IR model

search results for further analysis, often through human judgments. The output of the evaluation modules typically includes insights on improving the different search engine modules, such as suggestions on modifying the intent inference and ranking algorithms.

Search interaction, the topic of the thesis, may contribute to these three interrelated areas, enabling improvements of the search engine in different aspects. Improving the intent inference module through interpreting interaction data allows the improvement in the earliest stage, which may allow improvements in all the following stages, including selecting, ranking and presenting the search results and evaluating the result quality. Interaction data may also be used to improve the ranking module through direct estimation of document relevance, which may be the most effective in improving search result quality but the number of impacted areas

tends to be less than improving the intent inference module, which even though has broader impact but may suffer from error propagation to limit effectiveness. While interaction data allows for improving intent inference and ranking models more directly (e.g., through updating feature weights for a machine learning model), more substantial improvements on the underlying search engine components (e.g., crawling and indexing algorithms) may be possible through automatic evaluation and diagnosis based on the interaction data.

1.1 Contributions

The main contributions in the thesis are summarized as follows:

- **Techniques for inferring search information needs from interaction data:** The thesis develops models of rich searcher interactions, including query, click-through, time, and pre-click examination through interactions such as mouse cursor movements on a SERP, contextualized in a search session. These models are used to predict the immediate search intent along multiple dimensions, including general search intent [56, 55], commercial search intent, and search advertising receptiveness [57, 58] (Chapter 3).
- **Techniques for estimating result relevance from interaction data:** The thesis develops models of the rich post-click examination and interaction patterns, that are indicative of reading and skimming behavior. These models are used to estimate the “intrinsic” relevance of a visited landing page or a subsequently visited page on the search trail [60] (Chapter 4).
- **Techniques for predicting session-level search success from interaction data:** The thesis develops a principled framework (UFindIt) with different search success definitions [1] and reproducible large-scale remote user studies for modeling session-level search success. The thesis also introduces the *Fine-grained Session Behavior (FSB)* model that aggregates rich interaction data for the entire session and captures both pre- and post-click behavior [62]. In

addition, the thesis explores touch screen based interaction models for mobile devices [65] (Chapter 5).

- **Insights into modeling fine-grained search interaction for improving Web search:** The thesis provides insights into the characteristics of various behavioral signals and their effectiveness for different applications in the context of Web search. Finally, the thesis considers the integration of the fine-grained interaction models into a production search system, and the issues of deploying and evaluating such a system (Chapter 6).

1.2 Organization

The thesis is organized as follows: In Chapter 2, related work is reviewed to put the thesis in context. Then, Chapter 3 presents the techniques for detecting general intent, commercial intent, and future search behavior (e.g., ad click-through). In Chapter 4, the PCB model is introduced, which incorporates fine-grained post-click behavior to improve the estimation of document relevance and search result ranking. In Chapter 5, the techniques for predicting session-level search success are presented, including the UFindIt framework for large-scale remote user studies, QRAV success model, FSB fine-grained interaction model, and a touch screen based interaction model of mobile search success. In Chapter 6, the thesis is concluded with a summary of the findings, a discussion about system integration and limitations of the proposed techniques, and future research directions.

Chapter 2

Background and Related Work

The origins of user modeling research can be traced to library and information science research of the 1980s. An excellent overview of the traditional “pre-Web” user modeling research is available in [14]. With the explosion of the popularity of the Web and the increasing availability of large amounts of user data, Web usage mining has become an active area of research. The related work of the thesis centers around *user modeling* and spans the three key research areas in Web information retrieval, namely, *inferring search intent*, *estimating document relevance*, and *evaluating search experience*.

2.1 Inferring Search Intent

Inferring user intent in Web search has been studied extensively, which can be broadly categorized into three dimensions, namely, the general intent detection [94, 11, 79, 21], the topical intent detection [121, 25, 97, 30, 113] and commercial intent detection [42, 9, 8].

2.1.1 General Search Intent

For general intent detection, there is a broad consensus on the top 3 levels of intent taxonomy, namely the navigational, transactional and informational intents introduced by Broder [23]. Later on, a more specific query intent classification was presented in [119], where informational and resource (transactional) user goals are

further divided into specific sub-goals. Lee et al. [94] proposed a way to automatically classify queries into navigational and informational using user-click behavior and anchor-link distribution, while achieving high precision when considering only the “predictable” queries from their dataset of 50 queries, 40% of the queries were found “unpredictable”, suggesting substantial ambiguity lies in Web search queries. Baeza-Yates et al. [11] experimented with a larger set of 6000 popular queries and proposed to classify queries into informational, non-informational and ambiguous, using supervised and unsupervised machine-learning algorithms applied over query vector enriched by the clicked documents. To better understand query ambiguity in the general intent dimension, Wang and Agichtein [137] proposed to use click entropy and found that ambiguity can be considered as an orthogonal dimension to the general intent dimension – for example, query “people” may be both ambiguous and navigational as different users may want to navigate to different sites, while the query “lyrics” may be both clear and informational as the underlying search goal for different users is probably the same. While many of the proposed techniques do not aim to detecting transactional intent, Jansen et al. proposed [79] a rule-based approach that classifies queries into all the three classes (i.e., navigational, informational and transactional), which does not require expensive external sources to train. A more comprehensive survey and comparisons of the different general intent detection algorithms can be found in [21]. Bian et al. [19] proposed to use query-dependent loss functions and jointly learn the ranking function and query categorization (in the general intent dimension), achieving improvements over ranking functions without query categorization. Most recently, Lin et al. [98] investigated the non-navigational search intent at a finer level, proposing the notion of *actions*, which aim to capture specific intents that are performed on entities. They found that 28% of the queries are navigational (with corresponding action that is to visit a specific website) while 57% of the queries are either transactional or informational and bear needs regarding non-website entities. To better support the intent-based actions, Lin et al. developed a graphical model based on queries and clicks to recommend actions (e.g., reading reviews, shopping online) for the non-navigational

entity-bearing queries.

2.1.2 Topical Search Intent

The line of topical intent detection research is inspired by the KDDCup 2005 query classification challenge, where 67 query topics were given as the target categories for classification. The KDDCup 2005 winning team Shen et al. [122] proposed to build a bridging classifier using the Open Directory Project ¹ (ODP) taxonomy as the intermediate taxonomy, which can be used to map the search result enriched user queries to the target categories. Broder et al. [25] extended this work by significantly expanding the query topics to approximately 6000 and proposed to directly use the search result enriched queries without creating mappings between different taxonomies. Instead of enriching feature representation, Li et al. [97] focused on increasing the amounts of training data by semi-supervised learning with click graphs and demonstrated the effectiveness of the proposed approach in detecting product-related and job-related intent. Also utilizing the click graphs, Radlinsky et al. [113] proposed a three step approach, which begins with the expansion based on query reformulations and follows with random-walk on the click graph to filter and generate query intent clusters. Bian et al. [18] proposed a divide-and-conquer framework for ranking specialization, which defines a global ranking functions by combining risks from all query topics. Another extension of this line of work is the problem of *vertical selection*, introduced by Diaz [43] and further extended by Arguello et al. [7, 6], where the goal is not only to detect whether the search intent belongs to a particular vertical but also jointly decide whether the particular vertical contains high quality content to be incorporated into the organic search results.

2.1.3 Commercial Search Intent

The problem of commercial intent detection was firstly introduced by Dai et.al [42], where the authors proposed to enrich the queries with search results to detect whether

¹<http://www.dmoz.org/>

a query contains commercial intent. Ashkan et al. [9, 8] developed techniques to use both the ad click-through in addition to search result enriched queries to jointly detect general search intent and commercial intent and studied the effects of the number of displayed ads and ad click-through patterns. Related to this thread of research is the work on predicting ad click-through, most of which focuses on learning from the content of displayed ads (e.g., [37, 115]). Becker et al. [13] considered the result (and ad) relative position and presentation features to improve click-through estimation, within a single page.

Despite the success in detecting search intent, the above approaches are not sufficient as they only focus on predicting the majority intent and do not take into account the session-level search context and the individual user behavior. It has been shown that specific user goals and experience vary widely and have substantial effect on user behavior [140]. Some queries have substantial variation in intent [132], and searcher behavior can help distinguish user intent in such ambiguous cases. An early attempt in addressing this issue in the dimension of query topics was from Cao et al. [30], where the authors modeled the query and click sequence in the session using sequential models such as conditional random fields. Extending this work, White et al. [138] conducted a comprehensive analysis on finding the optimal weights to combine the query and context models (based on recent queries, clicks and subsequently browsed pages) to predict short-term search interests. This thread of research in detecting search intent is the closest to the proposed techniques described in Chapter 3, but, in contrast, addressed are two different dimensions in detecting search intent, namely, the general search intent and commercial search intent; also, the proposed techniques represent the user interactions in a much richer and finer-grained way, allowing for more effective models.

Recently, eye tracking has started to emerge as a useful technology for understanding some of the mechanisms behind user behavior (e.g., [41, 83, 91]). While informative, the eye-tracking instrumentation is not scalable. To address this issue, a model to estimate searcher's viewing behavior based on observable click data was introduced by Wang et al. [134]. Another thread of work aiming to address the

eye-tracking scalability issue was introduced by Rodden et al., where the authors observed the correlation between mouse and eye movements [118].

The proposed techniques in inferring search intent expands on the observations described above by exploiting the additional evidence from the fine-grained interactions and session-level context. In contrast to most of the previous work in intent detection, the focus is on predicting the *immediate search goal* in a search session, specifically in the dimensions of general search intent (e.g., navigational vs. informational) and the commercial search intent (e.g., research vs. purchase), instead of detecting the majority goal of a search query. The proposed model is also capable of predicting *future behavior*, e.g., whether the user is likely to click on an ad sometime during the current session, allowing search engines to adjust retrieval algorithms on the fly to customize for the immediate information needs of a searcher.

2.2 Estimating Document Relevance

Ranking or estimating document relevance is at the core of Web IR. Web search engines typically rely on various sources of evidence to generate search result ranking. The most common and basic ones are the similarity between the submitted query (or the corresponding inferred intent) with the result document (e.g., BM25 [116]) and the quality of the document (e.g., PageRank [22]). With increasing popularity, implicit feedback (i.e., search interaction) has become widely used to improve search result ranking. A good overview of different implicit measures studied in the previous research is described in Kelly and Teevan [89]. In particular, extensive research has been conducted in exploiting result click-through and document dwell time as implicit feedback while research in modeling examination (e.g., through eye-tracking and fine-grained interaction) in Web search has started only recently.

2.2.1 Click-through

One of the earliest research in modeling click-through is by Joachims [82], where the author proposed a Support Vector Machine (SVM) algorithm to learn ranking functions from result click-through. Later, Joachims et al. [83] presented empirical evaluation of interpreting click-through evidence through controlled eye-tracking studies in the laboratory setting and derived a variety of strategies (e.g., Click > Skip Above) to characterize search result preference based on the click patterns and the insights drawn from eye-tracking studies. One key finding was that the higher-ranked results tend to receive more attention and click-through and, without taking this position bias or *trust bias* into account, click-through would be a noisy relevance signal. Agichtein et al. [3] proposed to adjust the click-through rate (CTR, computed through aggregating over multiple users) by modeling behavior deviations for a document at a given rank from the expected behavior (i.e., interaction with all documents at the same rank) and to enrich click-through signals through model browsing and time patterns, demonstrating improved relevance estimation compared to the strategies proposed by Joachims et al. [83]. Agichtein et al. also incorporated the richer behavior representation into learning a Web search ranking function using neural networks and demonstrated substantial improvements over a commercial search engine in a large-scale evaluation [2, 4]. In addition to SVM and neural network, the most recent developments demonstrated the effectiveness of boosting based algorithms, such as GBRank proposed by Zheng et al. [150] and LambdaRank proposed by Burges et al. [26], for learning ranking functions.

To address the position bias (among the various presentation biases) of result click-through, much research was conducted to develop click models (i.e., models that estimate the probability of click) and estimate the examination. Aiming to predicting ad CTR, Richardson et al. [115] proposed to model click and viewing behavior separately, and assumed that a user has to view a document or ad before clicking on it, which, known as *examination hypothesis*, is served as the basis of many later developed click models. Crawswell et al. [39] extended the examina-

tion hypothesis and proposed the Cascade Model by assuming the linear traversal of result examination that is ended with the first click or abandonment. Dupret and Piwowarski proposed a User Browsing Model (UBM) which aimed to model the examination variations due to different search intents (e.g., navigational vs. informational [23]) and extended the model to handling scenarios with multiple clicks. Guo et. al [54] developed Dependent Click Model (DCM) which also extended the cascade model by handling multiple clicks in a query session and featured a simpler specification and more efficient algorithms. Srikant et al. [128] modeled the multi-click scenario from a different perspective, demonstrating the change of behavior due to the change of relevance caused by prior clicks. Lagun and Agichtein [92] verified the change of viewing behavior due to the change in overall result quality through an eye-tracking study, showing viewing pattern such as the time spent on viewing the first result is a good indicator of overall result quality. To enable large-scale remote studies of Web search examination, Lagun and Agichtein [91] proposed the *viewer* system, which only allows view port around the mouse cursor and blurs the rest of the search results. As shown in the study, the viewing behavior through the constrained view port in the viewer system is a good approximation of natural viewing behavior, resulting in better estimation of document attractiveness (i.e., perceived relevance) through the click over view (COV) measure. Most recently, Huang et al. [76] incorporated pre-click mousing and scrolling behavior into a click model, demonstrating higher accuracy in predicting future click-through. Another related effort in estimating examination is by Wang et al. [134], where the partially observable markov (POM) model was developed to estimate (un-clicked) result viewing behavior from click patterns and the POM model was further extended by He and Wang [70] through incorporating time information.

Another major bias of click-through lies in the *attractiveness* of search result snippet [33] – in other words, the user click on a result based on its *perceived relevance* other than the *intrinsic* relevance [47] – therefore, a click does not necessarily imply satisfaction with a result document [33]. Chappelle and Zhang [33] proposed the Dynamic Bayesian Network (DBN) click model to model both the attractive-

ness and satisfaction of a result document with the assumption that a user ends his or her query session when he or she is satisfied – that is, only the last click in a query session indicates satisfaction or actual relevance. A similar model was developed by Guo et al. [53] under the Bayesian framework, featuring more efficient probabilistic inference as compared to the iterative inference algorithms for the DBN model. Another effort in this direction is by Zhong et al. [151], where the authors incorporated the dwell time information into a click model to improve estimation of document relevance. Extending this thread of research, Dupret and Liao [47] proposed the utility model to better estimate the intrinsic relevance of a document, assuming multiple clicks in a session as a utility accumulation process. While search ending often indicate satisfaction, sometimes it may also imply a failure, in which case, no utility may be gained through any of the clicks in the session. To address this problem, Hassan et al. [69] developed session-level success prediction model to improve relevance estimation, which exploits click utility only when a search session was successful. Other advancements in click models addressed behavioral variations introduced from various aspects, such as branching (i.e., multi-tab usage) [74], vertical aggregation [35, 34], different intents [73] and different users [123].

2.2.2 Dwell Time

Dwell time, as a post-click implicit feedback measure, is another possible solution (other than modeling behavior sequence in a session) to the presentation biases of result click-through. As shown in previous work [99], dwell time approximately follows a Weibull distribution and different factors (such as content and layout) may have influence on the dwell time on a document. Using document dwell time for inferring relevance has a long history in the information retrieval community, with mixed conclusions about its utility. Some of the first research done in the area of implicit feedback in information retrieval was that of Morita and Shinoda [106]. They conducted a study where participants were asked to provide explicit feedback

about interestingness of news articles that they have read. The study focused on the correlation between reading time and explicit feedback while considering document length and additional textual features. They noted that there is a strong tendency to spend more time on interesting articles rather than on uninteresting ones. Similar findings have also been reported in [38] and [51]. Furthermore, Morita and Shinoda found only a very weak correlation between the lengths of articles and associated reading times, indicating that most articles are only read in parts, not in their entirety.

Interestingly, dwell time does not always correlate with relevance. Kelly and Belkin [87] tried to reproduce the results of Morita and Shinoda in a different, more complex information retrieval scenario, yet found no correlations between display time and explicit relevance ratings for a document. In a subsequent, naturalistic study, Kelly and Belkin [88] found again no general relationship between display time and the users' explicit ratings of the documents' usefulness. Instead, they observed high variation of display time with respect to different users and different tasks. Following this study, White and Kelly [144] reported that adjusting display time thresholds for implicit feedback according to task type leads to improved retrieval performance, while adjusting the thresholds according to individual users degraded performance. This stands in contrast to findings of a prior study by Rafter and Smyth [114] who showed for one specific task type that display time is correlated with user interest, especially after individually adjusting the measure. In summary, while dwell time clearly contains some relevance signal, numerous previous studies have found almost as many different interpretations of it with no clear consensus of the relationship to relevance of the document.

2.2.3 Examination

The proposed techniques in estimating document relevance build on previous research on connecting searcher examination patterns to user interest and document relevance. In particular, eye tracking studies have been helpful for understand-

ing common patterns in search result examination (e.g., [83, 41]). To operationalize these insights, the proposed techniques exploit the coordination between the searcher gaze position and mouse movement over the search results, shown previously in references [117, 118, 59, 75]. The mouse cursor movements have shown to be useful in various applications, such as inferring searcher intent [56, 58] and search result preferences [77, 76, 139], and inferring user attention in complex Web pages containing images, text and varied content [108].

Additional implicit measures have been examined on the object level (e.g., document paragraph or page item) as well. On one hand, it has been found that good indicators of interest include the amount of scrolling on a page [38], click-through [51, 83], and exit type for a Web page [51]. On the other hand, mouse movements and mouse clicks while viewing a document do appear to provide some correlation to user interest [38]. Furthermore, user behavior on the SERP, when combined with page dwell-time and session level information, can significantly improve result ranking in the aggregate (e.g., [2]), and can be further improved by personalizing these measures (e.g., [105]).

Other previous efforts focused on modeling more explicit user interactions on the page. Golovchinsky et al. [52] focused on user-created annotations on documents such as highlightings, underlinings, circles, and notes in margin. They used this kind of feedback to infer relevance of document passages. In a document search scenario utilizing query expansion, they reported a significant improvement of the annotation-based feedback technique over explicit relevance feedback on the document level. Ahn et al. [5] followed a similar idea but used the concept of a personal notebook where users could paste text passages worth remembering. On the basis of the text passages they built up term-based task profiles which were then used for re-ranking search result lists. Compared to a baseline ranking function not considering any feedback, the task-profile-based ranking performed significantly better. The previous two approaches both need more or less explicit and therefore rare user interactions (i.e., annotating, copying and pasting) to work properly. Buscher et al. [27] only rely on implicit data and determine which parts of a doc-

ument have been read, skimmed, or skipped by interpreting eye movements. Read and skimmed parts were taken as relevant while skipped document parts were ignored. They report considerable improvements concerning re-ranking of result lists when including gaze-based feedback on the segment level compared to relevance feedback on the document level. Gyllstrom and Soules [66] follow a similar idea, but consider all text that has been visible on the screen for building up term-based task profiles. They use such profiles for task-based indexing of documents on the desktop and show that re-finding documents that way is more effective compared to simple desktop search.

2.2.4 Personalization

Web search personalization is the problem of customizing search result ranking according to individual user's interests. The earliest personalization in Web search IR (e.g., early version of Google personalized search) depends on users to specify the interested topics. Aiming to automatically detect such topics of interest for individual users, Liu et al. [100] proposed to collect user search history and profile users by mapping their previous search queries into categories in the Open Directory Project (ODP). To enable scalable computation of personalized PageRank scores [22], Jeh et al. [81] proposed a technique that encodes personalized views as partial vectors and allows the construction of personalized views at query time. Dumais et al. [46] developed Stuff I've Seen (SIS) system to facilitate personal information re-use by indexing various entities such as emails, web pages, documents that a user has seen before. Sugiyama et al. [129] proposed to profile user browse history and adapted Collaborative Filtering algorithms to construct profiles and personalize search results. Teevan et al. [131] proposed a richer user representation to re-rank search results, which models information such as documents and emails a user has viewed and created in addition to his or her search and browse history. Tan et al. [130] proposed statistical language modeling based methods to represent and mine the long-term search history to more accurately estimate the current query language model

for more effective search result ranking. Dou et al. [45] performed a large-scale evaluation of different personalized strategies, finding click-based personalization strategies perform consistently well while profile-based ones are unstable. The importance of short-term search context was found to be also important in addition to the long-term modeling or user profiling. Matthijs and Radlinski [104] proposed to model long-term browsing history using rich representations of documents (e.g., title, content) and evaluated their re-ranking methods using an interleaving technique which measures result click-through on SERPs with merged results from the original and personalized rankings. Sontag et al. [127] propose a generative model of relevance based on long-term search history to personalize result rankings and evaluated the proposed techniques with history search data at a large scale.

2.2.5 Search Context

Context-aware Web search ranking aims to exploit the context (e.g., short-term search history) in the current session as compared to exploiting the long-term user search and browse history as in personalization. While personalization helps in retrieving documents that are of general interest to the user, modeling context enables more accurate retrieval in response to the immediate information need. Also, personalization may result in higher concern in privacy as it requires the storage of larger amount of personal data [126].

One of the earliest efforts in context-aware Web search ranking is Shen et al. [125], where the authors proposed several retrieval algorithms based on statistical language models to combine the previous queries and clicked document summaries with the current query to improve search result ranking. Extending this study, Shen et al. [124] introduced a theoretic framework of context-aware user modeling and developed a client-side Web search agent UCAIR that can perform eager implicit feedback in query expansion based on previous queries and result re-ranking based on click-through information. To address various types of search context (e.g., reformulation, specialization, and generalization), Xiang et al. [147] proposed various

heuristics for context-aware Web search result ranking and validated the proposed techniques with both human judgments and user click data at a large scale. Aiming to bridging the gap between the two related areas, Bennett et al. [17] studied how the modeling of session context (i.e., short-term search history) interact with personalization (i.e., long-term search history), finding that the two sources of information are complementary and that the long-term history is particularly valuable at the start of a search session while the short-term context is particular useful for an extended search session.

Most closely related to the proposed techniques, Huang and White [77] found correlations between cursor hovering over some of the results on the Search Engine Result Page (SERP) and result relevance. Complementary to previous efforts, the proposed work is the first to analyze the examination patterns, and relevance, from rich *post-click* searcher behavior such as cursor movements on landing pages and subsequently viewed documents, and the first to develop a predictive model, PCB, that captures these patterns. As Chapter 4 demonstrates, PCB can provide significant improvements for estimating document relevance and consequently for improving search result rankings. While the proposed techniques focus on modeling the implicit feedback and search context in a session, the modeling of examination through fine-grained interactions can be also applied to improve long-term personalization and combining short-term and long-term history is likely to achieve further improvements in Web search ranking as suggested by the previous studies [45, 17].

2.3 Evaluating Search Experience

The research on automatic techniques in evaluating can be categorized in three different levels, namely: query-level, session-level and the level of using multiple search engines.

2.3.1 Query-level

Research on query-level evaluation has been conducted to understand differences in the quality of search results for individual queries in aggregate. Such predictions can be used to devote additional resources or alternative methods to improve search results for difficult queries. While it has been shown that using different query representations [15] or retrieval models [12] improves search performance, it is more challenging to accurately predict which methods to use for a particular query. Measures such as query clarity [40], Jensen-Shannon divergence [31], and weighted information gain [152] have been developed to predict performance on a query (as measured by average precision, for example). Yom-tov et al. [149] proposed to estimate query difficulty through histogram-based and tree-based approaches and demonstrated the effectiveness of the difficulty estimation in improving information retrieval, detecting missing content and merging results for distributed IR systems. Leskovec et al. [95] used graphical properties of the link structure of the result set to predict the quality of the result set and the likelihood of query reformulation. Teevan et al. [132] developed methods to predict which queries could most benefit from personalization. While most of the techniques in predicting query performance depends on analyzing query, returned results, and document collection, Guo et al. [63] developed techniques that incorporate searcher interactions such as clicks and time, demonstrating valuable information carried in the interaction data. To more accurately quantify the result quality in response to a query, Wang et al. [135] proposed the “pSkip” metric, which estimates the probability of skipping based on the user click-through, and validated their proposed metric with eye-tracking data collected from user studies.

2.3.2 Session-level

As search query tends to be ambiguous and typically only represents a fraction of the overall information need, evaluating the query-level performance may not provide a full picture about the search experience. To this end, Hassan et al. [68]

developed Markov models to predict search success at the session-level, demonstrating the additional benefits of session-level evaluation over the Feild et al. [50] developed methods to predict user frustration, and showed that features capable of accurately predicting engine switching events were also highly predictive of frustration. To utilize unlabeled data, Hassan [67] proposed semi-supervised methods to predict Web search success. In addition, Hassan et al. [69] that the success prediction models can also be applied to improve document relevance estimation.

A special case of studying search success is the research on search abandonment, where the searcher leaves the SERP without clicking on any search results. The abandonment could be good, in which case the searcher found the needed information on the SERP, or bad, in which case the searcher did not find anything relevant and gave up [96]. Diriye et al. [44] studied the reasons why searchers abandoned their searches and developed techniques to predict abandonment rationales using features from query, result and interaction in a session.

Community Question Answering (CQA) sites such as Yahoo! Answers² and Quora³ are served as alternatives to address information needs that might not have answers on the Web. Often, the answers on those sites are returned as results in Web search. Some other times, searchers may even become askers on CQA sites when Web search fails [102]. While the asker and answerers usage history on the CQA sites were found to be major indicators of asker satisfaction on CQA sites [103], the characteristics of query clarity, query-to-question match, and answer quality were found to be effective in predicting the satisfaction of the Web searchers that end up visiting a CQA website in a search session [101].

2.3.3 Multi-engine level

At the level of multi-engine usage, research has examined search engine switching behavior. Early research by Mukhopadhyay et al. [107] has used economic models of choice to understand whether people developed brand loyalty to a particular

²<http://answers.yahoo.com/>

³<http://www.quora.com/>

search engine, and how search engine performance (as measured by within-session switching) affected user choice. They found that dissatisfaction with search engine results had both short-term and long-term effects on search engine choice. Juan and Cheng [85] described some more recent research in which they summarize user share, user engagement and user preferences using click data from an Internet service provider. They identify three user classes (loyalists to each of the two search engines studied and switchers), and look at the consistency of engine usage patterns over time.

Heath and White [71] and Laxman et al. [93] developed models for predicting switching behavior within search sessions using sequences of user actions (e.g., query, result click, non-result click, switch) and characteristics of the pages visited (type of page and dwell time) as the input features. Heath and White [71] used a simple threshold-based approach to predict a switch action if the ratio of positive to negative examples exceeded a threshold. Using this approach they achieved high precision for low recall levels, but precision dropped off quickly at higher levels of recall. Working with the same data, Laxman et al. [93] developed a generative model based on mixtures of episode-generating hidden Markov models and achieved much higher predicative accuracy.

White et al. [146] developed methods for predicting which search engine would produce the best results for a query. For each query they represented features of the query, the title, snippets and URLs of top-ranked documents, and the results set, for results from multiple search engines, and learned a model that predicts which engine produced the best results for each query. The model was learned using a large number of queries for which explicit relevance judgments were available. One way in which such results can be leveraged is to promote the use of multiple search engines on a query-by-query basis, using the predictions of the quality of results from multiple engines. White and Dumais [141] characterized search engine switching through a large-scale survey and built predictive models of switching based on features of the pre-switch query, session, and user. White et al. [143] modeled long-term engine usage over a six-month period, and identified three user classes:

(i) those who do not switch, (ii) those who switch at some time, and (iii) those who switch back and forth between different search engines. Guo et al. [64] studied in depth the reasons why users switch search engines and developed techniques to predict in-situ engine switching rationales using features of query, and pre-switch and post-switch interaction.

The thesis extends this thread of research in developing techniques for predicting session-level searcher success from rich interaction data, including a principle framework for formally studying the success prediction problem, an infrastructure for conducting remote large-scale user studies, and fine-grained interaction models for both desktop and mobile settings. As Chapter 5 demonstrates, the proposed techniques are more effective than previous methods that exploit only a limited set of search behavior.

Chapter 3

Inferring Search Intent

An improved understanding of searcher information needs is the crucial first step for search engines to generate satisfactory search results. As briefly mentioned in Chapter 1, what makes the problem particularly daunting is that the same query may reflect different goals not only for different users, but even for the same user at different times. For example, a user may search for “surface” initially to learn about the Microsoft Surface tablet; however, days or weeks later the same user may search for “surface” to identify the best deals on actually purchasing the device. Thus, identifying the most popular or majority meaning for a query is not sufficient; rather, the challenge is to identify the intent of the given search, contextualized within a search task (e.g., buying a tablet, which may involve goals such as researching the device, comparing data plans, lookup of customer reviews, and eventual purchase).

To infer the immediate search intent, the proposed solution is to model the fine-grained interactions such as mouse movements on the search engine result page (SERP) in a search session, which is inspired by the coordination between the searcher gaze position and mouse movement discovered in recent work [117, 118].

The hypothesis is that searcher interactions such as mouse movement, hovering and scrolling can help more accurately infer searcher intent and interest in the search results. That is, like eye movements, such interactions can reflect searcher attention. This would allow estimating which parts of the SERP the user is interested in (e.g., whether the searcher is paying more attention to the organic or the sponsored results), and provide additional clues about the search intent.

To test this hypothesis, a novel model was developed of inferring searcher intent that incorporates both search context and rich interactions with the results. The

model is operationalized by converting these interactions into features, which can then be used as input to classification algorithms to infer the search intent from the interaction data.

While many other dimensions of search intent have been studied (e.g., [119, 122, 42]), the thesis focuses on two representative dimensions of search intents, namely, general search intent and commercial search intent. The first dimension, originated by Broder [23] and refined by Rose and Levionson [119], includes three top-level intent classes: *navigational*, *informational* and *transactional*, where *navigational* searches aim to find a specific Web site that the user has in mind, *informational* searches aim to find information about a topic, and *transactional* searches aim to perform some web-based activity such as downloading, gaming or shopping. The second dimension, commercial search intent, consists of two broad classes of searches: *research* and *purchase*, illustrated in the examples above. While the first dimension is the most popular categorization of search intent, successfully distinguishing between the two commercial intent classes has significant practical applications for search advertising. For example, a searcher issuing a seemingly commercial query (e.g., “surface”) may not be interested in the search ads if the organic (non-sponsored) search results are sufficient for their needs. In this case, showing ads could annoy the searcher, and contribute to “training” them to ignore the ads [24]. Thus, knowing the searcher intent (and consequently, interest in viewing sponsored results) would allow search engines to target ads better; and for advertisers to better target the appropriate population of “receptive” searchers. So, if we could infer a user’s current interests based on her search context and behavior, a search engine may then show more or fewer ads (or none at all) if the current user is in the “research” mode.

The experiments in this chapter follow a similar progression. First, the proposed interaction model is shown to be helpful in distinguishing between “navigational” and “informational” intents. Then, the proposed interaction model is shown to be also helpful in distinguish between known “research” and “purchase” commercial intents. Next, the proposed model is applied to the large-scale search data of real

users, demonstrating that the searches predicted to have “purchase” intent indeed have significantly higher ad click-through rates than those predicted to have “research” intent. Finally, the proposed model is applied to the task of predicting “advertising receptiveness” (i.e., ad click-through for an individual user within the current search session), which could have significant practical applications for commercial search engines.

In summary, the contributions of this chapter include:

- A richer model of searcher intent, that incorporates searcher interactions with session-level *state* for jointly modeling searcher goals and behavior.
- Empirical evidence that the fine-grained searcher interactions improve the accuracy of predicting search intents along multiple dimensions.
- A large-scale experimental evaluation of the proposed model on predicting ad click-through, an important problem in its own right.

The bulk of this chapter has been published as [56, 58].

3.1 Motivation

In this section, examples are provided to further motivate the proposed model in improving the general intent detection (Section 3.1.1), commercial intent detection (Section 3.1.2), and behavioral targeting in search advertising (Section 3.1.3).

3.1.1 Inferring General Search Intent

While it has been shown previously that the most popular intent of a query can be detected for sufficiently frequent queries (e.g., [78, 97, 21]), the goal here is to detect the intent of the specific *search* – that is, for queries that could plausibly be navigational or informational in intent. Specifically, to detect the immediate search goals in a search session, fine-grained interactions such as mouse movements can provide additional clues. Figure 3.1 illustrates the representative mouse trajectories

of these two types of intents. Figure 3.1 (a) shows the mouse trajectory for the navigational query “facebook”, where the user directly moves the mouse to click on the first search result. In contrast, as we can see in Figure 3.1 (b), the informational query “spanish wine” exhibits a very different pattern of mouse trajectory: the user moves the mouse slowly and is likely to examine the first two results before clicking on the third one.

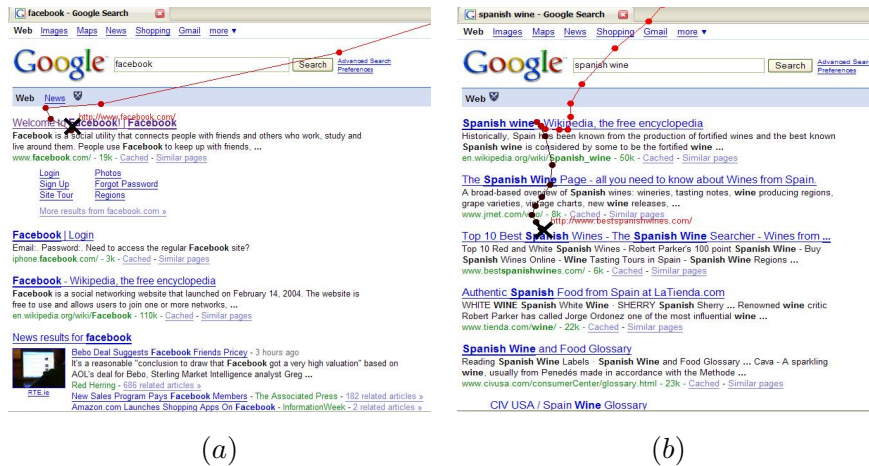


Figure 3.1: Searcher mouse trajectories on the search engine result pages for query with navigational intent: “facebook” (a) and query with informational intent: “spanish wine” (b)

3.1.2 Inferring Commercial Search Intent

In addition to the top-level categorization of general search intent, many other dimensions of search intent have been identified, including topical [119], exploratory vs. specific [119], commercial vs. non-commercial [42]. In particular, the thesis focuses on two important commercial intent categories, namely *Research* vs. *Purchase* intent.

As a concrete example, consider how users with research intent examine the SERP for a query “nikkor 24-70 review”. This query is commercial (the searcher is probably considering whether to buy this digital camera model), but could also be research-oriented (the searcher is interested in reviews, and not yet in making an

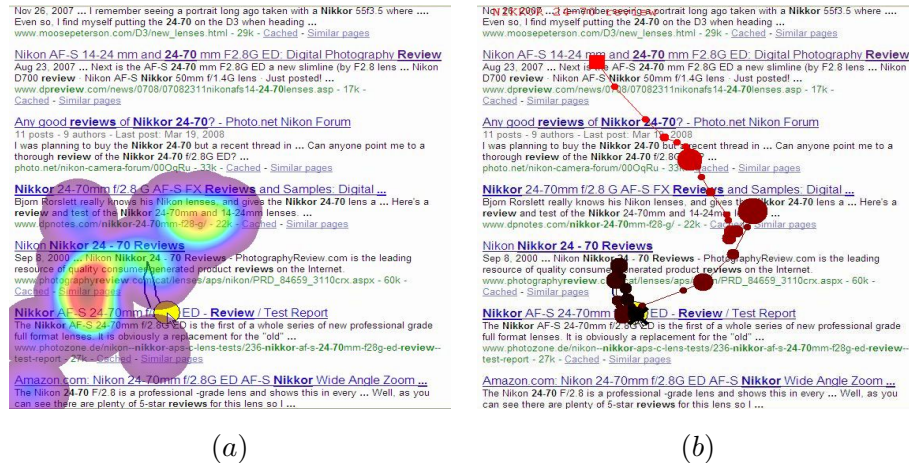


Figure 3.2: Searcher gaze position (a) and corresponding mouse trajectory (b) for query with research intent

immediate purchase). Figure 3.2 (a) shows the gaze position “heat map” (different colors represent amount of time spent examining the corresponding page position). Figure 3.2 (b) shows the mouse movements performed by the subject as they were examining the SERP. This example illustrates the possible connection between user interactions on the SERP and interest in the specific results. Thus, it is important to model not only the “popular” intent of a query, but also the searcher’s immediate intent based on the context (within a search session) as well as on the interactions with the search results. In addition to research importance, this capability has important practical applications to search advertising, as described next.

3.1.3 Application: Search Advertising

An important practical application of the proposed methods is predicting whether the user is likely to click on search ads shown next to the “organic” results. This problem is related to the research vs. purchase orientation of the user’s goals: a user is more likely to click on a search ad if they are looking to make a purchase, and less likely if they are researching a product. This observation was empirically verified by comparing the ad click-through of “research” and “purchase” searches

as classified by the proposed model.

Furthermore, one could *predict* whether the searcher is more or less likely to click on an ad in *future* searches within the current session. This idea is illustrated in Figure 3.3, which shows an example where the user hovers the mouse over the ads before she clicks on an organic result in her first search for the query “green coffee maker”. In her following search for the same query, in the same session, she clicks on an ad. This predisposition was referred to as “advertising receptiveness”, and, as shown later, the user’s interest in a search ad shown for a *future* search within the same session can be predicted based on the user interactions with the *current* search result page.

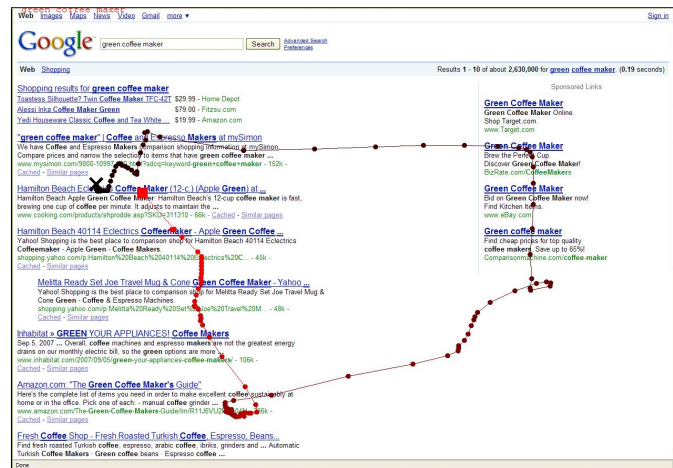


Figure 3.3: Mouse trajectory on a SERP for query “green coffee maker” with an ad click on the *next* search result page

If a search engine could be notified that a searcher is (or is not) interested in search advertising for their current task, the next results returned could be more accurately targeted towards this user. For example, if the user appears interested in buying a hybrid car, ads for hybrids as well as deals in the organic results should be returned. In contrast, if the user appears to be just *researching* hybrid technology, then the search engine should privilege customer reviews or technical articles. To achieve this real-time behavioral targeting, contextualized user interaction models

are needed.

3.2 Search and User Model

This section first describes the definitions of search tasks and search goals, then introduces the proposed approach to mine the contextualized fine-grained interactions.

3.2.1 Search Model: Tasks and Goals

The proposed work assumes a simplified model of search following recent literature (e.g., [84]), where a user is attempting to accomplish an overall search task by solving specific search goals, as illustrated in Figure 3.4.

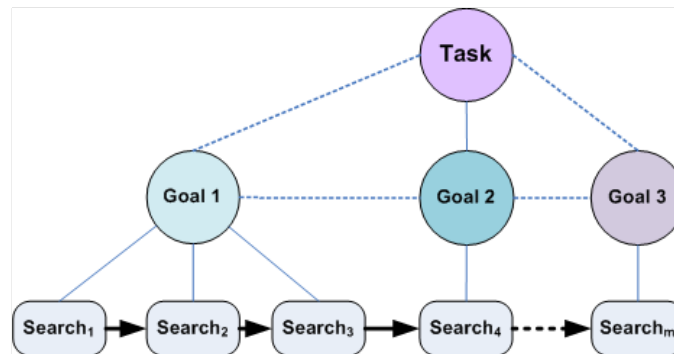


Figure 3.4: Relationship between a search task, immediate goals and specific searches to accomplish each goal.

Many user information needs require multiple searches until the needed information is found. Thus, it is natural to organize individual queries into overall tasks and immediate goals. For this, the idea of a search *task* (i.e., an extended information need) was used, which, in turn, requires more *immediate goals* (i.e., atomic information needs) to accomplish by submitting and examining related *searches*. The operational definition of a search task is that it consists of a consecutive sequence of queries that share at least one non-stopword term with any previous query within

the task. An example search session consisting of two search tasks is reported in Figure 3.5. This simple definition of a search task was verified manually, and out of more than 100 tasks examined, in all but 3 tasks the searches shared at least one non-stopword term with some other search in the task. In the dataset used in this study, while the 30-minute sessions tend to be 6.77 searches long on average, tasks tend to contain 2.71 searches on average, which is consistent with previous finding [140] that users perform on average only two or three query reformulations before giving up.



Figure 3.5: An example user session, consisting of two consecutive disjoint search tasks.

3.2.2 User Model: Goal-driven Search

The proposed user model naturally follows the search model. A user, while solving a search task, has a number of *immediate goals*. While these goals are “hidden” - that is, not directly observable, the user *searches* (queries) and their *interactions* on the corresponding search results can be observed. Thus, model a user as a non-deterministic state machine with *hidden states* representing user goals, and *observable actions* that depend on the user’s current state. The proposed model is illustrated in Figure 3.6: searcher actions such as queries, result clicks, and mouse movements are observations generated by the hidden states corresponding to the user’s goals. The interactions were restricted to those on the SERP to make this work more realistic: search engines are able to capture user interactions over their own results, but capturing actions on other pages require significant additional effort.

For example, if the immediate user goal is informational, then longer mouse trajectories are more likely to be observed on the SERP (as the user is likely to exam-

ine more results to decide which one is most relevant); in contrast, if the immediate user goal is navigational, the user can quickly recognize the target site, resulting in shorter mouse trajectory and faster response time. Similarly, ad clicks are more likely to be emitted if the user is in a receptive state to search advertising (e.g., has a Purchase goal), and less likely if the user is in a non-receptive state (e.g., has a Research goal). Hence, observations including the search context and user interactions are related to the states (goals) of the users. If the hidden states can be inferred using the observations, both the user’s immediate search goal and potentially the overall task may be recovered, as well as *predict future user actions* such as ad clicks, in *subsequent* searches.

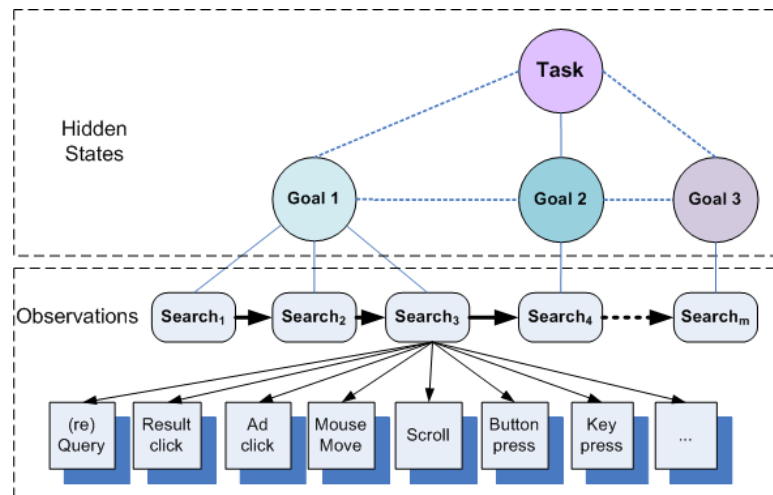


Figure 3.6: Sample states and observations for a single search within a task.

Note that the predictions in the proposed model are dependent on both the current and the *previous* goal state of the user, thus naturally allowing to maintain the user’s “mental state” across individual searches. Furthermore, this formalism allows for arbitrary number of hidden states which may correspond to different types of goals and more complex tasks. For example, this would allow modeling different variations of ad receptiveness. These sequential user models have been explored in some recent work (e.g., [30]). However, what makes the model unique is the rich repre-

sentation of user actions on the SERP *before* a click on the search result, allowing to potentially capture the mental state of the user while he or she is examining the search results.

3.3 Infrastructure, Features and Algorithms

This section describes the actual implementation of the proposed system, including: first, the infrastructure for extracting and storing user interactions (Section 3.3.1); then, the concrete representation of these interactions as features (Section 3.3.2) for incorporating the interaction information into classification algorithms (Section 3.3.3).

3.3.1 Infrastructure

The user interaction data was captured by the EMU [61] browser plug-in, which buffers the GUI events such as mouse movements, scrolling, and sends them to the server for logging. The EMU Web browser plug-ins were installed on approximately 150 public-use computers (mostly Windows PCs) at the Emory University library. The usage was tracked only for users who have explicitly opted in to participate in the study. No identifiable user information was stored.

As mentioned earlier, the instrumentation described above for the search result pages does not necessarily require a user download or installation: JavaScript code similar to what was run in the toolbar can be easily returned in the header of a Search Engine Result Page (SERP). Also, as the proposed techniques in this chapter only model searcher behavior on a SERP, and do *not* consider the external result pages visited by clicking on the results, all the data collected would be available to the search engine via light-weight server-side instrumentation.

In the current implementation, the mouse movements and scroll events are sampled at every 5 pixels moved, or every 50 ms, whichever is more frequent, and all other events (e.g., MouseDown) are kept without downsampling.

3.3.2 Features

This section describe the types of interactions captured and the corresponding feature representations.

Query group: Query text is perhaps the most intuitive and in-expensive feature available for inferring user intent. For example, an informational query is likely to contain more terms than a navigational or a transactional query, while a research query is more likely to contain the word “review” than a query with purchase intent. Query features have been used to predict search intent in previous research [122, 30], but were found to be of limited effectiveness as they tend to be sparse. For this group, the proposed features include query length in words and characters, and the unigram words of the query text, as well as binary features *IncludesTLD*, which indicates whether the query includes a TLD token such as “.com” or “.org”.

Click group: Click-through, both in aggregate and in an individual session, has shown to be useful in inferring user intent. As shown in previous research [94, 21], result click-through distribution in aggregate is indicative of different types of general search intent. For example, informational queries tend to have a more flat distribution while the click-through rate of navigational queries tend to be dominant by one URL. As for an individual session, click-through can be considered as implicit feedback from the user [11, 30], which is especially discriminative when the search query is ambiguous [30]. For this group, the aggregated features considered include the largest fraction of the result click-through, the average deliberation time (i.e., time before the first click on a search result), and the similarity between the most frequently clicked search result URL and the query (i.e., whether the query is a substring of the URL with the clicked URL, potentially indicating a navigational intent), computed over a large server-side search click log from a commercial search engine. For a individual search session, the captured features are the types and properties of result clicks, which include the unigram word tokens in the clicked URL (*ClickUrl*); the number of URLs visited after a result click (*NumBrowseAfterClick*), the average and total dwell time on each visited result URL, the number and posi-

tion in session of the satisfied (SAT) URL visits (those with dwell time greater than 30 seconds [51]) and dissatisfied (DSAT) or “bounce” visits (those with dwell time less than 15 seconds [120]), and a categorical feature *ClickUrl* indicating the type (e.g., organic result, menu item, search ad) of the click.

Interaction group: This group features aim to capture the fine-grained examination and interaction patterns before a click on the search result, which could provide additional clues about different search intents. For example, with an information intent, the searcher is more likely to move the mouse cursor more extensively and slowly before clicking on a relevant result as compared to the behavior with a navigational intent. The features considered include: the number of SERP GUI events, such as number of mouse events (*TotalMouse*), scroll events (*TotalScroll*) and keypress events (*TotalKeypress*); time features, such as SERP deliberation time, measured as seconds until first GUI event (*DeliberationTime*), the time until the first result click (*SERP dwellTime*); and hovering features, that measure how the time that the mouse hovers over an area of interest such as north ads, east ads, and organic results regions.

Also captured was the physiological characteristics hidden in mouse cursor movements, following reference [109]. In particular, the mouse trajectory representation is split into two subgroups, *Interaction (Global)* and *Interaction (Local)*, where the former include features such as the length, vertical and horizontal ranges of mouse trajectory, in pixels while the latter aims to distinguish the patterns in different stages of the user interactions. Specifically, for the *Interaction (Local)* subgroup, each mouse trajectory was split into five *segments*: initial, early, middle, late, and end. Each of the five segments contains 20% of the sample points of the trajectories. Then the same properties (e.g., speed, acceleration, slope etc.) were computed as above, but computed for each segment individually. The intuition is to capture mouse movement during 5 different stages of SERP examination (e.g., first two segments correspond to the visual search stage, and last segment corresponds to moving the mouse to click on a result). Also considered were the features describing the

general statistics of the trajectory, namely, the means and the standard deviations of the mouse coordinates, the difference in distance and time between two adjacent mouse points, the velocity, acceleration, slope and rotation (computed from the difference of slopes between neighboring points). The more sophisticated representation can capture more complex patterns of interactions. For example, with an informational intent, the mouse cursor is more likely to switch between speeding up (when the user finds something interesting and moves the mouse towards it) and slowing down (when the user begins reading or is about to click) several times and is more likely to move back and forth (rotation angles change several times) than for a navigational query. Similarly, other characteristics like the slope of the trajectory may also vary for different intents.

SERPContent group: As a popular way to enrich the query text, the search engine result page (SERP) that contains the top results of the query has been widely used as “pseudo” relevance feedback, which is shown to be effective in previous research for inferring both topical and commercial search intents [122, 42, 25, 9]. For this group, unigram word features were derived from the text content of the overall SERP (*SERPText*), and the organic results (*OrganicText*) and sponsored results (*AdText*), respectively (after frequency filtering).

ResultQuality group: These features aim to capture coarse information about the SERP relation to the query, namely how many words in the organic result summaries match the query terms (*SnippetOverlap*); how many words in the text of the ads match the query terms (*AdOverlap*); as well as the normalized versions of these features computed by dividing by the query length, in words. Also captured are the number of ads, number of ads at the top of the SERP (*NorthAds*), and number of ads on the side (*EastAds*). These features has been shown in previous work to correlate with the degree of commercial interest in the query [9, 8].

Context group: This group captures where the search belongs to within the search task, the features include: whether the query is initial in session (*IsInitialQ*), whether the query is identical to previous query (*IsSameQ*), whether the query overlap with

previous query submitted; respectively true if a word is replaced (*IsReformulatedQ*), or added (*IsExpansionQ*), or removed (*IsContractedQ*); whether the query was issued within same session (*RepeatQ*); the current position (progress) within a search session, e.g., whether this was a first, second, or 5th search in the session (*SERPIndex*).

3.3.3 Classifier Implementation

Now, the details of classifier implementations considered are provided. What was experimented with include three different families of classifiers, namely, decision trees, Support Vector Machines (SVMs) and Conditional Random Fields (CRFs). Decision trees and SVMs support flexible feature representation and model complex interactions, and CRFs naturally support modeling sequences. Note that the search-level predictions of decision trees and support vector machines can be aggregated to generate the predictions for an entire search session or task as illustrated in Figure 3.7.

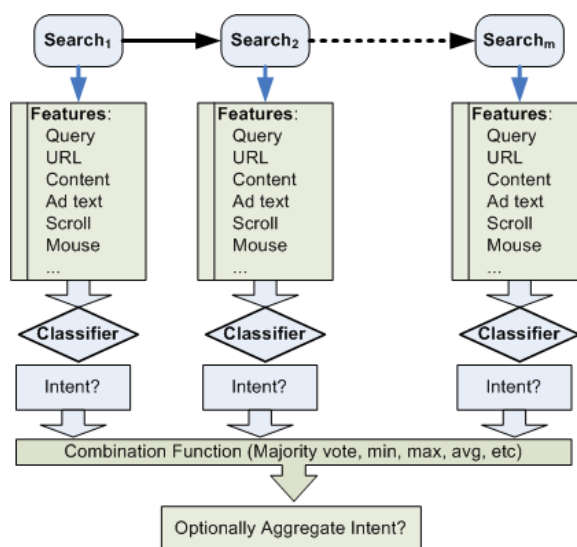


Figure 3.7: Session aggregation of search-level predictions

Decision Trees: The C4.5 algorithm was used to build the decision trees. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm avoids overfitting the data by determining how deeply to grow a decision tree. Also, it handles continuous attributes and training data with missing attribute values.

Support Vector Machine (SVM): The sequential minimal optimization algorithm was used for training a support vector classifier. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. Also, the polynomial kernel with degree 4 (chosen during preliminary development experiments) was used.

Conditional Random Field (CRF): The Decision Trees and SVM representations allow only limited representation of the search *state*: that is, whether the searcher is still in “exploratory” stage of the search or is now on the next goal of verifying specific information found during exploration “lookup” stage. To explicitly model these different states (goals) within a session, CRF allows defining a conditional probability over hidden state sequences given a particular observation sequence of searches. Take predicting future ad clicks as an example: at training, the hidden state is assigned according to whether an ad click was observed in the future searches within the task. Note that an ad click on *current* SERP is simply an observation, and is not necessarily an indication whether a user remains likely to click on future search ad in the same task. At test time, the intent sequence is identified that maximizes the conditional probability of the observation sequence.

Next, the experiments on predicting general search intent, commercial search intent and advertising receptiveness are described in three parallel sections. In each section, the problem statement, data, methods compared, metrics, results and findings are described in more details.

3.4 Inferring General Search Intent

Four variants of the general intent detection problem were considered:

- **Problem 1:** Classify a search into the three general intent categories [23, 119], namely, navigational, informational and transactional.
- **Problem 2:** Same as Problem 1, but do not distinguish between transactional and navigational intents. As we will see, the interaction pattern of a transactional search actually is similar to a navigational search, and there is often ambiguity between the two goals even for a human annotator. In this setting, all transactional queries are labeled as navigational. Note that similar formulations of binary intent classification have also been considered in previous research [94, 21].
- **Problem 3:** Same as Problem 2, but consider *re-finding* [133] searches (i.e., those searches where the user uses the query as a “bookmark” to return to a previously visited website) as navigational.
- **Problem 4:** Same as Problem 3, but identify and ignore the *abandoned* [96] searches (i.e., queries with none of the results clicked) as the interaction patterns differ substantially from searches followed by a click.

3.4.1 Data Collection

The data was gathered from mid-January 2008 until mid-March 2008 from the publicly-used machines in the Emory University libraries. The dataset statistics are reported in Table 3.1. The population was primarily undergraduate college students, graduate students, staffs and faculty, who agreed to opt-in for this study. The identity of the participants is unknown.

This study focused on only *initial queries* that is, avoiding follow-up queries in the same search session. The dataset statistics are summarized in Table 3.1, consisting of around 1500 initial searches, with their search engine result pages, clicked

URLs, and corresponding mouse move trajectories. From this set, 300 searches were randomly sampled (without replacement, only including the first search of each query) into the final sample. The number was chosen so that the sample is large enough to be interesting, and small enough to allow careful human labeling of the “correct” classification of the intent, according to the tasks, defined next.

<i>Statistic</i>	<i>Total</i>
Number of users	860
Number of search sessions	1,597
Number of queries	3,214
Average trajectory length (px)	1,068
Average vertical range (px)	324
Average horizontal range (px)	537

Table 3.1: Dataset statistics

The ground truth labels for general intent detection were generated through human annotators. To manually classify search intent, the annotators were presented with the search query, the clicked URLs, and the SERP snapshots overlaid with the corresponding mouse trajectories. Using these clues, the annotators then labeled the search intent into one of the classes, also marking searches that had ambiguous intent, when it was unclear whether a search was navigational or informational, for example.

The labeled dataset statistics are reported in Table 3.2. Note that 14% of the searches in the sample were ambiguous, and an additional 3% of the searches missed the corresponding search engine result pages. These 17% of the searches were excluded from the final dataset as there was no reasonable way of recovering the corresponding “true” search intent.

3.4.2 Metrics

Having obtained a set of manually labeled search intents as described above, the prediction accuracy of the various methods can be compared. In particular, standard

<i>Label</i>	<i>Count</i>	<i>Percentage</i>
Navigational	89	29.67%
Informational	147	49.00%
Transactional	13	4.33%
Error	9	3.00%
Ambiguous	42	14.00%

Table 3.2: Distribution of labeled general intents in the 300 search sample

information retrieval and classification metrics were used:

- **Accuracy:** The fraction of all the searches that were correctly assigned the intent label (compared to the manually assigned “true” label).
- **F1:** Macro-averaged F1 measure computed for each class, averaged across all classes. This complementary metric can help capture the difference in performance for skewed class distributions (where Accuracy might be misleading). The F1 measure for each class is computed as $2 \cdot PR / (P + R)$ where P is precision (i.e., fraction of predicted class instances that are correct) and R is recall (fraction of all true class instances correctly identified).

These two metrics give a complete picture of overall performance as well as performance for each intent class.

3.4.3 Methods Compared

The main methods used for general intent prediction in this section are summarized as below. C4.5 decision tree algorithm was used for the baseline as it performed the best for the query and click-through features and support vector machines were used to train the remaining classifiers as it achieved the best performance when interaction features were incorporated. The feature groups and the corresponding features are summarized in Table 3.3.

- **Query+Click:** C4.5 decision trees classifier trained using query and click-through features only, which can be obtained from the traditional server-side search logs. This method represents the state-of-the-art baseline [94].
- **Interactions (Simple):** Support vector machines trained using *Interactions (Global)* features only.
- **Interactions (Full):** Support vector machines trained using both the *Interactions (Global)* and *Interactions (Local)* features.
- **All:** Support vector machines trained on the combination of both the full client-side interaction features as well as the server-side query and click-through features, thereby using all the available information.

<i>Feature group</i>	<i>Count</i>	<i>Description</i>
<i>Query</i>	2	QueryLengthChars, QueryLengthWords
<i>Click</i>	3	TopFraction, DeliberationTime, IsSubstring
<i>Interaction (Global)</i>	3	TrajectoryLength, VerticalRange, HorizontalRange
<i>Interaction (Local)</i>	19	AvgSpeed*(5), AvgAcceleration*(5), Slope*(5), RotationAngle*(4)
<i>All</i>	27	All features and feature classes used for experiments

Table 3.3: Summary of the features used for general intent detection

3.4.4 Results and Discussion

Now, the experimental results for general intent detection are reported. First consider *Problem 1*, which aims to classify the three classic intent classes, with corresponding results, produced using 4-fold cross-validation, summarized in Table 3.4. As we can see, the naïve representation of interaction *Interactions (Simple)* consistently outperforms *Query+Click*, for a modest gain on both of the accuracy and F1 metrics. Furthermore, the enhanced representation *Interactions (Full)* substantially outperforms *Interactions(Simple)*, indicating the benefit of modeling the detailed

properties of pre-click mouse trajectory on SERPs. Finally, the integrated full system *All*, that combines query, click, and fine-grained interaction features, has the highest accuracy and F1 measure of all systems. Interestingly, the improvement of *All* over *Interactions (Full)* is not large, suggesting that the most benefit comes from the search behavior in the session, and not from the information aggregated across all users issuing the same query (e.g., click-through distribution) – allowing *All* to have higher accuracy than the *Query+Click* baseline by as much as 17%. Among the three intent classes, the informational class was the most easy to be detected, followed by the navigational class and transactional class, supporting the hypothesis that the interaction patterns of the navigational and transactional intents are similar. Next, the results for *Problem 2* are discussed, which provide further insights about this hypothesis.

<i>Method</i>	<i>Accuracy (%)</i>	<i>F1</i>			
		<i>Nav</i>	<i>Info</i>	<i>Trans</i>	<i>Macro Average</i>
Query+Click	65.46	46.20	76.60	0	40.93
Interactions (Simple)	67.70 (+3%)	57.90	76.3	0	44.73(+9%)
Interactions (Full)	75.50 (+15%)	69.00	82.40	0	50.47 (+23%)
All	76.31 (+17%)	71.30	83.10	0	51.47(+26%)

Table 3.4: Accuracy and F1 for different methods (Problem 1)

Table 3.5 reports the accuracy of the different methods if transactional searches are re-labeled as navigational. As we can see, the accuracy of all the methods increases for this problem, which further supports the hypothesis that navigational and informational intent exhibits similar behavior patterns – combining the two resulting in more accurate classification. The gain of the full fine-grained interaction model (*Interactions (Full)*) and the full model (*All*) over the query and click baseline or the simple interaction classifier remains consistent and substantial.

As reported in Table 3.2, 14% of the searches in the sample were ambiguous. Further analysis and re-labeling (Table 3.6) revealed that 27 of the ambiguous searches are likely to be with *re-finding* intent [133] – that is, searches that appear infor-

<i>Method</i>	<i>Accuracy (%)</i>	<i>F1</i>		
		<i>Nav</i>	<i>Info</i>	<i>Macro Average</i>
Query+Click	67.87	49.40	76.50	62.95
Interactions (Simple)	70.28 (+4%)	68.60	71.80	70.20 (+12%)
Interactions (Full)	78.71 (+16%)	72.30	82.70	77.50 (+23%)
All	79.92 (+18%)	76.60	82.40	79.50 (+26%)

Table 3.5: Accuracy and F1 for different methods (Problem 2)

mational based on the text of the query, but are really “bookmarks” to re-retrieve previously found website. The guess was based on the observation that in these cases the users went directly to click on a search result, which is very similar to the user behavior of typical navigational searches. This hypothesis was further verified by confirming that a portion of the clicked URLs were previously visited (not all of the URL revisits could be verified due to the limited timespan of the collected logs). Although according to the query text, a search might look like informational, such as the behavior with query “rpi rankings” and query “emory financial aid” as illustrated in Figure 3.8, the user intent might actually be navigational since she had visited the page or she was aware of such a page.

<i>Label</i>	<i>Number</i>	<i>Percentage</i>
Re-finding/Navigational	27	9.00%
Ambiguous (Unknown)	15	5.00%
Abandonment	85	28.33%

Table 3.6: Distribution of the re-labeled ambiguous searches

If the re-finding searches are considered as navigational in intent (referred to as *Problem 3*), the behavior of the classifiers changes drastically, as reported in Table 3.7.

As we can see, the fine-grained interaction model *Interactions (Full)* substantially outperforms the combined method *Interactions (All)*. This result illustrates that when search intent is indeed *personalized* – that is, for the current user session,

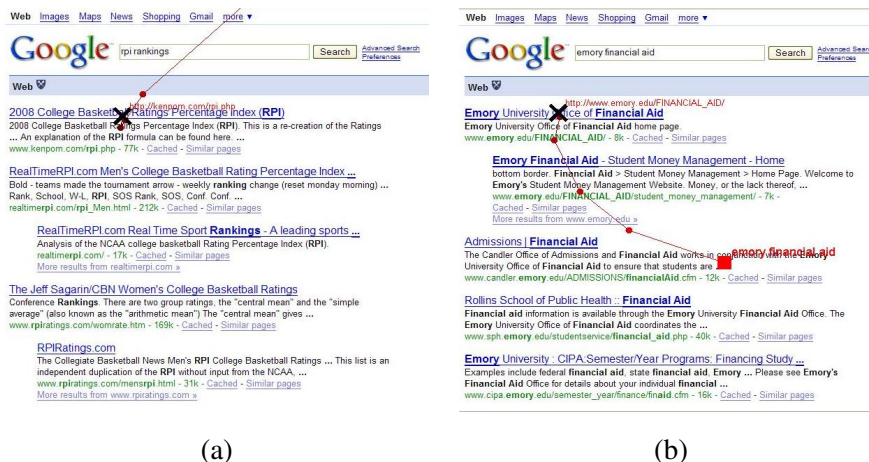


Figure 3.8: Mouse trajectories for searches with navigational/re-finding intent: query “rpi rankings” (a) and query “emory financial aid” (b)

<i>Method</i>	<i>Accuracy (%)</i>	<i>F1</i>		
		<i>Nav</i>	<i>Info</i>	<i>Macro Average</i>
Query+Click	64.49	49.00	72.80	60.90
Interactions (Simple)	71.38(+11%)	72.70	70.00	71.35(+17%)
Interactions (Full)	79.71(+24%)	78.50	80.80	79.65(+31%)
All	77.53(+20%)	77.40	77.70	77.55(+27%)

Table 3.7: Accuracy and F1 for different methods (Problem 3)

the normally informational query may actually be navigational – then the classifier session-level interaction is more accurate, and incorporating the “majority” intent in fact degrades performance. Interestingly, in the analysis, also found was that for easier informational RPI queries users may exhibit similar trajectory patterns as navigational searches. To address this problem, user history may be utilized to help teasing out the “re-finding” behavior, which may result in further improvement in prediction. In contrast, a navigational search may exhibit patterns resemble typical informational searches when the search engine fails to return the target website. Nevertheless, the examination patterns for navigational searches still tend to differ from the informational searches – with a navigational intent, the examination is

more likely a glance at the title, which may result in faster mouse movement or scrolling behavior, while for an informational intent, the examination is more likely to be scrutinizing on the snippets, which tends to result in slower mouse movements or scrolling.

Another major source of ambiguity comes from the *abandoned* searches (i.e., those with no click on any result). As shown in Figure 3.9, the mouse cursor trajectories appear similar as a navigational search, which introduces noise into the classifier. For example, if a user misspells the query, or none of the results appear relevant, the user may immediately click the “did you mean” feature, fine the query, or even give up the search task. Conversely, the user may have gotten the needed information from the result summaries, which is referred to as “good abandonment” in the literature [96]. To address these issues, including features of the subsequently page visit in the same task session could be useful. Nevertheless, as we have seen, even without these additional features, the fine-grained interaction features already result in an effective search intent classifier.

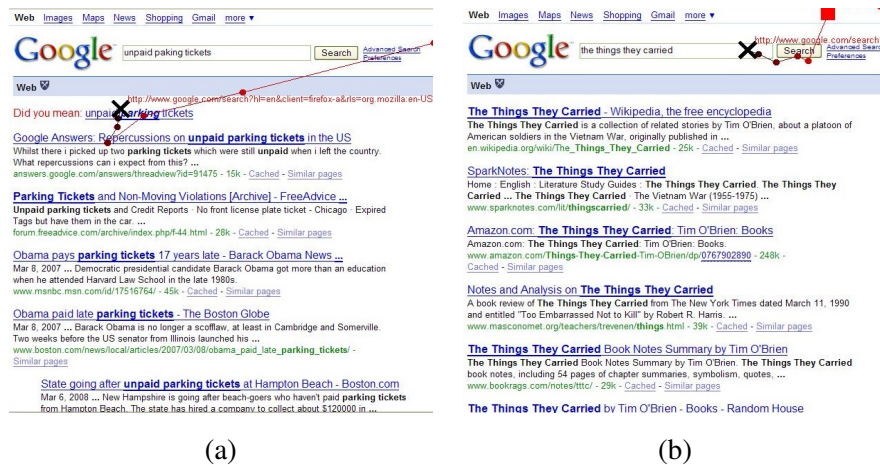


Figure 3.9: Mouse trajectories for abandoned searches: query “unpaid paking ticket” followed by a click on “did you mean” (a) and query “the things they carried” followed by a reformulation (b)

For the sake of exploration, one may consider simply discard the abandoned

searches (referred to as *Problem 4*). The results for this problem are reported in Table 3.8.

<i>Method</i>	<i>Accuracy (%)</i>	<i>F1</i>		
		<i>Nav</i>	<i>Info</i>	<i>Macro Average</i>
Query+Click	68.21	76.20	52.30	64.25
Interactions (Simple)	76.41(+12%)	70.10	72.80	75.30(+17%)
Interactions (Full)	83.59(+23%)	86.70	78.70	82.70(+29%)
All	82.56(+21%)	86.00	77.00	81.50(+27%)

Table 3.8: Accuracy and F1 for different methods (Problem 4)

As we can see, the Accuracy and F1 of all the methods further increase substantially as compared to Problem 3, suggesting much additional room for improvements if the abandoned searches can be handled appropriately. Interestingly, the *Interactions (Full)* classifier is still the most accurate, indicating that when re-finding queries are treated according to the *individual* user intent and when the abandoned queries are not considered, information about behavior of other users is not helpful for individual search intent identification.

Feature Contributions:

To better understand the contribution of the different features, the information gain of each feature (computed for Problem 2) are reported in Table 3.9. As we can see, the most important features represent different aspects of the mouse trajectories (e.g., speed, acceleration, rotation) but also include query length and deliberation time – the more traditional user modeling features.

<i>Information Gain</i>	<i>Feature</i>
0.2043	AvgAcceleration (segment 3)
0.197	AvgAcceleration (segment 2)
0.1705	AvgSpeed (segment 3)
0.1509	AvgSpeed (segment 4)
0.1451	VerticalRange
0.1449	AvgAcceleration (segment 4)
0.1425	AvgAcceleration (segment 1)
0.1275	TrajectoryLength
0.1146	TopFraction
0.1125	RotationAngle (segment 0)
0.0922	AvgSpeed (segment 2)
0.0843	QueryLengthWords
0.0781	IsSubstring
0.075	AvgAcceleration (segment 0)
0.0708	DeliberationTime

Table 3.9: Most important features for general intent detection (ranked by Information Gain)

3.5 Inferring Commercial Search Intent

The problem is to detect, given a user’s behavior on a SERP, whether the query had *research* or *purchase* intent.

3.5.1 Data Collection

A user study was performed with 10 subjects, who were graduate and undergraduate students and university staff, that is, were technically savvy and had some experience with Web search. The subjects were asked to perform two search sessions. Each subject was asked first to research a product of interest to them for potential future purchase. Then, the same subject was asked to attempt to “buy” an item of

immediate interest to the subject, which may or may not be the same item the subject was researching in the previous stage. The subjects were not restricted on time, and could submit any queries (usually, to the Google search engine) and click on any results.

All the interactions were also tracked using the proposed EMU Firefox plugin (Section 3.4). At the same time, the searcher gaze position was tracked using the EyeTech TM3 integrated eye tracker at approximately 30Hz sampling rate, for subsequent analysis. Additionally, each search and corresponding SERP interactions were labeled as parts of a *research* or *purchase* session, according to the explicitly stated intent of the corresponding session.

3.5.2 Methods Compared

Support vector machines (SVM) were used to train the commercial intent detector. The major feature groups and representative features are summarized in Table 3.10.

- **Baseline:** always guesses the majority class (Research).
- **SVM (Query):** similar to the state-of-the-art models using query features (e.g., [111]), implemented using Query group features described in Section 3.3.2, and trained using the SVM model.
- **SVM (All):** the SVM classifier implemented using the features described in Section 3.3.2 to infer the user goal for each search (independently of other searches in the session).

3.5.3 Results and Discussion

In this experiment, the intent of each search is predicted independently of other searches in the session. The data was split by time, using the first 90% of searches for each subject’s data for training, and the rest for testing (recall, that each subject had two sessions, one research, and one purchase). To evaluate classification

<i>Feature group</i>	<i>Count</i>	<i>Description</i>
<i>Query</i>	4	QueryTokens* (unigram), QueryLengthChars, QueryLengthWord, IncludesTLD (1 if contains “.com”, “.edu”).
<i>Click</i>	7	ClickUrl* (unigram), NumBrowseAfterClick, AverageDwellTime, TotalDwellTime, SAT, DSAT, ClickType
<i>Interaction</i>	99	MouseRange, MouseCoordinates, MouseSpeed, MouseAcceleration, TotalMouse, TotalScroll, TotalKeypress, SERPDwellTime, DeliberationTime, HoverEastAd, HoverNorthAd, HoverOrganic, etc (see main text)
<i>SERP Content</i>	3	AdText* (unigram), OrganicText* (unigram), SERPText* (unigram). Each feature contains 100 most frequent terms from each area of the SERP (e.g., 100 most frequent tokens in the ads).
<i>Result Quality</i>	7	TotalAds, NorthAds, EastAds, SnippetOverlap, SnippetOverlapNorm, AdOverlap, AdOverlapNorm
<i>Context</i>	7	IsInitialQ, IsSameQ, IsReformulatedQ, IsExpansionQ, IsContractedQ, RepeatQ, SERPIndex
<i>All</i>	127	All features and feature classes used for experiments

Table 3.10: Summary of the features used for representing searcher context and interactions in inferring search intent

performance, the standard Precision, Recall and Macro-averaged F1 were used. Table 3.11 shows that the proposed system, SVM (All), outperforms both baselines, resulting in accuracy of almost 97%.

To identify the most important features contributing to the classification, feature ablation was performed by removing one feature group at a time from the classifier (Table 3.12). All the feature groups provide significant contributions, but the most important features appear to be SERPContent and Interaction features: with these features removed, accuracy degrades to 86.7% from 96.7% with these features included. This makes sense since the SERP content can help enrich the context of a query, while the Interaction features provide additional clues about the searcher interest. However, since this user study was done over a rather small number of subjects, further investigation and additional user study is needed to fully understand the connection between various feature groups. To complement these results, the proposed model was validated on an objective ad click-through metric on a much

larger user population, as described next.

<i>Method</i>	<i>Acc.</i>	<i>Research</i>		<i>Purchase</i>		<i>F1</i>
		<i>Prec.</i>	<i>Recall</i>	<i>Prec.</i>	<i>Recall</i>	
Baseline	56.7	56.7	100	0	0	36.2
SVM (Query)	86.7	93.3	82.4	80.0	92.3	86.6
SVM (All)	96.7	100	94.1	92.9	100	96.6

Table 3.11: Classification performance for research vs. purchase.

<i>Method</i>	<i>Acc.</i>	<i>Research</i>		<i>Purchase</i>		<i>F1</i>
		<i>Prec.</i>	<i>Rec.</i>	<i>Prec.</i>	<i>Rec.</i>	
SVM (All)	96.7	100	94.1	92.9	100	96.6
SVM (-Query)	93.3	94.1	94.1	92.3	92.3	93.2
SVM (-Click)	90.0	93.8	88.2	85.7	92.3	89.9
SVM (-Interaction)	86.7	100	76.5	76.5	100	86.7
SVM (-SERPContent)	86.7	93.3	82.4	80.0	92.3	86.6
SVM (-ResultQuality)	93.3	100	88.2	86.7	100	93.3
SVM (-Context)	93.3	100	88.2	86.7	100	93.3

Table 3.12: Feature ablation results for intent classification.

3.5.4 Ad Click-through on Real Search Data

To better understand its effectiveness, the proposed model was evaluated on a large dataset of real user searches collected in the Emory University libraries using the infrastructure described earlier. The hypothesis is that for *research* searches, click-through on search ads should be lower than for *purchase* searches. Therefore, the effectiveness of the intent classification model can be evaluated by comparing the ad click-through on the searches classified as *research* by the model, to those classified as *purchase*. To avoid “cheating”, no click group or result URL features were used, as they could provide information to the classifier about the ad click on the SERP.

The data was gathered from mid-August through mid-December 2008. To ensure data consistency, a longitudinal dataset was generated, consisting of the usage for 440 opted-in users, who clicked a search ad at least once during this period. For this universe of users all the search sessions attempted during this period were included. The resulting dataset contains 4,377 login sessions, comprising 6,476 search sessions, 16,693 search tasks and 45,212 searches.

The predicted *purchase* searches have substantially higher ad click-through rates (9.7%) compared to *research* searches (4.1%), and all searches with at least one ad displayed (5.9%). These statistics are summarized in Table 3.13. As hypothesized, the *research* and *purchase* predictions indeed correlate with ad click-through of real users. What makes this result remarkable is that the proposed model was trained on a small dataset compiled from just 10 subjects in the user study (with clear intent labels), yet still provides promising performance on unconstrained user data obtained “in the wild”.

<i>Search class</i>	<i>#ACLK (%)</i>	<i>#SERP with Ads</i>	<i>Ad CTR (%)</i>
All	854	14545	5.9
Research	417	10054	4.1 (-29%)
Purchase	437	4491	9.7 (+66%)

Table 3.13: Search ad click-through statistics on all search pages (All), and for searches classified as “Research” and “Purchase”.

3.6 Inferring Advertising Receptiveness

This problem of predicting future ad click-through for the *current user* is distinct from predicting ad click-through in aggregate for many users. The problem is defined as follows: *Given* the first i searches in a search task $S(s_1, \dots, s_i, \dots, s_m)$, and the searcher behavior on these first i SERPs, *predict* whether the searcher will click on an ad on the SERP within the current search task S , for any of the future searches $s_{i+1}, s_{i+2}, \dots, s_m$.

3.6.1 Methods Compared

Conditional Random Fields (CRF) were used to train the future ad click-through predictor. Same as training the commercial intent detector, the major feature groups and representative features are summarized in Table 3.10.

As a realization of the problem statement, the CRF was configured to have two hidden states, $A+$ and $A-$, corresponding to “Receptive” (meaning that an ad click is expected in a future search within the current session), and “Not receptive” (meaning to not expect any future ad clicks within the current session). Figure 3.10 illustrates this configuration, as we can see the first two searches are labeled as $A+$ as there is an future ad click (during the third search) and the third and fourth searches are labeled as $A-$ as there is no future ad click in the remainder of the task session.

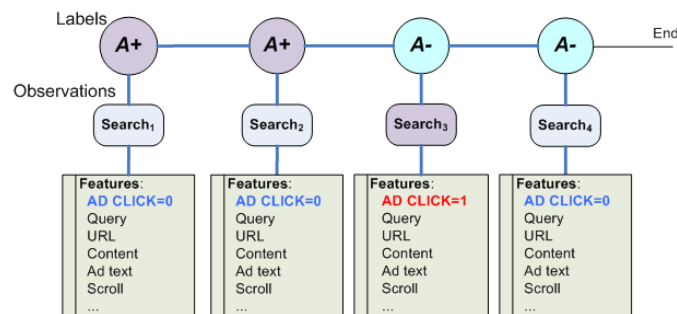


Figure 3.10: CRF model configuration with two hidden states, $A+$ (receptive), and $A-$ (non-receptive), with labels assigned according to the observed *future* ad click-through - here on the third search result page within the session.

The methods compared are summarized as follows:

- **CRF (Query)**: CRF model, implemented using the Query group features as described in Section 3.3.2, which represents the state-of-the-art methods that use the query signals, such as [111].
- **CRF (Query+Click)**: CRF model, implemented using Query group and Click group features as described in Section 3.3.2, which represents the state-of-the-art models that use the query and click signals, such as [30].

- **CRF (All)**: CRF model, implemented using all features as described in Section 3.3.2, which include fine-grained interaction, SERP content, result quality and context features in addition to the query and click features.
- **CRF (All-Interaction)**: Same as above, but with the Interaction group features removed, which is to gauge the additional evidence lies in the interaction feature group when other feature groups presented.

3.6.2 Data and Evaluation Metrics

For this problem, the dataset was based on the interaction data collected from the opted-in users in the Emory Libraries, and consists of the same log data as described in Section 3.5.4.

Evaluation Metrics: To focus on the ad click prediction, the results are reported for the positive class, i.e., the “advertising-receptive” state. Specifically, Precision (P), Recall (R), and F1-measure (F1) are reported and calculated as follows:

- **Precision (P)**: Precision is computed with respect to the positive (receptive) class, as fraction of true positives over all predicted positives. Specifically, for each search task, the precision is the fraction of correct positive predictions over all positive predictions for the task, averaged across all the search tasks.
- **Recall (R)**: for each task, the recall is computed as the fraction of correct positive predictions over all positive labels in the task. This value is then averaged over all the tasks.
- **F1-measure (F1)**: F1 measure, computed as $\frac{2P \cdot R}{P + R}$.

3.6.3 Results and Discussion

To simulate an operational environment, the data was split by time, and the first 80% of the sessions was used for training the system, while the remaining 20%

of the sessions was used for test. The results on the test set are reported in Table 3.14. As we can see, the proposed system achieves the highest performance on all metrics, compared to the baselines. Specifically, CRF (Query+Click) outperforms CRF (Query) on the ad receptiveness prediction task by incorporating the click information, and the CRF (All) system further increases both precision and recall by incorporating additional behavior features. Interestingly, removing the Interaction group of features from the full system (CRF (All-Interaction)) degrades the recall and overall F1 performance of the system, while precision is somewhat improved. This suggests that interaction features help detect additional cases (compared to query and click information alone) where a searcher may be interested in the ads, while occasionally introducing additional false positives. Discuss of the performance of the system in more detail are provided, next.

<i>Method</i>	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
CRF (Query)	0.05 (-)	0.11 (-)	0.07 (-)
CRF (Query+Click)	0.14 (+153%)	0.12 (+11%)	0.13 (+77%)
CRF (All)	0.15 (+170%)	0.21 (+99%)	0.17 (+141%)
CRF (All-Interaction)	0.16 (+206%)	0.14 (+32%)	0.15 (+112%)

Table 3.14: Precision, Recall, and F1 for predicting ad receptiveness within a search task

Potential limitations: While the ad click prediction experiments were performed over a relatively large dataset collected over thousands of real search sessions for hundreds of users, some limitations exist for the study. Specifically, the user population is relatively homogeneous (college and graduate students, and faculty and staff), and substantially more training data may be required to achieve this performance for the general population. Another limitation is lack of conversion data: ad click-through is just one evaluation metric, and may not be predictive of the ultimate intent of the searcher (e.g., a searcher may click on an ad out of curiosity). Despite the limitations above, the population is large enough that useful conclusions could be drawn. To better understand the system performance and guide follow-up research, the representative case studies are presented next to provide better under-

standing of the system performance:

Not using a mouse as a reading aid: this is the most frequent source of error introduced by the interaction features: when mouse is not used to mark or focus user interest, interaction information could be misleading. One example is given in Figure 3.11, where the mouse cursor keeps still around the search box (indicated by the black cross, the red square and dots) while the searcher scans various places on the search engine result page (SERP), which is indicated by the widely distributed gaze heatmap. One possible approach is to classify users into different groups according to their mouse usage patterns, and train separate prediction models for each group.

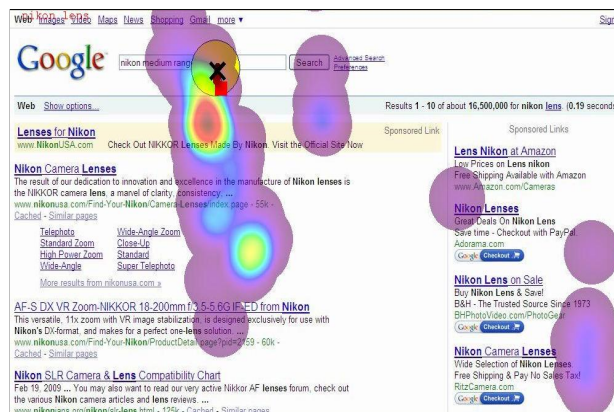


Figure 3.11: Mouse trajectory and gaze heatmap on a search engine result page when mouse is not used to mark or focus user interest

Long difficult research sessions with ad clicks: in such cases, the searcher began with a research intent, and in her first several searches, no interest in ads were shown. However, as the session progresses, the user eventually clicks on an ad as the promising organic results are exhausted. For example, one searcher submitted a query “comcast basic cable channels” in her task to find Comcast’s basic cable line-up, and finally clicked on an ad because of the unsatisfactory organic results. Such ad clicks appear to be different from cases where a user clicks on ads because of a premeditated purchasing intent.

Commercial purchase sessions without ad clicks: in such cases, a searcher ex-

amined the ads but did not click on any. This could be due to poor quality of search ads, or to availability of more promising organic search results. For example, one searcher submitted a query “george boots” and clicked on a Google’s Product Search result. In this case, the searcher might be actually receptive to search advertising. However, such sessions were labeled as “non-receptive” since there’s no future ad click to use as evidence. One natural extension of the model is to expand the labels by considering clicks on product search results to be similar to ad clicks with respect to purchasing intent. Another possibility may be that particular users could be generally “less-receptive” to advertising. To tackle this problem, personalizing the user models is a promising direction for future work.

3.7 Summary

This chapter introduced techniques that captured not only the queries and clicks, but also the fine-grained interactions with the search results, contextualized within a search session. The experimental results on detecting two dimensions of Web search intent demonstrated the generalizability and flexibility of the proposed approach. This thread of research began with detecting general search intent such as predicting *navigational* vs. *informational* goal and demonstrated the benefits of using client-side fine-grained interactions in more accurately determining user needs in a search session. Then the prediction of *research* vs. *purchase* goal of the searcher was studied, and insights were obtained about the feature groups most important for distinguishing these variants of commercial intent from a controlled user study. Following up on the prediction in the commercial dimension, the model was validated by showing correlation of the predicted search intents with the ad click-through rates of real users. Finally, the proposed intent detection model was extended to address an important practical application of predicting *future* search ad click-through within the *current* search session of each user, demonstrating the effectiveness of the proposed model.

Chapter 4

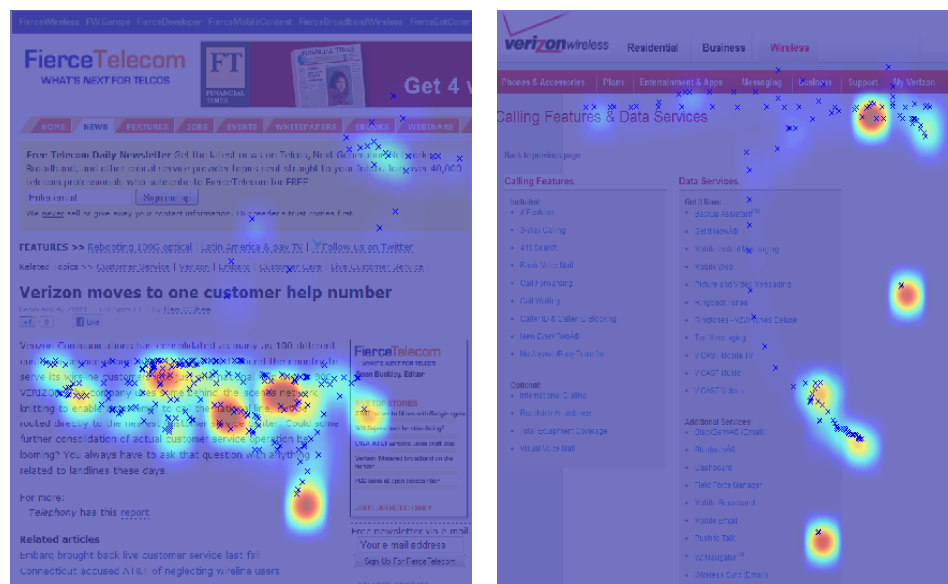
Estimating Document Relevance

In Chapter 3, the techniques for inferring search intent based on pre-click behavior are introduced. These techniques allow for more accurate retrieval of unseen documents through enriching and contextualizing the search query. However, if a document has been visited (i.e., its associated interaction data is available), direct estimation of its relevance may be more preferable, which allows directly re-ranking the search results. In this chapter, *post-click search behavior* is studied in depth for estimating the “intrinsic” page relevance for a search task. The bulk of this chapter has been published as [60].

To directly estimate document relevance, previous research has made great use of result click-through data (e.g., [3, 83, 48, 47]). However, the usefulness of click-through statistics is limited by a number of *presentation biases*, which strongly influence user click behavior. One of the most significant limitations of click-through data, is that clicks are based primarily on a document’s *perceived* relevance [47], where a searcher guesses the page’s relevance based on a short summary generated by the search engine. However, the “perceived” relevance may be inconsistent with the actual “intrinsic” relevance [47], where a searcher clicks on a result only to find out that it is not relevant. To address this problem, page *dwelling time* (the time spent examining the result document) has been proposed as a measure of intrinsic document relevance [106, 38, 51, 88, 144, 28]. The main intuition is that “short” dwelling time (typically, considered to be less than 30 seconds), indicates that a document is non-relevant. The most heavily studied scenario is that of a “bounce back”, which happens when the searcher returned to the Search Engine Result Page

(SERP) shortly after she clicked on a result, indicating low result relevance [120]. This heuristic and resulting metrics have been successfully adapted by the major search engines, and have undoubtedly improved search quality by detecting non-relevant or even detrimental results.

Unfortunately, the converse of the short dwell time rule is not true: a “long” page dwell time does not necessarily imply result relevance. In fact, a most frustrating scenario is when a searcher spends a long time searching for relevant information on a seemingly promising page that she clicked, but fails to find the needed information. Such a document is clearly non-relevant (and arguably one of the most detrimental to the searcher experience). Yet, based on dwell time alone, this document would be considered highly relevant, and remain high in the search ranking to frustrate future searchers.



(a) relevant (dwell time: 30s)

(b) non-relevant (dwell time: 30s)

Figure 4.1: Cursor-based “Reading” examination heatmap of a relevant document (a) compared to “Scanning” of a non-relevant document (b), both with equal dwell time (30 seconds).

To address this problem, the proposed approach is to use a rich set of *post-click searcher behavior* for more precisely analyzing how the searchers spend their time on the landing pages and the subsequently viewed documents, which would in turn allow for more accurate estimation of intrinsic document relevance. As an illustration, Figures 4.1(a-b) show the searchers' cursor movement on clicked result pages for the task of finding the phone number of the Verizon Wireless helpline for Massachusetts, where the user spends approximately 30 seconds examining each of the pages (i.e., both pages have almost equal dwell time). The color intensity in the figures indicates the amount of time the mouse cursor spent over the corresponding document regions, with the exact cursor coordinates indicated by the small crosses. The differences in the examination of a relevant page (Figure 4.1(a)) and a non-relevant page (Figure 4.1(b)) are striking. For the former, the searcher was carefully "reading" the text and using the mouse as a reading aid (examination of the page reveals that the answer of the search task indeed lies in the highlighted paragraph), while for the latter, the searcher appears to be "skimming" or "scanning" the page, without finding relevant information worth careful reading (indeed, the answer was not on the page). This example illustrates the underlying hypothesis: that page dwell time alone is not sufficient to distinguish between relevant and non-relevant pages, but post-click searcher behavior can provide the necessary additional evidence to distinguish the two.

Specifically, the hypothesis is that searcher interactions on landing pages such as cursor movements and scrolling can help more accurately interpret searcher viewing behavior, in turn, improve relevance estimation. That is, like eye movements, such interactions can reflect searcher attention. These interactions can be captured with Javascript code that is embedded in a browser Add-on (e.g., a search engine toolbar). This would allow estimating whether some parts of the landing page captured the searcher's attention and provide additional clues about the document relevance.

To test this hypothesis, the patterns of examination and interaction behavior are identified, that correspond to viewing a relevant or non-relevant document (Section 4.1), followed by developing a novel model of inferring document relevance that in-

corporates rich Post-Click Behavior (PCB) such as cursor movements and scrolling that could capture these patterns (Section 4.2). Similar to intent inference models, the relevance prediction model is then operationalized by converting these interactions into features, which can then be used as input to machine learning algorithms for tasks such as estimating personalized and aggregate document relevance, and improving result ranking (Section 4.3).

In summary, the contributions in this chapter include:

- Characterizing patterns of examination and interaction behavior that correspond to viewing a relevant or non-relevant document (Section 4.1).
- PCB, a novel model of relevance estimation that captures post-click behavior (Section 4.3).
- Empirical evidence that PCB is more effective than using dwell time information alone, both for estimating the explicit judgments of each user, as well as for ranking the documents using the estimated relevance (Section 7).

4.1 Landing Page Examination

This section describes the patterns of landing page examination and interaction that were identified. Overall, two basic patterns of viewing were observed, namely, “reading” and “scanning” (as illustrated in Figure 4.1). “Reading” tends to occur when relevant information (or seemingly relevant information) is found, and the searcher is consuming (or further verifying) the information. In contrast, “scanning” typically indicates that the searcher has not yet found the relevant information and is still in the process of searching. Typically, the viewing behavior is some mixture of these two basic components. Sometimes, the mixture is dominated by one of the two types. For example, Figure 4.1(a) is dominated by the “reading” behavior (suggested by the cursor heatmap overlaid on top of the answer of the search task [118]) while Figure 4.1(b) is dominated by the “scanning” behavior (suggested

by the more vertically spread-out cursor distribution on the right of the screen on this page that does not contain the relevant information [118]).

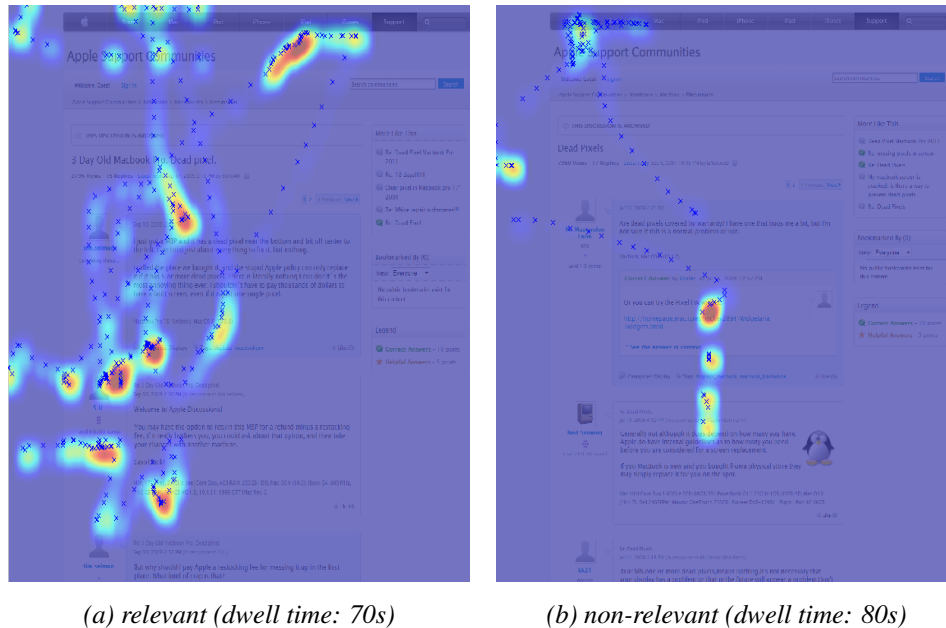


Figure 4.2: An example of “Reading” a relevant long document (a) vs. “Scanning” a non-relevant long document (b).

At other times, the viewing behavior is more complex, especially, when the relevance of the document is not obvious (e.g., the document is long and contains a mix of relevant and irrelevant information). Figure 4.2(a) shows an example of viewing behavior on a long relevant landing page, while Figure 4.2(b) shows an example of viewing behavior on a irrelevant long page. The search task for both of the pages were “How many pixels must be dead on a MacBook before Apple will replace the laptop? Assume the laptop is still under warranty.” and the dwell time on the two documents were roughly 70 seconds and 80 seconds, respectively. The two documents in this example are both from Apple’s support forum and are much longer than the example documents in Figure 4.1. In such a case, using dwell time alone would suggest that the two are both relevant, and moreover the second document is slightly more relevant. However, the document examination patterns suggest that

the two are quite different. The cursor movements on the first document are more focused on the left side, with clustering around the top posts, which suggests “reading” behavior (indeed, closer examination shows that the top posts contain relevant information). However, the pattern indeed seems more complex than what we have seen in Figure 4.1(a) – the cursor positions are more spread-out vertically and no extensive horizontal cursor movement was observed. In contrast, on the non-relevant document in Figure 4.2(b), the searcher keeps the mouse still and scrolls – which indicates “scanning” behavior. Interestingly, here too the cursor positions are clustered on the left (indicating slowing down of cursor movements) over the top post, which may indicate “reading” behavior. Examination reveals that the page indeed contains on-topic information, that initially seems relevant, but does not contain the needed answer. Thus, in this example, the initial “reading” behavior is followed by a series of “scanning” before the searcher exits the page without finding her answer.

As these examples indicate, in addition to the variety in combining the two basic viewing patterns, the corresponding behavioral signals of these two patterns vary too. For example, when reading, searchers might keep the cursor still or use the cursor to mark the text (eye and cursor are coordinated only vertically) without actively moving the cursor horizontally (eye and cursor are coordinated both vertically and horizontally). Nevertheless, in both cases, the searcher tend to slow down the cursor movements or scrolling, especially, in the vertical direction. As for the “scanning” behavior, in contrast, the searcher tend to move the cursor and/or scroll faster as they are searching for the relevant information or sometimes also keeps the mouse still.

In summary, after examining many viewing sessions, common patterns across all the post-click page examinations that correlate with document relevance were observed and listed below:

- *Periods of horizontal reading indicate relevance:* The searchers are more likely to slow down and move mouse horizontally to read when the document is relevant, as opposed to only quickly scanning the document when it

is non-relevant.

- *Focused attention indicates relevance*: searchers tend to focus on only one or a small number of areas for a relevant document, while distribute time more evenly throughout an non-relevant document. In contrast to the previous case, if searchers exhibit “reading” behavior (e.g., slowing down) *multiple times*, it is more likely that she still did not find the right information that satisfies her need – for more complex task, or documents with denser text, such “reading” behavior are likely to be triggered.
- *Left-prevalence*: On relevant pages, searchers tend to keep the cursor towards the left half of the screen, where typically most of the content laid out on a Web page, to help reading or prepare to click on a link for more details.
- *“Scanning” followed by “reading” indicates relevance*: Often, a “scanning” behavior followed by focused, careful “reading” behavior at the end of the examination indicates relevance, while “reading” behavior in the beginning followed by “scanning”(i.e., the searcher is still not yet satisfied with what he or she has found so far), indicates non-relevance of the document.
- *“Skipping” indicates non-relevance*: Periods of reading or scanning, interspersed with periods of quick scrolling (“skipping” document sections) indicates lower relevance than continuous examination – searchers may become impatient, and accelerate “scanning” to an even faster pace.

These patterns can be captured by post-click behavioral signals such as sequences of cursor and scroll speeds and ranges. In the next section, the features designed to model these examination patterns are described, which can subsequently be used to better estimate document relevance.

4.2 Post-Click Behavior (PCB) Features

This section describes the proposed *Post-Click Behavior (PCB)* features to capture the the page examination patterns that could indicate a difference in document relevance. In addition, also included are dwell time, task-level information (which is also shown to be useful in estimating document relevance in recent studies [69]), and the original search engine result ranking, as features in the PCB model. The full list of PCB features and their brief descriptions are reported in Table 4.1, and expanded below.

4.2.1 Dwell Time

Dwell time, or document viewing time, has been previously used as the basic indicator of document relevance. As typically done, dwell time is defined as the interval, in seconds, between the time the page is loaded and the time the searcher leaves the page. Dwell time is used both as a baseline to compare against and as a feature in the full PCB model.

4.2.2 Result Rank

The rank of search result is the belief in its relevance that the search engine holds, which is typically obtained by combining hundreds of ranking signals. Presumably, the smaller the rank value (i.e., the higher the document was ranked), the more relevant the document is likely to be. However, if the search engine fails in accurately estimating the document relevance, the rank would become uninformative. For the viewed documents in the search trail that were not ranked in a search engine result page, the rank of the landing page (i.e., the origin of the search trail the document was on) is used ¹.

¹The ranks are set to be 11 for a small portion of the documents whose ranking information is missing or cannot be recovered.

<i>Group (30)</i>	<i>Feature</i>	ρ
Dwell (1)	<i>dwell</i> : time of the page view in seconds	0.167**
Rank (1)	<i>rank</i> : the rank of the document or the rank of the origin (i.e., the landing page) of the search trail that the document is on if its rank is not available	-0.073
Cursor (14)	<i>cursorcnt</i> : num. of cursor movements	0.164**
	<i>cursorfreq</i> : cursorcnt/dwell	-0.082*
	<i>dist</i> : total overall distance the cursor traveled in pixels	-0.137**
	<i>xdist</i> : total distance the cursor traveled horizontally in pixels	0.101**
	<i>ydist</i> : total distance the cursor traveled horizontally in pixels	0.172**
	<i>speed</i> : dist/dwell	0.101**
	<i>xspeed</i> : xdist/dwell	-0.143**
	<i>yspeed</i> : ydist/dwell	-0.124**
	<i>xmin</i> : minimal x coordinate	0.112**
	<i>ymin</i> : minimal y coordinate	0.093*
	<i>xmax</i> : maximal x coordinate	0.067
	<i>ymax</i> : maximal y coordinate	0.243**
	<i>xrange</i> : xmax-xmin	-0.006
	<i>yrange</i> : ymax-ymin	0.172**
Scroll (5)	<i>scrlcnt</i> : num. of vertical scrolls	-0.008
	<i>scrlfreq</i> : scrlcnt/dwell	-0.206**
	<i>scrlldist</i> : total vertical scroll distance	-0.092*
	<i>scrlspeed</i> : scrlldist/dwell	-0.212**
	<i>scrlmax</i> : maximum scroll top	-0.026
AOI (3)	<i>dwell_aoi</i> : total time the cursor spent in the pre-defined Area of Interest (AOI)	0.227**
	<i>cursorcnt_aoi</i> : cursor count in AOI	0.189**
	<i>cursorfreq_aoi</i> : cursorcnt/dwell	-0.195**
Task (6)	<i>avg_dwell</i> : average dwell time of preceding page views in the task	0.081*
	<i>querycnt</i> : number of preceding queries	-0.138**
	<i>serpcnt</i> : number of preceding search engine result page (SERP) views	-0.142**
	<i>clkcnt</i> : number of preceding clicks	-0.171**
	<i>ctr</i> : clkcnt/serpcnt	0.085*
	<i>tasktime</i> : total time elapsed in seconds since the task started	-0.046

Table 4.1: Feature descriptions and Pearson's correlations with relevance Levels (** indicates statistical significance at $p < .01$ level; * indicates statistical significance at $p < .05$ level).

4.2.3 Cursor Movements

As suggested in the previous section, characteristics of cursor movements such as speed and range could indicate the searcher's reading behavior, and consequently the relevance of the document. For example, low speeds may indicate that the searcher was carefully "reading", while a long vertical range may indicate that the searcher found the document relevant and was willing to explore. The number and frequency of the cursor movements, distance, speed, and the range the mouse cursor travels in pixels (both overall, and its horizontal and vertical components) are measured, as well as the minimum and maximum of horizontal and vertical cursor coordinates.

4.2.4 Vertical Scrolling

Previous research (e.g., [38]) found that the amount a user scrolls correlates with the "interestingness" of a Web document in a non-Web search setting, while in a Web search scenario, another study [51] did not find a strong correlation between the amount of scrolling and the "satisfiability" of a clicked document. In this study, in addition to modeling the overall amount of scrolling, the frequency and speed of scrolling behavior are also modeled, as well as the overall scroll distance and range in pixels. The intuition behind is to capture the searcher's examination patterns. For example, high frequency and speed of scrolling may indicate that the searcher was "scanning" or skipping parts of the document, while a moderate range of scrolling with low speeds may indicate that the searcher was "reading".

4.2.5 Interactions in the Areas of Interest (AOI)

It has been proposed that searchers are more willing to interact with the content when it is relevant. To capture this idea, an "Areas of Interest"(AOI) is defined, as the region in a document where the main content lies, and model the searcher behavior within the AOI. In particular, the number and frequency of cursor movements within an AOI are measured, in addition to these measures for the document

as a whole. Since a typical Web page has its main content on the left half of the page, one AOI was defined as the region of the document with the X-coordinates between 100 and 400 pixels, and the Y-coordinates larger than 100 pixels. More sophisticated estimation of AOI's can be done, but as we will see later, even this simple AOI appears to improve the correlation between the features and document relevance (Section 4.5.1).

4.2.6 Task/Session-level Context

As shown in the recent work [69], task-level information could be valuable for improving relevance estimation. The intuition is that a page viewed in a successful search task is likely to be more relevant while a page viewed in a unsuccessful task, is likely to be less relevant. To detect task success, previously proposed features are incorporated, such as the number of queries, number of clicks, click-through rate (CTR), average dwell time, overall task time, and the number of page views. These features have been shown to be effective in detecting success or frustration in previous studies [68, 50, 1] and are potentially useful in improving document relevance estimation [69].

4.2.7 User Normalization

Previous work has identified significant variation in behavior across different searchers (e.g., [88, 144, 59]). Three methods were proposed to normalize feature values for individual searchers. The first method subtracts the mean of the feature values for a user, from the original feature values (most common approach); the second method subtract the median feature values for the user, as it is typically more robust to outliers than the first approach; the third method uses z-score normalization, which transforms the original feature distribution into normal distribution by scaling the difference between the original value and mean by the standard deviation.

4.3 Relevance Estimation Models

This section describes the machine learning algorithms used. The relevance estimation is formulated as a regression problem, and two popular regression algorithms were experimented with.

4.3.1 Ridge Regression (RR)

The first algorithm is Ridge Linear Regression, which is a variant of ordinary Multiple Linear Regression, whose goal is to circumvent the problem of predictors collinearity and overfitting. Furthermore, the M5's method is used to select attributes for use in the linear regression for each run. Specifically, the algorithm steps through the attributes and removes the one with the smallest standardized coefficient until no improvement is observed in the estimate of the error given by the Akaike information criterion. The advantages of using such a linear regressor lie in the easy interpretability and time-efficiency in training, which is potentially favorable in a large scale setting. And the disadvantage mainly lies in the less expressive power of the model, which does not capture the non-linear interaction among different features.

4.3.2 Bagging with Regression Trees (BRT)

The second algorithm is Bagging[20], which is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets. The single predictor or weak learner used was the C4.5 regression tree. The advantage of this non-linear regressor is, in contrast, the advanced expressiveness, which can help model the complex relationships among the features, and not surprisingly it suffers from longer training time and may not be applicable in certain large-scale

scenarios.

4.4 Experimental Setup

This section describes the experimental setup on estimating document relevance and re-ranking the documents.

4.4.1 Data

The data set used for the experiments, which has hundreds of search tasks and explicit relevance judgments of visited Web pages, is from a user study conducted by researchers at the University of Massachusetts [50]. The usage data of the participants was tracked, specifically, containing the URLs the searchers visited, the fine-grained interactions with the browsed pages, such as clicks, cursor movements, and scrolling, the time-stamp of each page view and interaction is also recorded. The search tasks in the user study were designed to be representative of Web search and difficult to solve with a search engine (i.e., the answer was not easily found on a single page). This is particularly valuable, since these more difficult and long-tailed search tasks are the main challenge for the state-of-the-art search engines. To distinguish oneself from the others, a search engine provider should ensure that they do a good job on such search tasks, and as we will see later in Section 4.5, the proposed techniques indeed improve relevance estimation and ranking for such difficult search tasks.

The original dataset is publicly available online ². Similarly, the processed data and source code for this study is available at <http://ir.mathcs.emory.edu/data/WWW2012/>. Next, the details of the user study and the collected data are described (additional information can be found along with the original dataset).

User study: The study relied on a modified version of the Lemur Query Log Tool-

²<http://ciir.cs.umass.edu/~hfeild/downloads.html>

bar³ for Firefox browser. To begin a task, participants had to click a ‘Start Task’ button. This prompted them with the task and a brief questionnaire about how well they understood the task and the degree to which they felt they knew the answer. They were asked to use any of four search engines: Bing, Google, Yahoo!, or Ask.com and were allowed to switch at any time. Links to these appeared on the toolbar and were randomly reordered at the start of each task. Users were allowed to use tabs within Firefox.

Explicit Judgments: Each time the participants navigated away from a non-search page, they were asked the degree to which the page satisfied the task on a five point scale (“1” indicates the page “did not satisfy the information need at all” and “5” indicates that the page “completely satisfied the information need”), with an option to evaluate later.

This self-reported explicit judgment was used as the ground truth for document relevance. A total of 211 tasks were completed, feedback was provided for 463 queries and 694 visited pages. For the experiments, the set of page views used have dwell time at least one second and at least one cursor coordinate recorded so as to exclude artificial URL visits (e.g., URL redirections) that are recorded in the dataset and focused on modeling the initial visit of a document in each session as subsequent visits of the same document typically exhibit larger variance in behavior and the dataset consists of only a very small portion of such subsequent page visits. As a result, the final dataset contains 666 page views with relevance judgments.

4.4.2 Evaluation Metrics

Given a feature vector \mathbf{x} of post-click page view, the explicit judgment of page relevance y , and a regression function $f(\mathbf{x})$ (where (\mathbf{x}, y) is an instance of the test dataset D), the performance on predicting document relevance is evaluated using the standard measure of correlation, and evaluated its performance on re-ranking documents using the standard measure of normalized discounted cumulative gain.

³<http://www.lemurproject.org/querylogtoolbar/>

Correlation: Pearson’s correlation $\rho_{f(\cdot),S}$ between the document relevance predicted by $f(\cdot)$ and true document relevance y across all instances in the test data D is given by:

$$\rho_{f(\cdot),S} = \frac{\sum_{(\mathbf{x},y) \in D} (f(\mathbf{x}) - \mu_{f(\cdot)})(y - \mu_y)}{(|D| - 1)\sigma_{f(\cdot)}\sigma_y}$$

where μ is the observed sample mean and σ is the observed sample standard deviation. This correlation coefficient is helpful for detecting the presence of informative predictions, even in the presence of shifting and scaling. The ideal value for correlation is 1.0, with a value of 0 showing no observed correlation.

Normalized Discounted Cumulative Gain at K (NDCG_k): as a standard metric of search engine providers, given a ranked list of documents for a search task, $NDCG_k$ [80] measures the quality of a ranked list at position k , as follows:

$$NDCG_k = \frac{DCG_k}{IDCG_k}, DCG_k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(1 + i)}$$

where $IDCG_k$ is the DCG_k value of the ideal ranking with respect to the actual document relevance, and the rel_i is the relevance judgment, which is at a five point scale. DCG_k aims to penalize the ranked list with highly relevant documents appearing at lower positions, with the graded relevance value reduced logarithmically proportional to the position of the result. $NDCG_k$ of 1.0 indicates a perfect ranking that is identical to $IDCG_k$ and smaller values indicates worse rankings. The $NDCG_k$ were first computed for each individual search task and then averaged into one $NDCG_k$ to summarize the quality of the ranked list provided by each method. Various k values were evaluated.

4.4.3 Methods Compared

The following different methods for estimating document relevance were considered, including methods using individual feature groups, combined feature groups,

with or without user normalization, for both the linear regressor RR and non-linear regressor BRT.

DTR Baseline: a strong baseline model was developed that utilizes signals from click Dwell time, Task-level context, and the search engine original Ranking (*DTR*). This model is representative of the state-of-the-art methods using dwell time [144] and task-level information[69].

Post-Click Behavior (PCB): the full model with all the feature groups combined, which include cursor movements, scrolling, interactions in areas of interest (AOI) and dwell time, task-level context and rank.

PCB with User Normalization (PCB_User): the full PCB model with user normalization for all feature groups, as described above.

Single Feature Group Runs: models trained on the six individual feature groups were evaluated, namely, dwell time, search engine original ranking, task-level context, cursor movements, scrolling, and interactions in the areas of interest (AOI). In particular, the dwell time, task-level context, and rank feature groups can be considered as three additional baselines to gauge the performance of the models. The three remaining behavioral feature groups are the proposed variants in modeling post-click interactions, and serve as the main building blocks of the full PCB model.

Combined Feature Group Runs: also evaluated were the PCB model with each single individual feature group removed from the full model to test the contribution of different feature groups when other groups are presented. This is important, as some features in different groups could be correlated.

4.5 Results and Discussion

This section reports the experimental results and discuss the findings. The first experiment was to analyze the association between each individual feature and the

explicit relevance judgments, and the second experiment was predict document relevance and re-ranking results based on the estimation, where each individual feature group and some combinations of the different groups were evaluated.

4.5.1 Feature Association with Relevance

To study the association between each individual feature and the explicit relevance judgements, Pearson's Correlation for each feature were computed and conducted statistical significant testing. The results are summarized in Table 4.1, along with the descriptions of the features, significant associations are highlighted: * indicates significance at $p < .05$ level and ** indicates significance at $p < .01$ level. As follows, the discussion is organized by feature groups.

Dwell Time As we can see from Table 4.1, there is a moderate correlation of 0.167 between dwell time and document relevance, which is consistent with previous findings [144], since longer dwell time typically indicates searcher interests in the page. However, as we can see later, some other post-click behavioral signals are actually correlated better with document relevance, suggesting the potential of improving upon dwell time information.

Rank The correlation between search engine result ranking and document relevance is -0.073, which matches the intuition that smaller rank values correspond to higher relevance. However, the correlation is low and insignificant. One explanation is that all the visited documents on a search trail following a click typically share the same rank (as some of which were not ranked) but vary in their relevance levels. This assumption is supported by the observation of a higher though still insignificant ρ of -0.094 when the correlation is computed over only pages that were ranked by the search engines. This low correlation of the search engine result ranking with relevance reveals the difficulty of the search tasks in the dataset.

Cursor Movements As suggested in the previous section, characteristics of cursor movements are indicative of searcher's reading behavior. Interestingly, no such

tendency between the cursor movement features and the document relevance was observed. Starting from the beginning of the list, the amount of cursor movements (i.e., *cursorcnt*) exhibits a similar level of correlation of 0.164 as dwell time, which makes sense, as the longer the time the searcher spent on a page the larger amount she might move the cursor.

A more interesting question then is, whether the cursor movements provide some additional information about document relevance – as discussed later, cursor movements and dwell time provides complementary information – and based on the results in this section alone, stronger associations from some of the cursor features we were actually already observed. For example, the maximal y coordinate of the cursor (i.e., *y_{max}*) exhibits a stronger correlation of 0.243 with relevance, which suggests that the further down the searcher moves the cursor the more likely she found the page to be relevant. This is consistent with the observation from the case studies (Section 4.1) – searchers tend to use mouse more actively and “read” when the page is relevant while the page is not relevant, keep mouse still and “scan”, in which case, it is less likely that she would move mouse further down. Note that there is a difference between scrolling down and moving mouse down – as we can see from the table, the correlation between maximal scrolling and relevance is only a insignificant -0.026, one possible interpretation may be that searchers tend to scroll when “scanning” and keep mouse still, while more likely to move the cursor to interact when interesting information is found.

Another interesting observation is about cursor movement speed: while overall the amount of cursor movement is correlated positively with document relevance, the speeds, both in vertical and horizontal directions, have negative correlation, which matches the observation and intuition: lower speed of cursor movements is indicative of “reading”, which is more likely to happen when the page is relevant. As for horizontal movements, the distance cursor travels exhibits a significant positive correlation of 0.101. This feature captures the horizontal movement of reading aid behavior illustrated in Figure 4.1, the possible explanation of lower correlation based on the case studies is that the horizontal movement behavior happens less fre-

quent than vertical moves, but when it does happen, it typically is a strong indicator of “reading” [118].

Vertical Scrolling In agreement with previous research [51], no significant correlation between the amount of scrolling (i.e., *scrlcnt* and *scrlmax*) and relevance was observed. However, interestingly, significant negative correlations were found of scrolling frequency and scrolling speed of -0.206 and -0.212 respectively, which well supports the hypothesis that high frequency and speed of scrolling indicate “scanning” behavior, which in turn, suggests lower document relevance.

Interactions in Areas of Interest (AOI) The intuition behind the AOI features is that searchers are more likely to interact with the content when it is relevant. Therefore, the expected position of the main content of Web page was specified as the AOI, hypothesizing that the interactions within the AOI are more indicative of document relevance. As we can see from Table 4.1, AOI features exhibits higher correlations as compared to their overall counter-parts. For example, the correlation of AOI dwell time, which is the dwell time accumulated when the cursor is within the area of interest, increases substantially from correlation of 0.167 to 0.227 while the correlation of AOI cursor frequency increases even more significantly from -0.082 to -0.195.

Task/Session-level Context In agreement with previous work [69], the task-level information is indeed found valuable in inferring document relevance. In particular, a document in a more successful search session is indeed more likely to be relevant, which is supported by statistically significant correlations between *CTR* and relevance, as well as the average dwell time and relevance. In contrast, a document is found less likely to be relevant in a less successful search session, which is indicated by the significant negative correlations between relevance and features representing task length (e.g., query count and dwell time). This makes intuitive sense, since a long session typically indicates the more efforts searchers have to put in finding the information, a claim supported by previous studies [1, 50].

Next, the findings in predicting documents relevance are discussed, and the performance of each individual feature groups as well as different feature group combinations are compared.

4.5.2 Predicting Document Relevance

This section reports the results and findings in predicting document relevance explicitly judged by the users. For training and testing, 10-fold cross-validation was used with 100 randomized experimental runs. The reported overall correlation was aggregated over all the folds and runs (note that, each instance occurs only once in exactly one fold for each run). The six single feature groups, different combinations of these groups, and the effects of adding user normalization information were evaluated.

Single Feature Group Runs: The results of the single feature group runs are summarized in Table 4.2. As we can see, all the three post-click interaction feature groups outperform the three baseline feature groups using dwell time, task-level information and search engine ranks, as well as the stronger *DTR* baseline that combines these three groups of signals; but none of them is comparable with the full model *PCB*. This trend is consistent across both the linear ridge regressor (RR) and the non-linear bagging regressor (BRT). Specifically, the correlation with relevance for the cursor feature group is the highest, followed by the scrolling feature group, aoi feature group, the task-level, dwell time and rank feature groups. Interestingly, BRT improves the performance of the cursor feature group over RR substantially. One possible interpretation is that the features within the cursor group have complex interactions with each other, which can not be successfully captured using a linear model such as RR.

Combined Feature Group Runs: The results are summarized in Table 4.3. As we can see, all the combined feature groups again outperform the *DTR* baseline that does not incorporate the post-click interaction features. For the ridge linear regression (RR) setting, the best performing model is the combination of all feature groups

<i>Single Feature Group</i>	<i>RR</i>	<i>BRT</i>
<i>PCB</i>	0.399*+	0.411*+
<i>cursor</i>	0.326*+	0.389*+
<i>scroll</i>	0.277+	0.268*+
<i>aoi</i>	0.261*+	0.177*
<i>task</i>	0.201*	0.146*
<i>dwell</i>	0.184*	0.136
<i>rank</i>	0.04	0.136
<i>DTR</i>	0.211	0.231

Table 4.2: Pearson’s correlation between the predicted and actual document relevance for the single feature groups. The groups are listed in descending order of the BRT performance. (* indicates a significant improvement over all the worse-performing groups in the same column at $p < .05$ level, + indicates a significant improvement over the *DTR baseline* in the same column at $p < .05$ level)

PCB and removing any one of the groups decreases the performance significantly. Among the six groups, the contributions of the cursor and scroll groups are the most significant while removing each of the other groups only results in decrease with a small margin. As for the non-linear bagging regression (BRT) setting, only the cursor, scroll, and rank groups contribute significantly when other groups are presented and the additive contribution from the ranking information is the least substantial among the three. The three remaining groups, namely, *dwell*, *task*, and *aoi*, do not seem to contribute additional information when the other groups are presented. One possible explanation is that the non-linear BRT regressor was able to capture the complex relationships among different features and induce the information carried by the features in dwell time, task-level context and AOI interactions, making it unnecessary to incorporate these features when other groups are presented, even though all the feature groups tend to be useful in combination when only a linear regressor such as RR is used.

<i>Combined Feature Group</i>	<i>RR</i>	<i>BRT</i>
<i>PCB</i>	0.399+	0.411+
<i>no.cursor</i>	0.326*+	0.336*+
<i>no.scroll</i>	0.353*+	0.379*+
<i>no.aoi</i>	0.394*+	0.412+
<i>no.task</i>	0.394*+	0.413+
<i>no.dwell</i>	0.395*+	0.414+
<i>no.rank</i>	0.393*+	0.409*+
<i>DTR</i>	0.211	0.231

Table 4.3: Pearson’s correlation between the predicted and actual document relevance for the combined feature groups. The groups are listed in ascending order of the BRT performance. (* indicates a significant decrease in performance from *PCB* in the same column when removing the feature group at $p < .05$ level, + indicates a significant improvement over the *DTR* baseline in the same column at $p < .05$ level)

User Normalization: The effects of adding the user normalization information were further evaluated. The results are summarized in Table 4.4. As we can see, adding user information to the full model (*PCB_User*) further improves the performance in predicting document relevance, which was the best-performing model among all the other feature combinations, and as expected, the model also outperforms the *DTR* baseline. In particular, the improvement with a linear regressor was smaller compared to that of the non-linear bagging regressor. This result indicates the existence of variation in behavioral signals across different users, a claim supported by previous research [144, 59, 29]. However, as we have seen, even without the user information, the behavioral patterns seem sufficiently consistent to achieve improvement in estimation performance.

Next, the discussion moves on to be about the results and findings on improving result ranking in aggregate using the estimated document relevance from the proposed models.

<i>Combined Feature Group</i>	<i>RR</i>	<i>BRT</i>
<i>PCB_User</i>	0.420*	0.447*
<i>PCB</i>	0.399*	0.411*
<i>DTR</i>	0.214	0.231

Table 4.4: Pearson’s correlation between the predicted and actual document relevance when adding user normalization features (* indicates a significant improvement over the *DTR* baseline in the same column at $p < .05$ level)

4.5.3 Re-ranking

This section reports the results on re-ranking documents using the estimated relevance from the regressors. For training and testing, 10-fold cross-validation was again used. $NDCG_k$ averaged over all the search tasks across different users was reported. Specifically, combined feature groups with one feature group removed at a time, the *DTR* baseline, and the full models *PCB* and *PCB_User* were compared. As BRT generally performs better than RR, BRT was used for the rest of the experiments.

The feature ablation results are summarized in Table 4.5. The trend is the same as what we have observed in Table 4.3: cursor and scroll feature groups tend to contribute the most, while the rest of the groups contribute marginally when other groups are presented. One interesting difference in this setting is that for smaller K , the contribution of scroll features appears larger than that of the cursor features.

The results of the post-click behavior models, with and without user normalization (*PCB* and *PCB_User*) are reported in Figure 4.3. Both variants of the *PCB* model again outperform the *DTR* baseline, and adding user normalization features *PCB_User* provides moderate improvements in ranking, especially for smaller values of K .

Next, the performance of *PCB* was evaluated on the subset of documents that were ranked by the search engines (i.e., landing pages). The results are summarized in Figure 4.4. For the landing pages, *PCB* and *PCB_User* still consistently

<i>Combined Feature Group</i>	<i>K=10</i>	<i>K=20</i>
<i>PCB</i>	0.579	0.675
<i>no.scroll</i>	0.515 (-11.0%)	0.630 (-6.7%)
<i>no.cursor</i>	0.548 (-5.2%)	0.619 (-8.3%)
<i>no.aoi</i>	0.570 (-1.5%)	0.671 (-0.7%)
<i>no.rank</i>	0.576 (-0.5%)	0.669 (-0.9%)
<i>no.dwell</i>	0.578 (-0.1%)	0.677 (+0.2%)
<i>no.task</i>	0.587 (+1.5%)	0.681 (+0.8%)
<i>DTR</i>	0.515 (-10.9 %)	0.598 (-11.4 %)

Table 4.5: NDCG at K for the combined feature groups with one feature group removed at a time, the groups are listed in ascending order of NDCG@10.

outperform the *DTR* baseline at all values of K , indicating that PCB predictions could be directly usable by a search engine for improving search ranking quality.

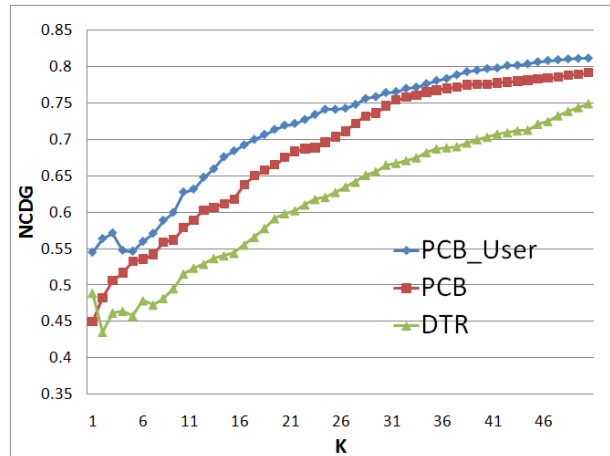


Figure 4.3: NDCG at K for the *DTR* baseline and the full models with (*PCB_User*) and without (*PCB*) user normalization features in re-ranking all the pages.

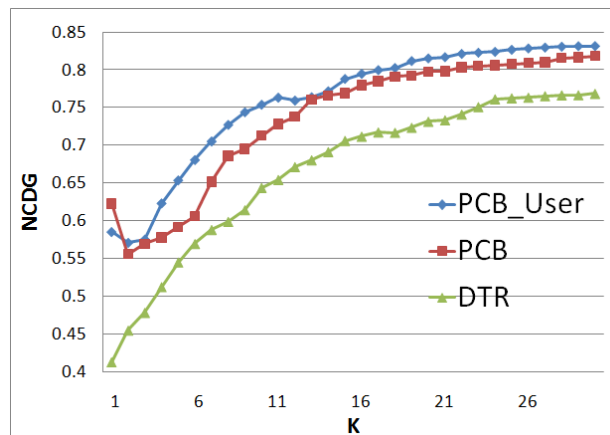


Figure 4.4: NDCG at K for the *DTR* baseline and the full model (*PCB*) in re-ranking only the landing pages.

4.6 Summary

This chapter introduced a new model for representing the searchers' post-click behavior (PCB) that captures not only dwell time and task-level information, but also fine-grained user interactions *after* clicking on a search result, such as cursor movements and scrolling. To the best of my knowledge, PCB is the first successful attempt to exploit such "low-level" post-click behavioral signals to identify the basic patterns of "reading" and "scanning" behavior, as well as more complex combinations of these, coupled with expressive features to capture these examination patterns automatically (Section 4.2).

The experimental results show that these behavioral signals indeed correlate with searchers' explicit judgments of document relevance, and provide additional valuable information beyond dwell time and session-level information. Specifically, the distance and range the cursor travels, as well as movement speed, especially its vertical component, are was found to be among the most predictive signals of document relevance; while the amount of scrolling itself was not found to be strongly correlated with document relevance, the frequency and speed of scrolling was. In combination, these signals enable PCB to exhibit significant improvements of relevance estimation, as well as significant improvements in re-ranking the documents based on this relevance estimation. Finally, when user information is available (e.g., for long-term users of a search engine), adjusting the PCB model for each user's "normal" profile can further improve the prediction performance.

In summary, this study has laid the groundwork for exploiting fine-grained post-click search behavior for document relevance estimation, identifying common page examination patterns and operationalizing the insights in a novel PCB model for effective relevance prediction. Together, the proposed methods enabled substantial improvements of relevance estimation, and the resulting document ranking over and beyond dwell time alone.

Chapter 5

Evaluating Search Experience

In Chapter 3 and Chapter 4, techniques were introduced on improving the information retrieval effectiveness through better intent inference and direct estimation of document relevance. This chapter focuses on developing techniques for the complementary problem of automatically evaluating the search experience, which is important to gauge the performance of a search engine, identify the potential weak areas to improve, and intervene and provide additional assistance in real-time when users are detected to be struggling. Three studies were conducted, focusing on different aspects of this problem.

In the first study, a principled framework of studying Web search success and a novel data collection infrastructure was proposed for performing controlled, yet realistic, scalable, and reproducible crowd-sourcing studies of search behavior. These techniques addressed the tension between the relatively small-scale, but controlled lab studies, and the large-scale log-based studies where the search goals are unknown. In addition to the relative large scale (hundreds of users) and the known search goals for each task, the proposed techniques also benefit from the objective and well-defined search success metrics that can be used to evaluate whether the search goal was actually achieved. A search success model based on CRFs was developed and trained on the behavior data collected from the proposed infrastructure, and was demonstrated to be effective under different definitions of Web search success. The bulk of this part has been published as [1].

In the second study, additional fine-grained behavioral evidence was modeled to improve the prediction of search success. While result click-through, dwell time

and sequences of searches [68, 50, 64, 1] are effective in predicting searcher success, the prediction accuracy is largely limited as such models are agnostic about how users actually view and interact with the visited pages. For example, as we have seen in Chapter 4, the dwell time does not provide a full picture of the search experience – spending a long time on a landing page might suggest that the searcher was struggling and could not find the relevant information if she was actually scanning instead of reading during the stay. Similarly, spending some time carefully reading a result snippet before clicking on it is different from quickly scanning the whole search result page within the similar amount of time – in the latter case, the user appears to be less satisfied with the returned results and is less likely to be successful. As an example, Figure 5.1 shows the mouse cursor heat maps overlaid on the visited search engine result pages from a successful and an unsuccessful search sessions respectively. As we can see, the mouse cursor positions in the unsuccessful session (Figure 5.1(b)) are more spread-out than in the successful session (Figure 5.1(a)) and spread to the lower part of the result page, suggesting that when the searcher examines multiple search results (especially the lower-ranked ones) before click, she is more likely to be unsuccessful in finding the needed information. The bulk of this study has been published as [62].

In the third study, further extension in this thread was conducted in the mobile search space. Recently, as mobile devices, such as smart phones, have become an increasingly popular platform for browsing and searching the Web, it is becoming crucial that the Web search experience on a mobile phone is satisfactory. However, to the best of my knowledge, no work has been done to understand how the behavioral patterns on a mobile device can reveal the success and satisfaction of a search task. There are many differences in the usage of computers and mobile devices [36, 86], and so it is unclear whether the models of search success developed for the desktop setting (e.g., using fine-grained user signals, such as scrolling and mousing [28, 58, 77]) would translate to the mobile search setting. As an attempt to close this gap, this study aims to automatically predict search success and satisfaction in mobile Web search from behavior on mobile phones with touch screens. In



Figure 5.1: Example mouse cursor heat maps: (a) - search result page in a successful search session; (b) - search result page in an unsuccessful search session.

in addition to previously studied behavioral signals of search success, client-side interaction features were investigated, which include zooming (an example of zooming interaction was given in Figure 5.2), scrolling/sliding, and orientation changing - increasingly common in modern smart phones - and show that these features significantly improve prediction accuracy. The bulk of this study has been published as [65].

In summary, the main contributions of this chapter include:

- A flexible and general informational search success model for in-depth analysis of search success and failure for different definitions of success.
- Effective machine learning-based techniques for predicting and analyzing different types of search success.

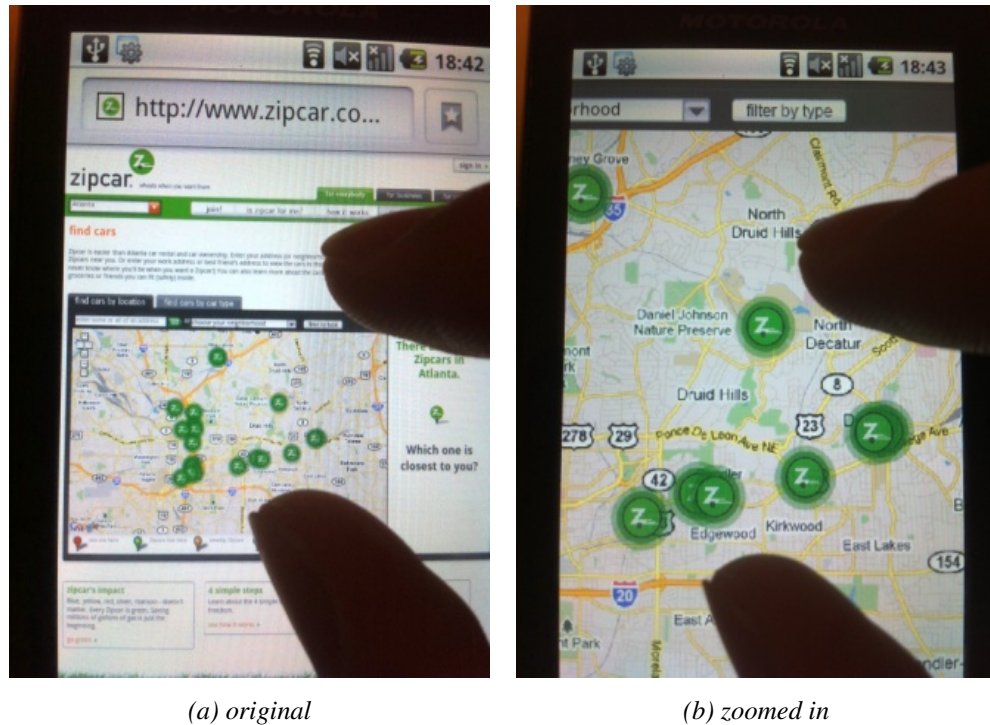


Figure 5.2: An example of zooming interaction with an Android smart phone with a touch screen: (a) the user was viewing the original picture; (b) the user zoomed in using two fingers on the touch screen.

- FSB, a effective novel model of success prediction that jointly models fine-grained pre- and post- click behavior at the session-level.
- Effective machine learning-based techniques for predicting mobile search success using client-side interactions.

5.1 Predicting Search Success with UFindIt

This section introduces a principled search framework and a novel competition-based data collection methodology. The proposed approach is based on enticing participants to compete in a game-like setting, to find answers to real informational questions, while tracking the resulting search behavior.

This approach has a number of advantages over previously reported search evaluation methods: i) the information needs are real, well-defined questions selected from community forums such as wiki.answers.com and Yahoo! Answers. ii) the goals (needed information) are known to both the searchers and the assessors, and are well-defined, allowing for objective measures of success; and iii) sufficient amount and diversity of search behavior can be acquired for difficult queries (which are relatively rare among all queries submitted to search engines), thus enabling in-depth study of the behavior characteristics for these difficult and rare queries that were not previously available through passive log analysis.

Next, the search success model is presented, which allows formulating hypotheses that motivate the study design, and the experiments described in the rest of the section.

5.1.1 Search Success Model

To better analyze the search process, a simple, yet powerful four-stage QRAV (Query-Result-Answer-Verification) model of an informational search success was proposed. This model is primarily geared towards analyzing and describing searches with specific, direct information needs, such as those phrased as factoid questions.

The process of successfully answering factual queries was conceptually divided into several parts. First, the user should correctly understand the question and issue a relevant query (*Q for Query formulation*). If the query retrieves at least one target document in the top 10 results, it is considered a *Good Query*. Then, the user has to find the correct result on a search engine result page (SERP), and click on it to examine the document (*R for Result identification*). If a result document contains the correct answer, it is considered a *Good URL*, indicated by R^+ . Furthermore, how many clicks away this document was from the SERP was considered, indicated by the subscript. For example, if a document containing a correct answer was the last in the search session, it is denoted as R_L^+ . The next step is extracting an answer from a document (*A for Answer extraction*). Finally, the answer is verified that

it correctly answers the question and is in fact supported by the document (*V for Verification of the answer*).

Thus, the final success of a user in finding an answer to a question depends on successfully performing each stage in the QRAV model, represented as $Q^+R^+A^+V^+$. In a case where a user issues a good query and clicks on a good document, but submits an incorrect answer, the outcome would be represented as $Q^+R^+A^-V^-$. Finally, if the success in a particular stage is unknown (or not considered), the “?” mark is used. The model is illustrated graphically in Figure 5.3, where the states in the process are represented by circles and the arrows represent possible state transitions. Note that some transitions in the model are not possible, e.g., by definition it is not possible to directly go from a bad query (Q^-) to a good result (R^+).

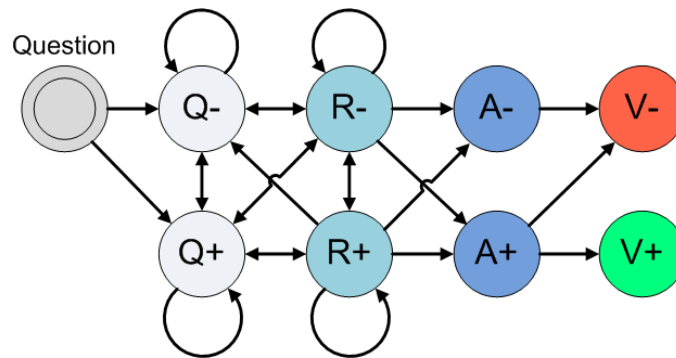


Figure 5.3: Possible state transitions in QRAV model.

The QRAV model can describe (and estimate) the success factors at each stage of the search process, and naturally represent previously posed definitions of search success:

- $Q^+R_*^+A^+V^+$: The correct answer was found and validated to be supported by a good document - it is a search success in the strictest sense, most similar to the definition of a correct answer in a TREC question answering track [72].
- $Q^+R_*^+A^+V^?$: An answer was found on a good result, and submitted - which means that the participant was satisfied with the session and believed that she

found an answer (but the answer could still be incorrect). This definition of success matches the definition of Aula et al. [10], where the users submitted the answer to a difficult search task, but the answer was not validated.

- $Q^+ R_*^+ A^? V^?$: A good URL was found - the user visited a relevant page, but did not necessary extract the correct answer. This definition matches the model based on analysis of search engine click logs by independent assessors as in [68].
- $Q^? R_L^+ A^? V^?$: A good URL was found and it was the last in the search session. This follows the ideas of marginal document utility in a search session [47] - after viewing the last document in the search session, the user is satisfied and stops searching.

Having defined the measures of success, the sets of actions (behavior) that correlate with each of the above definitions of success now can be analyzed. In other words, at each stage of the process, searchers perform actions, some of which are indicative of a success or failure, or a continuation at one of the QRAV stages above. The proposed study, described next, was designed with the goal of being able to validate, and isolate the success or failure of the searcher at each stage of the process. Using this model one can also analyze the corresponding behavioral clues for predicting the success or failure at each stage of the search process.

5.1.2 Acquiring Search Behavior Data

The overall design of the study was modeled on a game, more precisely as a search competition: the participants played a game consisting of 10 search tasks (questions) to solve, with a timer displaying the number of seconds remaining, shown to the subject (see Figure 5.4). The stated goal of the game was to submit the highest possible number of correct answers within the allotted time. Overall, four game rounds were used in this study, with participants recruited and scored separately for each round. The top “players” in each round (typically, those successfully posting

a correct answer to 7 or 8 out of the 10 questions) received a bonus payment. More details about the search tasks, study procedure, and system implementation can be found in the original paper of this study [1].

Figure 5.4: An example search game interface, which has the question, the search query window, and a dropdown box for choosing a search engine to use (Google, Yahoo! Search, or Bing). When the answer is found, the participant submits it together with the supporting URL. The query result page is opened in new tab, allowing natural querying and browsing.

5.1.3 Predicting Search Success

This section describes the models, features, and algorithms used for analyzing and predicting the success of searchers.

Algorithms

Markov Model (MML+Time): As the baseline state-of-the-art model, the Markov Model approach introduced in the reference [68] was adapted. As in the original model, the states are the types of visited page “Q” for SERPs, “ R_1 ” for pages clicked from SERP or “E” for end-of-session. The only difference from original

model is that the data in this study does not contain sponsored-search clicks and search engine-specific links like “related search”, but information about pages visited from hyperlinks – those pages correspond to additional model state “ $R_{>1}$ ”.

The transitions of Markov Model are events of new page visit. Given the session success $B \in \{1, 0\}$ (success or fail), the transition probability between two states $s_i, s_j \in \{Q, R_1, R_{>1}, E\}$, is estimated on the training set:

$$P(s_i \rightarrow s_j, \Delta t|B) = \frac{N_{s_i, s_j, B}}{N_{s_i, B}} \Gamma(\Delta t, k, \theta) \quad (5.1)$$

where $N_{s_i, s_j, B}$ is a frequency of transitions $s_i \rightarrow s_j$ in sessions with a given result B , $N_{s_i, B}$ is a frequency of state s_i in those sessions, and Δt is a time delta between events s_i and s_j . The model in reference [68] assumes that has the Gamma distribution $\Gamma(\Delta t, k, \theta)$ with parameters k and θ , estimated from the training set. Parameters k and θ also depend on s_i, s_j , and B .

The trained Markov Model is used to predict session success from the search behavior data. For the sequence of states with known time deltas $S = s_0 \xrightarrow{\Delta t_1} s_1 \xrightarrow{\Delta t_2} \dots \xrightarrow{\Delta t_n} s_n$ the log likelihood of success and failure is estimated as:

$$LL_B(S) = \sum_{i=1}^n \log P(s_i \rightarrow s_j, \Delta t|B) \quad (5.2)$$

and the session success is defined as:

$$Pred(B) = \begin{cases} 1 & \text{if } LL_1(S) \geq LL_0(S) \\ 0 & \text{otherwise} \end{cases} \quad (5.3)$$

The performance of the Markov Model was tested, with and without the time delta distribution features, and the experiments confirmed that incorporating time delta distribution indeed improves performance. This agrees with the results described in reference [68] and validates the implementation.

Conditional Random Fields (CRF): an extension of the Markov Model approach above was used, by adapting the CRF model [90] for the task. The benefit of CRF

is that it allows to augment the Markov Model with additional search behavior features, derived from previous works in references [50, 68, 145], and described next. The Mallet¹ implementation of CRF was used, freely available for research.

A CRF allows defining a conditional probability $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\lambda})$ over the hidden state sequences $\mathbf{y} = \{y_1, \dots, y_n\}$ given a particular observation sequence of n page views $\mathbf{x} = \{x_1, \dots, x_n\}$, and the CRF parameters $\boldsymbol{\lambda}$ that are estimated at training. At training time, the hidden state of an observation (i.e., a page view) can be assigned a “+” or “-” label, depending on whether the user was successful in the task. Alternatively, one can also assign “ Q^+ ”, “ Q^- ”, “ R^+ ”, “ R^- ” labels to the intermediate stages in the session to allow more fine-grained modeling. At test time, given an observed page view sequence \mathbf{x}' , the most likely state sequence \mathbf{y}' can be inferred by maximizing the conditional probability $P(\mathbf{y}'|\mathbf{x}', \boldsymbol{\lambda})$ using the formula:

$$P(\mathbf{y}|\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_j \lambda_j F_j(\mathbf{y}, \mathbf{x})\right) \quad (5.4)$$

where $\frac{1}{Z(\mathbf{x})}$ is a normalization factor and $F_j(\cdot)$ is the j^{th} feature function, which could be either a state feature function or a transition feature function. An example configuration is shown in Figure 5.5. Each observation corresponds to a page view (i.e., either a search engine result page, Query _{i} ; or the landing page of a clicked result, Result _{i}), and is represented by a vector of features introduced next, (Table 1) such as dwell time on the page, query length in words, and number of queries in a session.

To experiment with the tradeoff between the precision and recall, the marginal probability of the last hidden state $y_n = “+”$ was used as the classification confidence, since the last state indicates whether a searcher is successful or not across all potential CRF configurations. The marginal probability is computed by summing over the probabilities of all labeled sequences Y^+ that end with label “+” in their last states, according to the following formula:

¹Available at: <http://mallet.cs.umass.edu/>

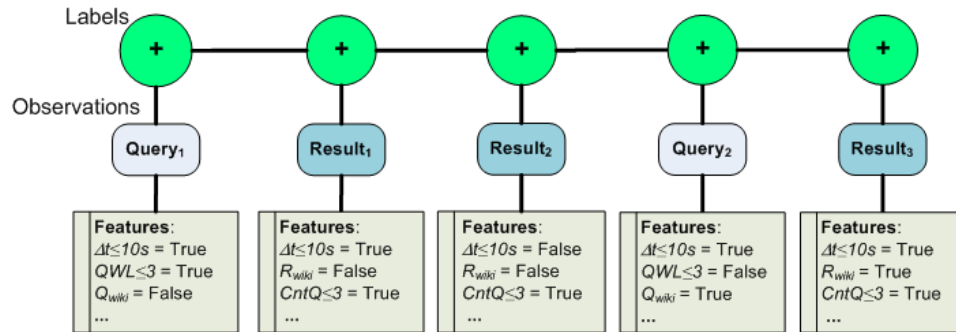


Figure 5.5: CRF implementation of session-level model. The labels represent overall session success; the observations at each step in the sequence are the features in Table.

$$conf = \sum_{\mathbf{y} \in Y^+} P(\mathbf{y} | \mathbf{x}, \lambda) \quad (5.5)$$

For the experiments the Mallet implementation of CRF was used, which allows only nominal features. Therefore, the numeric features were discretized using increasingly large thresholds. For numeric features, the used discretization thresholds are shown in the column “Bins”. The complete set of behavioral features used for the CRF model is reported in Table 5.1. Also used were aggregated behavioral features for analysis of user success, as described in the next section. The aggregation function is shown in the column “User” for the features that have a reasonable interpretation for describing an individual searcher.

Search Behavior Representation: Features

Search behavior was represented by adapting and extending the features introduced in previous studies. Specifically, the browsing features from [145] were used. Additional features, such as session duration and number of viewed pages were adapted from [50]. Also added was a feature for the average page trail length.

For the analysis, only used features were those could be reasonably matched to user’s search skills or expertise. For example, one hypothesis was that highly successful searchers view more results (i.e., CntR is higher), use more advanced syntax

(i.e., QADV is higher), and perform search faster (i.e., the duration of session in seconds Δt is lower). As will be shown, the first two hypotheses are supported by the collected data, but the third is not. To compute these features, the base feature values in Table 5.1 are aggregated for each participant according to the rules presented in the last column, and then averaged over all the sessions that the user has completed.

5.1.4 Results and Discussion

This section presents the results of analyzing and predicting search success. First, the participant data and descriptive statistics are described. Then, the behavior patterns of successful vs. unsuccessful searchers are contrasted. Then, the behavior patterns associated with search task difficulty were analyzed in order to automatically predict search success for the various definitions of success in the proposed model.

Participant and Search Behavior Data

A total of 200 MTurk participants finished at least one of the game rounds. The user sessions were both automatically and manually checked to detect violations of game rules. For example, some users did not use the search interface provided by the study, or used unsupported browsers, despite being warned not to do so in the task instructions. The users who did not answer even the easy, effectively trivial questions were filtered out, as it indicated either poor understanding of the game rules, or an attempt to make a quick buck without effort. After this filtering, 159 users (79.5%) remained in the dataset. The data for these users consists of 14873 search sessions, distributed among 40 distinct questions in 4 game rounds. All these 159 users were paid the base \$1 payment. The top 25% of the searchers (ranked by the number of correct answers submitted) were paid an additional \$1 bonus. *In total, the user payments cost less than \$250.*

Overall, there were from 30 to 50 valid search sessions collected for each ques-

tion. For each session the browsing events were extracted from the log of the deployed proxy. Each session was finished by either submitting an answer (87% of sessions), or by passing to the next question. The submitted answers were manually marked as either correct (65% of all sessions) or incorrect. An answer was considered correct if the page of the submitted URL indeed contained the submitted answer. There were 4382 search engine queries in the collected log data, and 14676 page visits. This data is available at (<http://ir-ub.mathcs.emory.edu/uFindIt/>). Additional data statistics, including the overall average session times, number of actions, average query length, and others are reported in Table 5.2. These statistics largely agree with the published statistics from a similar study of web search success reported by Aula et al. [10]. The small quantitative differences can be expected, as the study focuses on relatively difficult information gathering tasks.

Prediction of Search Session Success

This section reports results on the prediction of session success by using the behavioral features as input. The proposed Conditional Random Fields (CRF) algorithm described in section 5.1.3 was compared to both the naïve baseline algorithm that always predicts success (the majority class), as well as to the state-of-the-art Markov Model method that incorporates time distribution between actions (MML+Time) described in [68] and summarized in section 5.1.3.

Four-fold cross-validation was used in the following manner: for each of the four game rounds, the model was trained on all sessions from the other three games, and apply the trained model to predict the search success of the current game. Thus, there are four folds of roughly equal size. For each fold, neither the users nor the questions intersect. Algorithms are compared by accuracy and F-measure, macro-averaged over the positive (successful) and negative (unsuccessful) classes. The results are presented in Table 5.3, which shows that the proposed CRF model exhibits significant improvement over both the baseline and MML+Time models proposed

in [68] for all definitions of success except the first one.

Feature significance: The relative significance of behavioral features used for training CRF are now explored by using a subset of the features, starting from one feature, and use a greedy search to extend the used subset, one feature at a time, by adding the best of the remaining features. In each step the feature that gives the highest F1 performance of the CRF algorithm for predicting $Q^?A^?V^?$ success is chosen. The results are shown in Table 5.4, and the best results obtained by this greedy feature selection for each definition of success are shown in Table 5.3. As we can see, the first, most significant feature (State) is the same one as reported in [68] the search action itself. Interestingly, the time interval between the actions and the choice of the web search engine are the two next most useful features. This makes sense, as the faster searchers are also likely to be more advanced or experienced, and are also more likely to experiment with switching search engines (encouraged by a drop-down box in the proposed search interface). Finally, the position of the search result clicks provides additional indication of the search result quality – which in turn indicates the presence of a good query.

Next, the differences in the performance of the MML variants and the proposed CRF system are explored, for different definitions of search success (Section 5.1.1). Figures 5.6 (a-d) report the precision vs. recall plots of identifying the Successful class. CRF performs best, and significantly better than MML for the definitions proposed in [68] (b) and for the definition proposed in [47] (d). For the other definitions of success (e.g., the most strict one (a)), the improvement of CRF is less striking, while MML variants exhibit performance comparable to the reports in the original study [68].

5.1.5 Real World Success Prediction:A Log-based Study

We have seen that the proposed model can successfully predict success of over a hundred participants in the tournament-like setting. Can one use the resulting model, trained on the contest data, to predict search success in a real-world search

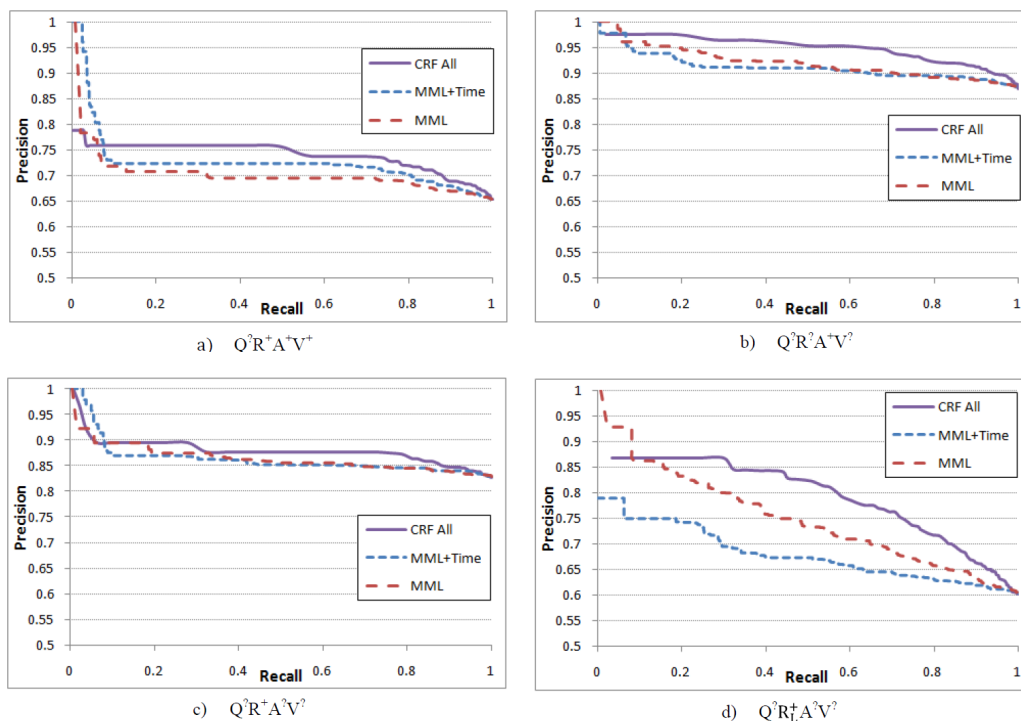


Figure 5.6: Recall-precision curves for compared algorithms, for different definitions of session success in QRAV model.

engine log? To answer this question, a large log of web searches is used, which is performed from hundreds of shared-access workstations in a major university library, to predict search success using the models trained on the collected contest data.

Experimental Setup

The data was collected by instrumenting over 100 shared-use workstations at the Emory University library using the EMU toolkit², with participants explicitly opting in to allow the searches to be tracked for library improvements (with roughly 60% opt-in rate). 16,693 search sessions (using almost primarily the Google search

²<http://ir.mathcs.emory.edu/EMU/>

engine) were collected over a period of 6 months. A sample of 175 search sessions consisting of more than one query (and thus more likely to be a non-trivial search) were manually labeled by the researchers to be successful or not, using the methodology and criteria outlined in Hassan et al. [68]. Specifically, the assessors used their best guess of the searcher intent based on the sequence of queries submitted, clicks on the results, and by manual examination of the visited result pages (recorded in the proxy log). 29% of the sessions were labeled as successful, 28% as unsuccessful, and 43% as unknown where the assessors were not able to infer the searcher intent or determine whether the visited results satisfy the search.

Methods Compared

The CRF was trained on all the contest data, using all of the features in Table 5.1. The algorithms were trained using the two most successful definitions of search success in QRAV model, namely $Q^?R^+A^?V^?$ (finding a good document) and $Q^?R_L^+A^?V^?$ (finding a good document as last in the search session). Then, the CRF model trained on the UFindIt game data was applied to log data.

Results and Discussion

Table 5.5 reports the results of predicting search session success. While the absolute accuracy and F1 values are lower than those on the original search contest data, the predictions significantly and substantially outperform the baseline. This experiment demonstrates that training a success model on search contest data can have significant practical applications, by directly applying the trained models to estimate search success of users of a production search engine.

<i>Feature</i>	<i>Description</i>	<i>CRF Bins</i>	<i>Aggregation</i>
state	Type of visited pages $s \in \{Q, R_1, R_{>1}, E\}$		
Δt	Time delta between previous state and current state	$\leq 3s, 10s, 30s$	$\sum \Delta t$
Q_{engine}	One of google, bing, yahoo		
$Q_{abandoned}$	True if no clicks for the query		
QWL	Query word length	$\leq 3s$	Avg
Q_{wiki}	True if wikipedia.org is on SERP		
Q_{ADV}	True if the query use advanced query syntax. (i.e., queries that use search operators – quotes, “+” operator, and field operators like “site:” and “allintext:”.)		Avg Count
Q_{DT}	Query Deliberation Time - minimum time delta between query and first click		Avg
R_{wiki}	True if visited page is on wikipedia.org		
$R_{Q_{serp}pos}$	Position of SERP click	$\leq 2, 5$	Avg
R_{trail}	Length of trail from search engine result page, defined as the number of clicks from SERP	≤ 1	Avg
$ref_{serp/start}$	True if visited page was clicked from the SERP or from the start of a game (these features are extracted from HTTP Referer header, and could catch some patterns of non-linear browsing, when user uses several browser tabs)		
<i>Session-level aggregates</i>			
CntQ/CntR	Count of queries and pages in the session	$\leq 1, 3$	Avg
QPS	$QPS = \frac{CntQ}{\sum \Delta t}$ – average number of Queries submitted by a user Per Second		Avg
CPQ	$CPQ = \frac{CntR}{CntQ}$ – average number of result Clicks Per Query		Avg

Table 5.1: Behavior features used for CRF. “Q*” features are defined only for SERPs (if the state=Q), “R*” features are defined only for non-SERP pages. Discretization thresholds are shown in the “Bins” column. For the features used in the search behavior analysis, the aggregation function is shown in the last column.

<i>Statistic</i>	<i>All</i>	<i>Successful</i>	<i>Unsuccessful</i>
Count	1487	971	518
Average duration, sec. ($\sum \Delta t$)	215 (223)	182 (176)	276 (384)
Average number of query terms/query (QWL)	6.0 (4.8)	5.8 (4.7)	6.6 (5.1)
Average number of queries per session (CntQ)	2.9 (6.7)	2.3 (5.0)	4.0 (12.4)
Ratio of queries with operators (QADV)	0.05 (0.07)	0.05 (0.06)	0.05 (0.13)

Table 5.2: Descriptive statistics for search sessions collected with the UFindIt game. The corresponding statistics from reference [10] are shown in parentheses.

Success Definition	<i>Baseline</i>		<i>MML+Time</i>		<i>CRF (All)</i>		<i>CRF (Selected)</i>	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
$Q^?R^+A^+V^+$	0.65	0.40	0.61	0.58	0.68	0.60 (+4%)	0.68	0.62
$Q^?R^?A^+V^?$	0.87	0.47	0.72	0.55	0.86	0.64 (+17%)	0.88	0.66
$Q^?R^+A^?V^?$	0.83	0.45	0.66	0.53	0.80	0.57 (+8%)	0.81	0.59
$Q^?R_L^+A^?V^?$	0.60	0.38	0.59	0.53	0.68	0.66 (+26%)	0.69	0.67

Table 5.3: Prediction of search session success for different levels of success in QRAV model. Relative improvement against MML+Time model is shown in parenthesis.

<i>Feature</i>	<i>F1</i>	<i>Accuracy</i>
State	0.624	0.675
$+\Delta t_{\leq 10}$	0.655 (+5%)	0.680
$+Q_{engine}$	0.666 (+1.7%)	0.680
$+R_1_{serp_pos} \leq 2$	0.670 (+0.6%)	0.687
$+R_1_{serp_pos} \leq 5$	0.671 (+0.1%)	0.686

Table 5.4: Prediction of search success by the CRF model, when adding one best-performing feature at a time.

<i>Success Definition</i>	<i>Baseline</i>		<i>CRF (All)</i>	
	<i>Acc</i>	<i>F1</i>	<i>Acc</i>	<i>F1</i>
$Q^? R^+ A^? V^?$	0.51	0.34	0.55 (+8%)	0.52 (+53%)
$Q^? R_L^+ A^? V^?$	0.51	0.34	0.53 (+4%)	0.44 (+29%)

Table 5.5: Prediction of search success for real-world log using CRF trained on contest data, for success definitions in [68] and [47] respectively.

5.2 Predicting Search Success with FSB

This session presents the *Fine-grained Session Behavior (FSB)* model, which captures fine-grained interactions to improve the prediction of search success. As we have seen earlier, fine-grained interactions such as mouse cursor movements and scrolling behavior on a Web page are valuable signals in inferring search intent (Chapter 3) and estimating document relevance (Chapter 4). As the rest of the section demonstrates, by incorporating these “low-level” signals, *FSB* also achieves significant improvements for predicting search success over the state-of-the-art methods. To the best of my knowledge, *FSB* is the first to jointly models the fine-grained behavioral patterns on both search engine result pages and pages on the search trails [142].

5.2.1 Fine-grained Session Behavior (FSB) Features

Next, the proposed *Fine-grained Session Behavior (FSB)* features are described, which aim to capture the page examination patterns that could be indicative of search success. In addition, also included are features from the queries, clicks and dwell time. The brief descriptions about some of the FSB features along their correlation with the success labels (Section 5.2.4) are reported in Tables 5.6, 5.7, and 5.8 and expanded below. Note that the features in the fine-grained interaction groups, namely, cursor and scroll, are first computed for each page view and then aggregated over the entire search session. These two groups can be further divided into pre-click and post-click sub feature groups, corresponding to behavior on the search engine result pages and the behavior on the pages in the search trail.

Query Features: Query features derived from the query string itself, include the query length in words and characters, average number of characters of query terms, the number of submitted queries, SERP views, and unique queries. Intuitively, the longer the query, the more likely the task is difficult and the searcher ends up unsuccessful. On a session-level, the larger the number of queries users have to submit the more likely that the user is struggling. Notice the subtle difference between

<i>Group</i>	<i>Feature</i>	ρ
Query	<i>avg_qwords</i> : average number of words of the queries in the session	-0.117
	<i>num_queries</i> : total number of queries in the session	-0.522**
Click	<i>num_clicks</i> : total number of clicks in the session	-0.363**
	<i>ctr_q</i> : total number of clicks over number of queries in the session	0.150*
	<i>ctr_s</i> : total number of clicks over number of SERP views in the session	0.358**
Time	<i>tasktime</i> : total time duration in the session	-0.473**
	<i>avg_time_s</i> : average deliberation time on SERP pages in the session	-0.107
	<i>avg_time_c</i> : average dwell time on clicked landing pages in the session	-0.119
	<i>satr_c</i> : the ratio of satisfactory clicks (clicks with dwell time at least 30 seconds)	0.059
	<i>dsatr_c</i> : the ratio of dis-satisfactory clicks (with dwell time at most 10 seconds)	-0.083

Table 5.6: Coarse-grained behavior feature descriptions and Pearson’s correlations with success ratings (** indicates statistical significance at $p < .01$ level; * indicates statistical significance at $p < .05$ level).

submitting a query and viewing a SERP – one query might correspond to multiple SERP views. The former can be captured from server-side logs while the latter can only be obtained with a client-side instrumentation.

Click Features: Click features include the number of clicks and click-through rate (over queries and SERPs). Click-through generally is an indicator of success as searchers click on a document when they think that the document can satisfy their information needs. However, a large number of clicks, especially when paired with an even larger number of submitted queries, might indicate that the clicked documents are actually not relevant and the search goal is unsuccessful. Here, the click-through rate may provide additional evidence about search success.

Time-related Features: Both the time users spend on the SERP and the pages on the search trail are considered. In the literature [51, 68, 60], the former is referred as “deliberation time” while the latter is referred as “dwell time”. Usually, these measurements of time are defined as the intervals, in seconds, between the time the page is loaded and the time the searcher leaves the page. To aggregate the time information across multiple page views in the session, computed are the total time span during the session, averaging time spent on different types of pages, and the

ratio of clicks that result in SAT (dwell time ≥ 30 seconds) and DSAT (dwell time ≤ 10 seconds) [51], following previous research [68, 64].

<i>Group</i>	<i>Feature</i>	ρ
Cursor (SERP)	<i>avg_ymax_s</i> : average maximum y coordinate on SERPs	-0.384**
	<i>min_ymax_s</i> : minimum maximum y coordinate on SERPs	0.163*
	<i>max_ymax_s</i> : maximum of maximum y coordinate on SERPs	-0.330**
	<i>med_ymax_s</i> : median of maximum y coordinate on SERPs	0.073
	<i>num_low_ymax_s</i> : number of maximum y coordinate on SERPs that are below 400 pixels	-0.138*
	<i>num_high_ymax_s</i> : number of maximum y coordinate on SERPs that are above 800 pixels	-0.031
Cursor (Trail)	<i>avg_ymax_t</i> : average maximum y coordinate on trail pages	0.179**
	<i>min_ymax_t</i> : minimum maximum y coordinate on trail pages	0.253**
	<i>max_ymax_t</i> : maximum of maximum y coordinate on trail pages	0.200**
	<i>med_ymax_t</i> : median of maximum y coordinate on trail pages	0.289**
	<i>num_low_ymax_t</i> : ratio of maximum y coordinate on trail pages that are below 400 pixels	-0.132
	<i>num_high_ymax_t</i> : ratio of maximum y coordinate on trail pages that are above 800 pixels	0.361**

Table 5.7: Sample fine-grained cursor feature descriptions and Pearson’s correlations with success ratings (** indicates statistical significance at $p < .01$ level; * indicates statistical significance at $p < .05$ level).

Cursor Movement Features: As suggested in the previous section, characteristics of cursor movements such as speed and range could indicate the searcher’s reading behavior, and consequently the success of the search goal. For example, on a landing page, low speeds may indicate that the searcher was carefully “reading”, while a long vertical range may indicate that the searcher found the document relevant and was willing to explore. The features include the number and frequency of the cursor movements, distance, speed, and the range the mouse cursor travels in pixels (both overall, and its horizontal and vertical components), as well as the minimum and maximum of horizontal and vertical cursor coordinates.

Vertical Scrolling Features: In addition to modeling the overall amount of scrolling, the frequency and speed of scrolling behavior were also modeled, as well as the

<i>Group</i>	<i>Feature</i>	ρ
Scroll (SERP)	<i>avg_ymax_s</i> : average speed of vertical scrolls on SERPs	-0.318**
	<i>min_ymax_s</i> : minimum speed of vertical scrolls on SERPs	0.069
	<i>max_ymax_s</i> : maximum speed of vertical scrolls on SERPs	-0.331**
	<i>med_ymax_s</i> : median speed of vertical scrolls on SERPs	-0.074
Scroll (Trail)	<i>avg_ymax_t</i> : average speed of vertical scrolls on trail pages	-0.087
	<i>min_ymax_t</i> : minimum speed of vertical scrolls on trail pages	-0.071
	<i>max_ymax_t</i> : maximum speed of vertical scrolls on trail pages	-0.068
	<i>med_ymax_t</i> : median speed of vertical scrolls on trail pages	-0.131

Table 5.8: Sample fine-grained scroll feature descriptions and Pearson’s correlations with success ratings (** indicates statistical significance at $p < .01$ level; * indicates statistical significance at $p < .05$ level).

overall scroll distance and range in pixels, following [60]. The intuition behind is to capture the searcher’s examination patterns. For example, high frequency and speed of scrolling may indicate that the searcher was “scanning” or skipping parts of the document, while a moderate range of scrolling with low speeds may indicate that the searcher was “reading”.

Aggregation of Interaction Features: Different strategies were explored in aggregating the page-level features, including computing the mean, median, minimum, and maximum of all the page views and counting the number of page views that meet some specific requirements. The four statistics are more generic treatments of aggregation, with the mean or average more frequently used, median more robust to outliers, and minimum and maximum capture the extreme behavior. The threshold-based counting aggregation is more customized towards individual features, which may result in more effective predictors when appropriately applied. This approach may require a deeper understanding of each individual feature to define meaningful thresholds. Further discussion will be provided about the different strategies in more depth in Section 5.2.4 and the comparison about their effectiveness.

Note that aggregation over an entire section might be problematic when a search session consists of multiple sub-goals, in which case the search behavior may exhibit larger variations. To address this issue, search goal boundary detection al-

gorithms (e.g., [110, 68]) can be applied to ensure aggregation is over single search goal. Alternatively, one may also consider aggregation over each search trail [142], which may also reduce the variance that comes from different types of pages. In this study, the proposed aggregation is on a single search goal as each session in the dataset consists of one single search goal (Section 5.2.2). While further improvement may be possible (e.g., through trail-level aggregation), as we will see later, this goal-based aggregation formalism already results in effective models (Section 5.2.4).

5.2.2 Data

The data set used for the experiments is the same as the one used in Chapter 4, which has hundreds of search tasks and explicit relevance judgements of visited Web pages, is from a user study conducted by researchers at the University of Massachusetts [50]. The usage data of the participants was tracked, containing the URLs the searchers visited, the fine-grained interactions with the browsed pages, such as clicks, cursor movements, and scrolling, the time-stamp of each page view and interaction was also recorded. The search tasks in the user study were designed to be representative of Web search and difficult to solve with a search engine (i.e., the answer was not easily found on a single page). As mentioned earlier, this is particularly valuable as these more difficult and long-tailed search tasks are the main challenge for the state-of-the-art search engines, and an accurate success prediction algorithm would enable search engines to evaluate and improve performance in these search tasks at a large scale.

Explicit Judgements: Each time the participants completed a search task, they were asked the degree to which their information need of the task was satisfied during the entire search session on a five point scale (“1” indicates the search session “did not satisfy the information need in any way” and “5” indicates that the search session “completely satisfied the information need”).

This self-reported explicit judgement was used as the ground truth for search suc-

cess. A total of 211 search tasks were completed and provided feedbacks from 30 participants, with 463 queries submitted and 711 pages visited.

5.2.3 Evaluation Metrics

Given a feature vector \mathbf{x} of search session, the explicit judgement of search success y , and a classification function $f(\mathbf{x})$ (where (\mathbf{x}, y) is an instance of the test dataset D), its performance on predicting the search success was evaluated using the standard measure of Accuracy (Acc), Precision (P), Recall (R) and F1-measure (F1) calculated as follows:

- **Accuracy (Acc):** The fraction of all search sessions \mathbf{x} in D that were correctly assigned the label $f(\mathbf{x})$ compared to the explicit judgement of search success y .
- **Precision (P):** Precision is computed as the fraction of the predictions $f(\mathbf{x})$ of the positive class that are correct.
- **Recall (R):** Recall is computed as the fraction of all true positive class sessions y that are correctly identified.
- **F1-measure (F1):** F1 measure, which is the harmonic mean of precision and recall, computed as $\frac{2P \cdot R}{P+R}$, provides a more complete picture of the performance especially when class distribution is skewed.

5.2.4 Results and Discussion

This section describes the experimental results and discusses the findings, by starting with analyzing the association between each individual session feature and the explicit success judgements, and then moving on to the results on success prediction, where each individual feature group and some combinations of the different feature groups are evaluated.

Feature Association with Search Success

The association between each individual session feature and the explicit success judgements were studied by computing Pearson's Correlation for each feature with statistical significant testing. The results are summarized in Tables 5.6, 5.7 and 5.8. We organize the discussion by feature groups and compare alternative session-level aggregation strategies and sources of evidence (e.g., pre-click vs. post-click) when appropriate.

Query, Click, Time As we can see from Table 5.6, the average query length is negatively correlated with the session length, though not significant, while the number of submitted queries exhibits much stronger negative correlation of -0.522, confirming the intuition that the longer the queries the user had to submit and the larger the number of queries, the more likely that the user was struggling and more likely to fail.

The number of clicks turns out to be negatively correlated with search success, which may seem counter-intuitive as click-through is typically considered as a signal of finding relevant information. One explanation is that the large number of clicks may come from the large number of queries. Indeed, divided over the number of queries, the click-through rate measures results in positive correlations, with the ratio computed over SERP views much more significant. This suggests benefits of client-side instrumentation.

As for the time measures, it turns out that the overall time span of a session exhibits the most significant correlation of -0.473, which makes sense as it characterizes the session length as the number of queries and clicks do. Somewhat surprisingly, the average dwell time on landing pages is negatively correlated with search success. One explanation is that as the task difficulty increases, users need to spend on average longer time to find the information on a page. The SAT and DSAT click-through rates, in contrast, match the intuitions and exhibit positive and negative correlations with success. However, the correlations are not significant, which may also be explained by the fact that the search tasks in the dataset are relatively

more challenging.

Cursor Movements The analysis of this feature group is given in Table 5.7. For simplicity, the discussion only focuses on analyzing the most discriminative feature of this group – maximum y coordinates and compare the different aggregation functions on this feature as well as two different sources of evidence, namely the pre-click behavior on search engine result pages (SERP) and the post-click behavior on the search trail pages.

For the SERPs, the averaging function for cursor (*avg_ymax_s*) appears to be most effective, which is substantially stronger than the deliberation time counterpart of cursor *avg_time_s* (Table 5.6). Interestingly, the minimum aggregation function (*min_ymax_s*) results in a significant positive correlation of 0.163. Note that very small maximum y coordinate suggests abandonment of search results and the minimum aggregation function of maximum y coordinate to some extent quantifies the likelihood of abandonment.

For the search trail pages, the averaging function (*avg_ymax_t*) only results in a moderate significant positive correlation of 0.179, which is stronger than its dwell time counter-part (*avg_time_s*) with a negative insignificant correlation of only -0.107 as shown in Table 5.6. Interestingly, the averaging function does not seem to be the most effective for the search trail pages. Instead, median function appears to be the most effective statistic, likely due to its robustness to outliers and larger variance. The best aggregation function for this feature turns out to be counting with meaningful thresholds. For example, the number of “SAT” page views, whose maximum y coordinate is on or above 800 pixels (*num_high_ymax_t*), exhibits a substantially stronger correlation of 0.361. The gain is significant compared to its dwell time based counterpart (*satr_t*), whose correlation is only an insignificant 0.09. Similarly, the number DSAT page views, whose maximum y coordinate below 400 pixels (*num_high_ymax_t*) exhibits stronger correlation than its dwell time counterpart (*dsatr_t*). These observations match the finding in [60] that post-click cursor features such as maximum y coordinates have stronger association with rel-

evance as compared to dwell time. However, as we have seen, careful selection of the session-level aggregation functions could have significant impacts on the predictive power of such a page-level feature.

Vertical Scrolling The analysis of this feature group is given in Table 5.8. Similar to the cursor feature group, the discussion focuses only on analyzing the most discriminative feature of this group – the scroll speed, to illustrate the differences in the various aggregation options as well as the patterns in pre-click and post-click pages. Overall, the correlations of different aggregations of sources of evidence all result in negative correlations for scroll speed, while the correlations are stronger for SERPs than trail pages. This may be explained by the larger variance in the types of trail pages – some of which may be shorter and do not require scrolling. As a result, scrolling may be more sparse and less reliable than the cursor features such as maximum y coordinate, as suggested by the overall weaker associations for search trail pages. In contrast, the correlations for scroll speeds on SERPs are much stronger with mean and maximum aggregation functions, resulting in significant correlations of -0.318 and -0.331, which is likely attributable to the relative fixed layout of search engine result pages, where user behavior tend to have smaller variance.

Predicting Search Success

This section reports the results and findings in predicting search success explicitly judged by the users using the different groups of features (Section 5.2.1). The success prediction problem is formulated as classification, and consider a search session with explicit success judgements (Section 5.2.2) equal to or larger than 4 as *successful* and *unsuccessful* otherwise. This definition of *search success* corresponds to the $Q^+R_*^+A^+V^?$ type of success according to the Query-Result-Answer-Verification (QRAV) model proposed in Section 5.1.1, where a participant was satisfied with her search session and believed that she found an correct answer, without a verification whether the submitted answer was actually accurate.

The underlying success prediction algorithm used was logistic regression, which is a widely used generalized linear model for classification [50, 64], where the predictors can take different forms, such as continuous, discrete, dichotomous, or a mix of these. The response variable, in this study, whether the search goal was *successful* or *unsuccessful*, is not a linear function of the predictors but a logit transformation of their linear combination. Logistic regression has the advantages of simple implementation, good interpretability, and time-efficiency in training at scale.

For training and testing, 10-fold cross-validation was used with 100 randomized experimental runs. The evaluated methods included the full model *FSB*, its four single feature group components: *query*, *click*, *time*, and *cursor*. Also evaluated were the *cursor_serp* and *cursor_trail* sub-groups, which are based on the *cursor* feature group computed for the SERPs and trail pages respectively. Two baselines considered are a naïve Majority Baseline (*MB*) that always guesses the majority class *successful* and a state-of-the-art baseline *QCT* model trained on the Query, Click and Time feature groups. Feature group ablation analysis was also conducted by removing the single feature groups one at a time. Finally, the three selected aggregation functions were compared, namely, average (*avg*), median (*med*), and the threshold based counting function (*thres*). The reported metrics are: Accuracy and weighted averages of Precision, Recall, and F1-measure over the two search success classes.

Single Feature Groups: The results are summarized in Table 5.9. The differences between different methods are statistically significant at .05 level under paired t-test except the difference between *cursor_serp* and *QCT* and the difference between *cursor_trail* and *click*. As we can see, the full model *FSB* significantly outperforms the two baselines as well as all the single feature groups. The *cursor* group performs the best among all the single feature groups and is the only single feature group that outperforms both of the two baselines. The remaining single feature groups outperform the *MB* baseline but underperform the *QCT* baseline. Among the two cursor subgroups, the *cursor_serp* group is significantly more predictive than the

cursor_trail group, which may be due to the more severe data sparsity and larger variance lies in the different search trail pages compared to the SERPs. Nevertheless, the combined feature group *cursor* significantly outperforms each of the two sources of evidence individually, suggesting the two are complementary.

<i>Methods</i>	<i>Acc (%)</i>	<i>P</i>	<i>R</i>	<i>F1 (% Imp.)</i>
<i>FSB</i>	77.1	77.9	77.1	77.5 (+7.6%)
<i>cursor</i>	75.3	76.0	75.3	75.6 (+5.1%)
<i>cursor_serp</i>	71.2	72.0	71.2	71.6 (-0.6%)
<i>cursor_trail</i>	65.8	65.8	65.8	65.8 (-8.6%)
<i>query</i>	68.9	68.7	68.9	68.8 (-4.4%)
<i>click</i>	66.3	66.2	66.3	66.3 (-8.0%)
<i>time</i>	70.1	70.5	70.1	70.3 (-2.4%)
<i>QCT</i>	71.8	72.3	71.8	72.0 (n/a)
<i>MB</i>	61.7	38.1	61.7	47.1 (-33.4%)

Table 5.9: Accuracy, Precision, Recall and F1-measure for the full FSB model, the single feature groups, and the QCT, MB baselines. The percentage of improvement over the QCT baseline is reported for the F1-measure.

Feature Group Ablation: The results are summarized in Table 5.10. The differences between the full model *FSB* and all the feature ablation methods are significant at .05 level under paired t-test except for *FSB-query*, suggesting all the feature groups except the *query* group contribute significantly to the full model even when other feature groups are presented. The largest decrease comes from removing the *cursor* feature group. Interestingly, even though *cursor_serp* seems to contribute more than the *cursor_trail* subgroup, the contributions from both of the two subgroups are statistically significant as supported by the fact that *FSB-cursor* significantly underperforms *FSB-cursor_serp* and *FSB-cursor_trail*.

Aggregation Functions: The results are summarized in Table 5.11. The differences between different methods are statistically significant at .05 level under paired t-test. As we can see, all the single aggregation functions underperform the full *FSB* model

<i>Methods</i>	<i>Acc (%)</i>	<i>P</i>	<i>R</i>	<i>F1 (% Diff.)</i>
<i>FSB</i>	77.1	77.9	77.1	77.5 (n/a)
<i>FSB-cursor</i>	71.8	72.3	71.8	72.0 (-7.3%)
<i>FSB-cursor_serp</i>	72.2	72.6	72.2	72.4 (-6.6%)
<i>FSB-cursor_trail</i>	73.7	74.5	73.7	74.1 (-4.4%)
<i>FSB-query</i>	77.3	78.0	77.3	77.6 (+0.2%)
<i>FSB-click</i>	74.8	75.7	74.8	75.2 (-2.9%)
<i>FSB-time</i>	73.3	73.9	73.3	73.6 (-5.0%)

Table 5.10: Accuracy, Precision, Recall and F1-measure for feature group ablation. The difference compared to the FSB full model is reported for the F1-measure.

that utilizes all the three functions. Among the individual functions, *FSB (thres)* performs the best, followed by *FSB (med)* and *FSB (avg)*, showing the importance in selecting the aggregation functions.

<i>Methods</i>	<i>Acc (%)</i>	<i>P</i>	<i>R</i>	<i>F1 (% Diff.)</i>
<i>FSB</i>	77.1	77.9	77.1	77.5 (n/a)
<i>FSB (avg)</i>	71.9	72.4	71.9	72.1 (-7.2%)
<i>FSB (med)</i>	72.8	73.4	72.8	73.1 (-6.0%)
<i>FSB (thres)</i>	73.1	73.8	73.1	73.4 (-5.5%)

Table 5.11: Accuracy, Precision, Recall and F1-measure for individual aggregation functions. The difference compared to the FSB full model is reported for the F1-measure.

5.3 Predicting Success in Mobile Search

In this section, the success prediction model is generalized to the mobile search space, where a smart phone with a touch screen was used by the searcher. Similar to the more traditional desktop setting, client-side interactions were found to be particularly predictive of search success in this study. The methodology is first introduced, followed by the results and findings.

5.3.1 Methodology

A controlled user study was conducted to collect data with known search success outcome. Ten subjects were recruited (6 male, 4 female, average age 26.2 ± 3.2). All subjects were undergraduate and graduate students or staff at the Emory University, and had some experience with Web search and smart phones. The search tasks (descriptions are given in Table 5.12) were designed for this study to be representative of common Web search tasks on mobile devices. The tasks have varying difficulty and topics, and highlight geographical intents that have been identified as a significant portion of mobile information needs [36].

The user study proceeded as follows: before the tasks began, the participants were given a tutorial of using the phone, including opening bookmarks, clicking, zooming, scrolling, and changing the physical device orientation. Next, a warm-up task was given to each participant to familiarize them with the task procedure. To begin each task, the participants were presented a task description and an initial query. For each task, the participants were instructed to first open the bookmark with the Google search engine result page (SERP) of the initial query. Once they reached the SERP, they could click the search results and/or reformulate the query if needed until the information was found or it took too long than they would spend in reality. After each task, the participants were asked a few questions, including whether they have successfully completed the task and how satisfied they were about the search experience during the task. Following the warm-up, eight search tasks were given to each participant.

Check the weather to determine how you would dress today (weather)
Find the MARTA routes and schedules from Georgia Tech to Emory on Tuesdays after 7PM (marta schedules)
Find out today’s hours of WoodPEC swimming pool
Find the address and the driving directions from Emory University to Lenox Square
Find the closest zipcar location to MathCS (zipcar locations)
Find out the names of the three Niagara Falls and read a bit about them
Find the earliest show times of “social network” after 7:00 PM today in the closest movie theater (social network)
Find out the nearby restaurants that are opened between 7:00-9:00 PM today, with approximately 15 dollars cost per person (restaurant).

Table 5.12: Task descriptions (initial queries).

To capture the client-side interactions, including the number of browsed pages, zooming, and sliding, a modified version of the Chrome browser application³ was developed for the Android phone. The events are encoded in a string and sent to the server as HTTP requests for analysis.

For each of the success and satisfaction dimensions, the prediction task was formulated as binary classification: that is, each search task was classified into two classes: successful/satisfied (user selected “very successful/ satisfied” in the post-task questionnaire) and unsuccessful/unsatisfied.

Each search task was then represented as a feature vector, with values corresponding to the server-side features and client-side features. The server-side features include the number of queries, clicks, click-through rate (CTR), average query length and task duration. The client-side features include the numbers of all browsed pages, search engine result pages (SERP), and non-SERP pages, as well as the event counts on these pages (e.g., scrolling, scaling). Additional details about the features are available on the project website referenced above.

³Available at <http://ir.mathcs.emory.edu/intent/data/sigir2011/>

The intuition was that these client-side features can provide additional insights about the search success and satisfaction level. For example, a small number of queries and clicks with moderate task duration might indicate a successful search task, as it seems that the user did not need to spend too much time and effort to complete the task; however, if the user, during this task, browsed a large number of pages and had to intensively scroll and rescale on the browsed pages, she might be actually not satisfied and not even successful, since she actually spent a lot of efforts and was keeping searching without finding the relevant information.

Various classification algorithms were experimented with, including Bayes Network (BN), Support Vector Machines (SVMs), decision trees, and others. Classification results are reported for BN only, as it performed best in this setup, even though other algorithms achieved similar performance.

5.3.2 Results and Discussion

To simulate the real scenarios in mobile search, where searchers are less likely to attempt time-consuming and complex searches, none of the assigned tasks were overly difficult. As a result, the means of success and satisfaction ratings are 3.25 (std=1.0) and 3.0 (std=1.2) respectively, on a 5-point scale (i.e., 0 represents very unsuccessful/unsatisfied, 4 represents very successful/satisfied). Interestingly, success rating has a noticeably higher mean and the correlation between these two are high but not perfect ($R=0.81$), which makes sense since users might feel unsatisfied about the experience even if they end up finding the information successfully. Also, the noticeably higher variance of the satisfaction ratings for each task suggests higher subjectivity of making satisfaction judgments.

As we can see in Figure 5.7, some tasks are better solved, while other tasks have more room for improvement. Interestingly, for easier tasks, the variance of both satisfaction and success ratings across users is smaller. In contrast, for more difficult tasks, the variance of both satisfaction and success ratings is larger, which suggests significant opportunities for personalization, or of exploiting successful “expert”

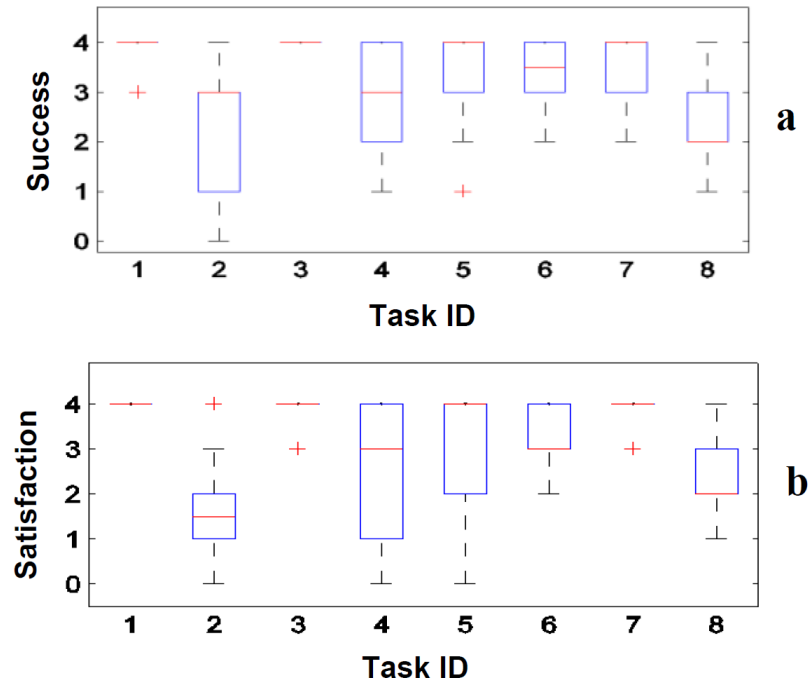


Figure 5.7: Success (a) and Satisfaction (b) by Task ID.

mobile searchers to help unsuccessful “novice” mobile search users.

Ten-fold cross validation was used: in each fold, tasks of nine participants were used for training and the remaining tasks of the left-out participant were used for testing. The average of the results across the folds are reported in Table 5.13. As we can see, the BN classifiers (the proposed models) significantly outperformed the majority baselines, exhibiting the accuracy of 79% compared to the baseline system (accuracy of 54% for predicting satisfaction and 58% for predicting success). Interestingly, using client-side features achieved better performance than using server-side features, and combining the two achieved optimal performance for predicting search satisfaction, while clients-side achieved the best performance in predicting search success.

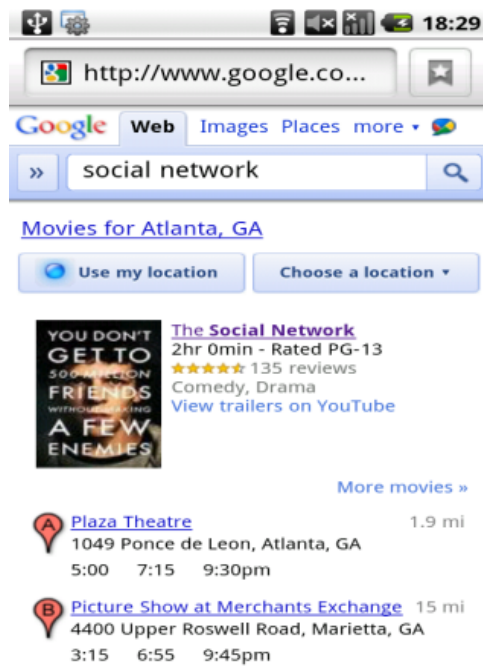
To understand the contributions of the various features, the χ^2 statistic was computed for each feature with respect to the class. The most significant features found

<i>Method</i>	<i>Both</i>		<i>Successful</i>		<i>Unsuccessful</i>	
	<i>Acc</i>	<i>F1</i>	<i>Prec.</i>	<i>Recall</i>	<i>Prec.</i>	<i>Recall</i>
Baseline	57.5	36.5	57.5	100	0	0
Server	66.3	64.2 [▲]	67.9 [▲]	78.3 [△]	63.0 [▲]	50.0 [▲]
Client	80.0 [△]	79.2 [▲]	80.0 [▲]	87.0 [△]	80.0 [▲]	70.6 [▲]
Full	78.8 [△]	78.0 [▲]	79.6 [▲]	84.8 [△]	77.4 [▲]	70.6 [▲]

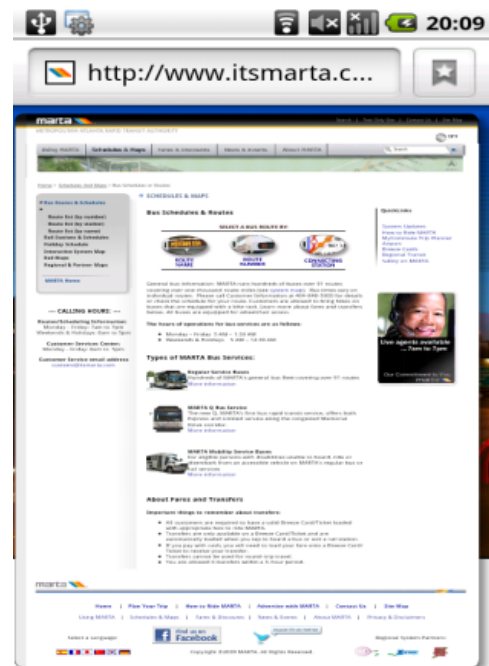
<i>Method</i>	<i>Both</i>		<i>Satisfied</i>		<i>Unsatisfied</i>	
	<i>Acc</i>	<i>F1</i>	<i>Prec.</i>	<i>Recall</i>	<i>Prec.</i>	<i>Recall</i>
Baseline	53.8	35.0	53.8	100	0	0
Server	71.5 [▲]	68.1 [▲]	66.1 [▲]	95.3	88.9 [▲]	43.2 [▲]
Client	76.3 [▲]	75.6 [▲]	74.0 [▲]	86.0 [▲]	80.0 [▲]	64.8 [▲]
Full	78.8 [▲]	78.5 ^{▲°}	78.3 ^{▲°}	83.7 ^{▲°}	79.4 [▲]	73.0 ^{▲°}

Table 5.13: Results of predicting success and satisfaction. Significance of differences is indicated between models and: Baseline: [△] $p < .05$, [▲] $p < .01$; Server: [°] $p < .05$, [•] $p < .01$.

include the number of browsed pages, number of non-SERP events, task duration, click-through rate, number of clicks and average query length. Generally, the more effort a user spent on searching, the less likely she were to be satisfied or successful – which makes sense in a mobile setting, where the screen size is small, the bandwidth is limited, and each user interaction requires effort. Two examples are given in Figure 5.8.



(a) successful



(b) unsuccessful

Figure 5.8: (a) An example page that leads to a successful task (“social networks”): the movie show times were presented on the top of the search engine result page; (b) an example page that leads to an unsuccessful task (“marta schedule”): no instant answer was presented on SERP and the official site was not optimized for the mobile setting (e.g., fonts too small)

5.4 Summary

This chapter introduced techniques on modeling interactions for evaluating search experience at the session-level based on three studies.

The first study presented a novel methodology that emphasizes realistic search tasks, yet allowing for well-defined, objective success criteria. This methodology enabled the development of a principled model of web search success that naturally encompasses previously proposed models. The success prediction models, based on CRFs and the proposed framework, were trained on the acquired behavior data and performed adequately on predicting search success on the regular search engine log, outperforming state-of-the-art baselines such as Markov models.

The second study of Web search success aimed to exploit and aggregate “low-level” behavioral signals on the session-level for success prediction. This study presented a new model for representing the searchers’ fine-grained session behavior (*FSB*) that captures not only information about the queries, clicks and time, but also fine-grained interactions both *before* and *after* clicking on a search result, such as cursor movements. The experimental results showed that these behavioral signals indeed correlate with searchers’ explicit judgements of search success, and provide additional valuable information beyond queries, clicks and the amount of time users spend on the pages in a search session. It is also found that the different sources of evidence (i.e. behavior *before* and *after* a click) carry valuable complementary information about search success and that the feature aggregation choice was crucial. In combination, these signals enable *FSB* to exhibit significant improvement of predicting search success over the state-of-the-art methods.

The third study of Web search success aimed to generalize the findings to different modalities, focusing on studying the feasibility of predicting search success and satisfaction in mobile search. The experiments showed that the proposed techniques can predict search success and satisfaction with accuracy of nearly 80%, by incorporating additional client-side interactions, such as zooming and sliding, outperforming the baseline methods based on server-side signals such as queries and

clicks.

The techniques presented in this chapter demonstrated the crucial information lies in the search interaction data in revealing the search experience and user satisfaction, allowing various important applications in diagnosing and improving search engine performance.

Chapter 6

Conclusions and Future Work

This chapter first summarizes the findings, then discusses how the proposed techniques could be integrated in a production search engine, and finally concludes by considering limitations and future research directions.

6.1 Summary of Findings

This thesis presented techniques for modeling examination through fine-grained interaction patterns, which, contextualized in a search session, can be used to improve the three fundamental and interrelated areas in Web search and provide more intelligent and tailored search experience.

The first improved area is intent inference, where the proposed techniques focus on inferring the immediate search goals of a searcher in a session and modeling the *pre-click* examination and interaction patterns on the search engine result pages (SERPs). The findings in Chapter 3 support the notion that information needs vary a great deal even for the same search query, and that users examine and interact with search results differently for distinct information needs. In particular, the findings suggest that user behavior tends to differ between exploratory and directed search goals. For example, users were found to use mouse more extensively with typically lower speeds for informational and research-oriented intents while using the mouse less actively when the search goals were more clear (e.g., for navigational or transactional intents). Also found was the influence of page content, result quality, and session context on the behavior. By incorporating evidence from these

sources, more accurate intent inference was achieved. For example, consider result quality. A user may need to use mouse more extensively to examine lower-ranked documents for poor quality results (even for a navigational intent) or may abandon the search without moving the mouse if the result quality was too poor to be worth further exploration (even for an informational intent).

The second improved area is estimating document relevance and ranking, where the proposed techniques focus on estimating the “intrinsic” relevance of a search document and modeling the *post-click* examination and interaction patterns on the landing pages and subsequently viewed pages. The findings in Chapter 4 support the hypothesis that the fine-grained interactions are associated with the various post-click viewing patterns that are indicative of the document relevance. In particular, the findings suggest that user behavior tends to exhibit “reading” patterns when visiting a relevant document and “skimming” patterns when visiting an irrelevant document, and these complex patterns are not fully captured by measuring the time user spent on the visited documents (i.e., dwell time). Also found was the variations in user behavior – by adjusting the behavioral variance across users, more accurate relevance estimation was achieved.

The third improved area is automatic evaluation of search experience, where the proposed techniques focus on inferring the search success by mining rich search interactions, including the examination patterns both “before” and “after” a click on the search result. The findings in Chapter 5 support the hypothesis that the fine-grained interactions provide additional crucial clues about search success beyond query, click and time information. The two sources of evidence, that is, the examination patterns both before and after a result click were found to provide valuable complementary information about the search experience. Among the two, the pre-click examination was found to be more predictive of search success. The benefits of modeling the fine-grained interactions were also found for the mobile setting where a touch screen is used and the interactions include gestures such as pinching, zooming and sliding. Also addressed in this thread of research was the tension between log analysis and controlled user studies, where the former lacks of accu-

rate understanding about search goals and objective measurement of search success while the latter suffers from the small scale. In particular, the proposed techniques enable reproducible large-scale remote user studies with pre-defined naturalistic search goals and objective measurements of search success with various definitions, resulting a principled framework to study search success – the collected data from this framework was successfully applied to improving search success prediction for real search sessions.

In summary, the modeling of both pre-click and post-click examination through fine-grained interaction, such as mouse movements and scrolling, was found to provide essential additional clues in improving the search engine performance in the three different areas, and adjusting for the variations across users and page types was found to be beneficial, resulting in higher accuracy in predicting the target variables.

6.2 Integrating Intent Inference, Ranking and Evaluation

As introduced briefly in Chapter 1, the three types of techniques naturally fit in different components of a search system and may be used to improve the system performance in a complementary way. In particular, the intent inference and document relevance estimation techniques can both be integrated to the ranking module while the evaluation techniques can be integrated into the monitoring and diagnosing module of the system, which also has impacts on improving the search result ranking, but from a more indirect fashion. In addition, more accurate intent inference may also help improving other modules such as query suggestion and result presentation, and real-time search experience evaluation may enable intervention for additional assistance.

Web search result documents are retrieved and ranked based on their relevance to the Web search query, which is measured or computed through a ranking function. Classic ranking functions typically measure textual similarities between documents and the given query (e.g., BM25 [116]). However, modern search engines

adapt the learning-to-rank paradigm [111, 2, 150, 26], where the ranking function is learned through features derived from the query (e.g., query length), documents (e.g., PageRank [22]) and query-document pairs (e.g., similarity computed via BM25 [116]) using machine learning algorithms.

6.2.1 Integrating Intent Inference

Techniques in improving intent inference from Chapter 3 can be used to enrich the query and query-document feature space and impact the selection of machine learning algorithms to train the ranker. In particular, the intent classes can be added as features in the query feature space and intent-document features can be computed in addition to the query-document features. For example, the ranking function may promote product review sites for a query “surface” with “research” intent feature triggered, while promote shopping sites when the “purchase” intent feature is on. Also, precision-oriented learning algorithms may be selected for navigational intent while recall-oriented learning algorithms may be selected to improve coverage and diversity for informational intent. Furthermore, better intent inference may also help improving query suggestion and result presentation. For example, if the search intent was informational, query suggestions may be provided to uncover different aspects of the topics and a more interactive and exploratory search interface may be preferred; in contrast, if the search intent was navigational, query suggestions may be provided for navigating to similar websites and a hub-like search result page highlighting different possible sites to navigate may be more appropriate.

6.2.2 Integrating Relevance Estimation

Techniques in improving relevance estimation from Chapter 4 can be used to enrich the query-document, intent-document and document feature spaces. For example, the estimated relevance of a document with respect to a query or an intent class can be added as a feature. Alternatively, the raw PCB features proposed in Chapter 4 may be incorporated directly to the learning-to-rank framework. These features

may be further aggregated at the document level, indicating the overall likelihood of relevance of the document. These techniques would probably have the most direct impacts on improving search result ranking.

6.2.3 Integrating Automatic Evaluation

While the proposed techniques in intent inference and document relevance may be more directly impacting the learning of the ranking function, the automatic evaluation techniques proposed in Chapter 5 enable the large-scale measurement and monitoring of the search engine performance, especially, when changes (such as those brought by applying the proposed intent inference and relevance estimation techniques) have been made to the search engine. The automatic evaluation techniques also enable deeper analysis and allow for more substantial improvements that cannot be achieved by improving intent inference and relevance estimation. For example, if the reason of failure of some search task is found to be the absence of documents in the index, then the improvements may be made to the crawling or indexing modules. Such cases are out of scope for the intent inference, relevance estimation techniques introduced in Chapter 3 and 4. In addition, in an online setting, the automatic evaluation techniques can be used to gauge the search experience in real-time and intervene when likely failure was detected.

6.2.4 Pre-click and Post-click Instrumentation

To enable the modeling of pre-click examination, the Javascript tracking code can be embedded in the returned SERPs without requiring additional installation, which makes the fine-grained interaction models applicable to all the general Web search engine users. To model the post-click behavior, browser plugin would be needed for general purpose Web search engines, as the landing pages are typically not owned by the search engine companies. As a result, only a fraction of users may enjoy these additional benefits brought by post-click examination modeling. However, for specialized search engines who own the search corpus (e.g., Amazon.com,

Yelp.com), the post-click modeling could also be applied without requiring additional installation.

Consider the instrumentation required for the techniques in the three different areas. For intent inference and success prediction, pre-click examination was found to be the major source of improvements, so even without post-click instrumentation, all the search engine users may be benefited from embedding the Javascript tracking code in the returned SERPs. For relevance estimation, the proposed techniques, while very effective, can only be fully applied to a fraction of users with the browser tracking plugin installed, as the post-click instrumentation is needed. However, the modeling of fine-grained interaction may still improve Web search result ranking for all users, as pre-click behavior, such as hovering, was also found to provide additional relevance signals than clicks in previous research [77, 76, 139].

In summary, the modeling of the fine-grained interaction can be applied to all general search engine users with the pre-click instrumentation for improving the different Web search applications. Also, when a browser plugin is available or the search engine owns the landing pages, the post-click instrumentation can also be enabled and additional improvements can be brought to the search engine users.

6.2.5 Infrastructures for Offline and Online Deployment

As mentioned in earlier chapters, the proposed fine-grained interaction models can be deployed in either an offline or an online setting, with different settings require different underlying infrastructures. In an offline setting, the rich interaction data can be collected and used to re-train the ranking function (Chapter 3 and 4) periodically and measure the search engine performance (Chapter 5). In an online setting, the rich interaction data need to be exploited “on-the-fly” to re-rank search results in real-time (Chapter 3 and 4) and/or detect struggling users to provide additional assistance (Chapter 5).

For an offline setting, the data collection can be implemented through instrumenting the SERPs for pre-click behavioral modeling and through instrumenting

a browser plug-in for post-click behavioral modeling, as discussed in the previous section. One possible implementation of such logging system is the EMU system [61] proposed in the thesis, which was implemented through sending asynchronous HTTP requests that encode logged usage data to avoid noticeable overhead and has been successfully deployed at Emory Libraries for years without causing any reported problems in degrading user experience. While feasible, orders-of-magnitude larger of interaction data needs to be collected to enable the fine-grained behavioral modeling described in the thesis, therefore, large-scale data processing capabilities are needed.

For an online setting, in addition to collecting the rich interaction data as in the offline setting, a real-time agent to interpret the collected data is needed. The real-time feedback interpreting agent can be deployed on the server-side, such as the underlying infrastructure of *Google Instant*¹ or *SurfCanyon*² that allows dynamic re-rendering of SERPs; or on the client-side as a browser plugin such as the UCAIR toolbar [125] that alters SERP locally. The former server-side approach allows benefiting a larger user base without requiring additional installation but increase the server-side computation burden. The latter client-side approach flips the pros and cons by distributing the centralized computation but limits its applicability due to the required installation. Other benefits of the client-side instrumentation include the availability of post-click behavior data (as mentioned above) and the increased privacy protection (user search history may be stored locally without being sent back to the server through the Web).

6.2.6 Evaluating the Deployed System

Before deployment, the effectiveness of the individual modules need to be tested offline and only deployed when the test results meet the requirements.

Different methods of offline evaluation were considered in the thesis. The most adopted method was evaluating over labels from controlled user studies with tens

¹www.google.com/instant

²<http://surfcanyon.com/>

of participants and hundreds of search tasks (either through self-report or pre-assignment). Alternatively, third party annotations were used for predicting general intent (Chapter 3), real user behavior (i.e., future ad click-through in a session) was used as proxy of “ad receptiveness” (Chapter 3), and remote user studies of hundreds of users were conducted for success prediction under the UFindIt framework, where self-reported and third party judgements were both considered (Chapter 5).

As shown in Chapter 3, 4, and 5, the accuracies of many of the proposed techniques fall in the 70-90% range based on the offline evaluation, well above the baseline models learned through a limited set of behavioral signals. While the absolute accuracy may not be as high as 95% or above (which may be deemed necessary for deployment in certain applications to avoid the risk of hurting search experience), their accuracy may actually be sufficient for production deployment, as the beaten baseline systems are very similar to the deployed state-of-the-art systems. That is, incorporating the proposed techniques to the existing systems is very likely to improve the system performance and the low absolute numbers may be due to the difficulty of the problems. Also, as the overall improvement of the proposed techniques may also result in hurting performance of the minority of the searches, risk-sensitive optimization based methods [136] may be adapted to minimize risk and ensure ranking robustness. In addition, for scenarios where high precision is needed, sacrificing recall may push the precision into the higher (more acceptable) ranges.

When deployed, the proposed intent inference and relevance estimation techniques can be further evaluated with A/B testing [16], while the proposed evaluation techniques provide novel metrics for such testing with the live search traffic.

With A/B testing, two different versions of search systems are deployed and compared. For example, one of the two systems may be the one with the proposed intent inference techniques integrated, which provides an interactive search interface if the search session is detected to be with a more exploratory intent, and the other system is the original system that provides a same search interface. After these two systems are deployed, different user behavior metrics may be computed over time and

be used to compare the user satisfaction with the two systems. Similarly, systems with and without the improved relevance estimation techniques integrated may be also compared through running A/B testing.

To measure the search experience, different metrics can be considered. A frequently used metric is the result click-through rate (CTR) [82, 112, 32], which indicates the attractiveness of the search result – the higher the CTR the better the search experience is likely to be. However, as we have seen in Chapter 4 and Chapter 5, click-through may not indicate the *intrinsic* relevance of a document; therefore, metrics, such as “success rate”, may be more meaningful, which can be computed through the success prediction algorithms introduced in Chapter 5. In addition, other metrics, such as monthly number of search sessions per user, that are indicative of long-term user engagement and loyalty, can also be used to provide a complimentary view about the search experience.

6.3 Limitations and Future Work

While the results and findings are promising, the techniques introduced in the thesis have a few limitations.

First, the additional effectiveness brought by the presented models may be limited for an unseen information need and/or unseen documents in the search log, when the corresponding interaction data is not yet available. This limitation is prevalent for any other user models that require searcher interactions as input (e.g., click models). To address this limitation, one possible solution is to utilize the interaction data from the similar information needs and or similar documents that are observed previously. Also note that, in an online setting, as the search session progresses, the interaction data accumulated from the initial search(es) and page view(s) is available to improve the search experience for later searches.

The second limitation lies in the lack of fine-grained interactions when the corresponding tracking instrumentation is not available – instrumenting result pages is at the discretion of the search engine designers but instrumenting landing pages would

require browser plug-in deployment and/or partnership with the landing page hosts, as mentioned earlier.

Third, the effectiveness of the proposed models may be limited to modalities other than personal computers (with a mouse and keyboard) and devices with a touch screen (e.g., tablet, smart phone). The fine-grained interaction modeling in the thesis mainly focuses on the mouse cursor movements and a smaller fraction of the research aims to model the gestures on a touch screen – in Chapter 5, the pinching, zooming and sliding interactions on a smart phone touch screen are shown to be useful in predicting search success. While the results are promising, further studies are needed to understand the effectiveness of the gesture-based behavioral models on other applications such as intent inference and relevance estimation. Also, with the rise of natural interfaces, such as that with the Xbox Kinect, more substantial adaption of the proposed models may be needed.

Outlined below are some interesting future research directions, including immediate extensions of the thesis that address some of the limitations listed above, and longer-term research directions that build on the presented techniques.

- **Extending supported intent inference classes:** As we have seen in Chapter 3, the rich interaction intent model was able to infer immediate information needs more accurately for various search intent dimensions, including the general intent and commercial intent. As a immediate extension, other dimensions of search intent may be considered, such as the topical intent categories (e.g., arts, computers, sports).
- **Addressing sparsity in the fine-grained interaction data:** As pointed out earlier as a limitation, the sparsity in the interaction data (e.g., unavailable for unseen queries and unseen documents) may limit the applicability of the presented models. To address this issue, techniques need to be developed to link unseen information needs and documents to the observed ones, possibly through modeling textual similarities and co-occurrences (e.g., in the same search results or sessions).

- **Adapting to different modalities:** While some initial success has been achieved in Chapter 5, further extensions in this direction include analyzing the fine-grained behavioral patterns with a touch screen in more depth, developing prediction models for other applications (e.g., ranking), modeling the behavior on customized mobile apps (other than Web searching in the mobile browser), and incorporating mobile context (e.g., location, personal information). Also, further extensions can be done with natural interfaces such as that of the Xbox Kinect.
- **Modeling reading behavior and sub-document level implicit feedback:** As discussed in Chapter 2, some recent success has been achieved in predicting gaze positions from cursor movements and modeling gaze movements for sub-document level implicit feedback. To extend this thread of research, reading behavior may be extracted on top of the predicted gaze positions, which can be used to obtain implicit relevance feedback at the level of the sub-documents.
- **Developing online interactive search systems:** An online interactive system can be developed by integrating the presented techniques as described earlier, which infers the immediate information needs, re-ranks search results and provides other search assistance while evaluating and monitoring the search experience in real-time.

In summary, the thesis has shown that modeling the rich search interactions enable improved understanding of the searcher information needs, more accurate estimation of document relevance, and better evaluation of search engine performance at the session-level, using machine learning and data mining techniques. The presented techniques and ideas allow for more effective and intelligent search systems, providing building blocks for more extensive research in the area of information retrieval and user behavior modeling.

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, 2011.
- [2] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 19–26, 2006.
- [3] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 3–10, 2006.
- [4] E. Agichtein and Z. Zheng. Identifying "best bet" web search results by mining past user behavior. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 902–908, 2006.
- [5] J.-w. Ahn, P. Brusilovsky, D. He, J. Grady, and Q. Li. Personalized web exploration with task models. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 1–10, 2008.

- [6] J. Arguello, F. Diaz, and J. Callan. Learning to aggregate vertical results into web search results. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 201–210, 2011.
- [7] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 315–322, 2009.
- [8] A. Ashkan and C. L. Clarke. Characterizing commercial intent. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 67–76, 2009.
- [9] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo. Classifying and characterizing query intent. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 578–586, 2009.
- [10] 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, 2010.
- [11] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. The intention behind web queries. In *Proceedings of the 13th international conference on String Processing and Information Retrieval*, SPIRE'06, pages 98–109, 2006.
- [12] B. T. Bartell, G. W. Cottrell, and R. K. Belew. Automatic combination of multiple ranked retrieval systems. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 173–181, 1994.

- [13] H. Becker, C. Meek, and D. M. Chickering. Modeling contextual factors of click rates. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2, AAAI'07*, pages 1310–1315, 2007.
- [14] N. J. Belkin. User modeling in information retrieval. *Tutorial at UM97*, 1997.
- [15] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. The effect multiple query representations on information retrieval system performance. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '93*, pages 339–346, 1993.
- [16] P. N. Bennett, B. Carterette, O. Chapelle, and T. Joachims. Beyond binary relevance: preferences, diversity, and set-level judgments. *SIGIR Forum*, 42(2):53–58, Nov. 2008.
- [17] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '12*, pages 185–194, 2012.
- [18] J. Bian, X. Li, F. Li, Z. Zheng, and H. Zha. Ranking specialization for web search: a divide-and-conquer approach by using topical ranksvm. In *Proceedings of the 19th international conference on World wide web, WWW '10*, pages 131–140, 2010.
- [19] J. Bian, T.-Y. Liu, T. Qin, and H. Zha. Ranking with query-dependent loss for web search. In *Proceedings of the third ACM international conference on Web search and data mining, WSDM '10*, pages 141–150, 2010.
- [20] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, Aug. 1996.

- [21] D. J. Brenes, D. Gayo-Avello, and K. Pérez-González. Survey and evaluation of query intent detection methods. In *Proceedings of the 2009 workshop on Web Search Click Data*, WSCD '09, pages 1–7, 2009.
- [22] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.
- [23] A. Broder. A taxonomy of web search. *SIGIR Forum*, 2002.
- [24] A. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras. To swing or not to swing: learning when (not) to advertise. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1003–1012, 2008.
- [25] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using web knowledge. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 231–238, 2007.
- [26] C. J. C. Burges, K. M. Svore, P. N. Bennett, A. Pastusiak, and Q. Wu. Learning to rank using an ensemble of lambda-gradient models. *Journal of Machine Learning Research - Proceedings Track*, pages 25–35, 2011.
- [27] 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, 2008.
- [28] 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, 2009.

- [29] G. Buscher, R. W. White, S. Dumais, and J. Huang. Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 373–382, 2012.
- [30] H. Cao, D. H. Hu, D. Shen, D. Jiang, J.-T. Sun, E. Chen, and Q. Yang. Context-aware query classification. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 3–10, 2009.
- [31] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. What makes a query difficult? In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 390–397, 2006.
- [32] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. Inf. Syst.*, 30(1):6:1–6:41, Mar. 2012.
- [33] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 1–10, 2009.
- [34] D. Chen, W. Chen, H. Wang, Z. Chen, and Q. Yang. Beyond ten blue links: enabling user click modeling in federated web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 463–472, 2012.
- [35] W. Chen, Z. Ji, S. Shen, and Q. Yang. A whole page click model to better interpret search engine click data. In *AAAI'11*, pages –1–1, 2011.

- [36] K. Church and B. Smyth. Understanding the intent behind mobile information needs. In *Proceedings of the 14th international conference on Intelligent user interfaces*, IUI '09, pages 247–256, 2009.
- [37] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 227–236, 2008.
- [38] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*, IUI '01, pages 33–40, 2001.
- [39] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 87–94, 2008.
- [40] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 299–306, 2002.
- [41] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 407–416, 2007.
- [42] H. K. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li. Detecting online commercial intention (oci). In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 829–837, 2006.
- [43] F. Diaz. Integration of news content into web results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 182–191, 2009.

- [44] A. Diriye, R. W. White, G. Buscher, and S. T. Dumais. Leaving so soon? understanding and predicting web search abandonment rationales. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, 2012.
- [45] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web, WWW '07*, pages 581–590, 2007.
- [46] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, SIGIR '03*, pages 72–79, 2003.
- [47] G. Dupret and C. Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *Proceedings of the third ACM international conference on Web search and data mining, WSDM '10*, pages 181–190, 2010.
- [48] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08*, pages 331–338, 2008.
- [49] H. A. Feild, J. Allan, and J. Glatt. Crowdlogging: distributed, private, and anonymous search logging. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11*, pages 375–384, 2011.
- [50] 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, 2010.

- [51] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2), 2005.
- [52] G. Golovchinsky, M. N. Price, and B. N. Schilit. From reading to retrieval: freeform ink annotations as queries. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 19–25, 1999.
- [53] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y.-M. Wang, and C. Faloutsos. Click chain model in web search. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 11–20, 2009.
- [54] F. Guo, C. Liu, and Y. M. Wang. Efficient multiple-click models in web search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 124–131, 2009.
- [55] Q. Guo and E. Agichtein. Exploring client-side instrumentation for personalized search intent inference. In *Proc. of ITWP*, 2008.
- [56] Q. Guo and E. Agichtein. Exploring mouse movements for inferring query intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 707–708, 2008.
- [57] Q. Guo and E. Agichtein. Exploring searcher interactions for distinguishing types of commercial intent. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1107–1108, 2010.
- [58] 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, 2010.

- [59] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 3601–3606, 2010.
- [60] 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, 2012.
- [61] Q. Guo, R. P. Kelly, S. Deemer, A. Murphy, J. A. Smith, and E. Agichtein. Emu: the emory user behavior data management system for automatic library search evaluation. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '09, pages 389–390, 2009.
- [62] 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, 2012.
- [63] Q. Guo, R. W. White, S. T. Dumais, J. Wang, and B. Anderson. Predicting query performance using query, result, and user interaction features. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 198–201. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2010.
- [64] 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*, SIGIR '11, pages 335–344, 2011.
- [65] Q. Guo, S. Yuan, and E. Agichtein. Detecting success in mobile search from interaction. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, pages 1229–1230, 2011.

- [66] K. Gyllstrom and C. Soules. Seeing is retrieving: building information context from what the user sees. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '08, pages 189–198, 2008.
- [67] A. Hassan. A semi-supervised approach to modeling web search satisfaction. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 275–284, 2012.
- [68] 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, 2010.
- [69] A. Hassan, Y. Song, and L.-w. He. A task level user satisfaction metric and its application on improving relevance estimation. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, CIKM '11, 2011.
- [70] Y. He and K. Wang. Inferring search behaviors using partially observable markov model with duration (pomd). In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 415–424, 2011.
- [71] A. P. Heath and R. W. White. Defection detection: predicting search engine switching. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 1173–1174, 2008.
- [72] E. V. Hoa, E. M. Voorhees, and H. T. Dang. Overview of the trec 2005 question answering track. In *In TREC 2005*, 1999.
- [73] 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, 2011.

- [74] J. Huang, T. Lin, and R. W. White. No search result left behind: branching behavior with browser tabs. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 203–212, 2012.
- [75] J. Huang, R. White, and G. Buscher. User see, user point: gaze and cursor alignment in web search. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, pages 1341–1350, 2012.
- [76] J. Huang, R. W. White, G. Buscher, and K. Wang. Improving searcher models using mouse cursor activity. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 195–204, 2012.
- [77] 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, 2011.
- [78] B. J. Jansen, D. L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1149–1150, 2007.
- [79] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3):1251–1266, May 2008.
- [80] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 41–48, 2000.

- [81] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web, WWW '03*, pages 271–279, 2003.
- [82] 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, 2002.
- [83] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2), 2007.
- [84] 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, 2008.
- [85] Y.-F. Juan and C.-C. Chang. An analysis of search engine switching behavior using click streams. In *Proceedings of the First international conference on Internet and Network Economics, WINE'05*, pages 806–815, 2005.
- [86] M. Kamvar, M. Kellar, R. Patel, and Y. Xu. Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices. In *Proceedings of the 18th international conference on World wide web, WWW '09*, pages 801–810, 2009.
- [87] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01*, pages 408–409, 2001.
- [88] D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. In *Proceedings of the 27th annual international ACM SIGIR*

conference on Research and development in information retrieval, SIGIR '04, pages 377–384, 2004.

- [89] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, Sept. 2003.
- [90] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
- [91] D. Lagun and E. Agichtein. Viewser: enabling large-scale remote user studies of web search examination and interaction. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11*, pages 365–374, 2011.
- [92] D. Lagun and E. Agichtein. Re-examining search result snippet examination time for relevance estimation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '12*, pages 1141–1142, 2012.
- [93] S. Laxman, V. Tankasali, and R. W. White. Stream prediction using a generative model based on frequent episodes in event sequences. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08*, pages 453–461, 2008.
- [94] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web, WWW '05*, pages 391–400, 2005.
- [95] J. Leskovec, S. Dumais, and E. Horvitz. Web projections: learning from contextual subgraphs of the web. In *Proceedings of the 16th international conference on World Wide Web, WWW '07*, pages 471–480, 2007.

- [96] 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, 2009.
- [97] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 339–346, 2008.
- [98] T. Lin, P. Pantel, M. Gamon, A. Kannan, and A. Fuxman. Active objects: actions for entity-centric search. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 589–598, 2012.
- [99] C. Liu, R. W. White, and S. Dumais. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 379–386, 2010.
- [100] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 558–565, 2002.
- [101] Q. Liu, E. Agichtein, G. Dror, E. Gabrilovich, Y. Maarek, D. Pelleg, and I. Szpektor. Predicting web searcher satisfaction with existing community-based answers. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 415–424, 2011.
- [102] 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, 2012.

- [103] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08, pages 483–490, 2008.*
- [104] N. Matthijs and F. Radlinski. Personalizing web search using long term browsing history. In *Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, pages 25–34, 2011.*
- [105] M. Melucci and R. W. White. Discovering hidden contextual factors for implicit feedback. In *CIR'07, pages –1–1, 2007.*
- [106] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '94, pages 272–281, New York, NY, USA, 1994. Springer-Verlag New York, Inc.*
- [107] T. Mukhopadhyay, U. Rajan, and R. Telang. Competition between internet search engines. In *Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 8 - Volume 8, pages 80216.1–, Washington, DC, USA, 2004. IEEE Computer Society.*
- [108] V. Navalpakkam and E. Churchill. Mouse tracking: measuring and predicting users' experience of web-based content. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, CHI '12, pages 2963–2972, New York, NY, USA, 2012. ACM.*
- [109] J. G. Phillips and T. J. Triggs. Characteristics of cursor trajectories controlled by the computer mouse. *Ergonomics*, 2001.

- [110] B. Piwowarski, G. Dupret, and R. Jones. Mining user web search activity with layered bayesian networks or how to capture a click in its context. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 162–171, 2009.
- [111] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 239–248, 2005.
- [112] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 43–52, 2008.
- [113] F. Radlinski, M. Szummer, and N. Craswell. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1171–1172, 2010.
- [114] R. Rafter and B. Smyth. Passive profiling from server logs in an online recruitment environment. In *Proc. of ITWP*, 2001.
- [115] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 521–530, 2007.
- [116] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.
- [117] K. Rodden and X. Fu. Exploring how mouse movements relate to eye movements on web search results pages. In *Web Information Seeking and Interaction Workshop*, 2006.

- [118] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 2997–3002, 2008.
- [119] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 13–19, 2004.
- [120] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1325–1334, 2009.
- [121] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Query enrichment for web-query classification. *ACM Trans. Inf. Syst.*, 24(3):320–352, July 2006.
- [122] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. Building bridges for web query classification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 131–138, 2006.
- [123] S. Shen, B. Hu, W. Chen, and Q. Yang. Personalized click model through collaborative filtering. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 323–332, 2012.
- [124] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 824–831, 2005.
- [125] X. Shen, B. Tan, and C. Zhai. Ucair: a personalized search toolbar. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 681–681, 2005.

- [126] X. Shen, B. Tan, and C. Zhai. Privacy protection in personalized search. *SIGIR Forum*, 41(1):4–17, June 2007.
- [127] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais, and B. Billerbeck. Probabilistic models for personalizing web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 433–442, 2012.
- [128] R. Srikant, S. Basu, N. Wang, and D. Pregibon. User browsing models: relevance versus examination. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 223–232, Washington, DC, 2010.
- [129] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 675–684, 2004.
- [130] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 718–723, 2006.
- [131] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 449–456, 2005.
- [132] J. Teevan, S. T. Dumais, and D. J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 163–170, 2008.

- [133] S. K. Tyler and J. Teevan. Large scale query log analysis of re-finding. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 191–200, 2010.
- [134] K. Wang, N. Gloy, and X. Li. Inferring search behaviors using partially observable markov (pom) model. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 211–220, 2010.
- [135] K. Wang, T. Walker, and Z. Zheng. Pskip: estimating relevance ranking quality from web search clickthrough data. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1355–1364, 2009.
- [136] L. Wang, P. N. Bennett, and K. Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 761–770, 2012.
- [137] Y. Wang and E. Agichtein. Query ambiguity revisited: clickthrough measures for distinguishing informational and ambiguous queries. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 361–364, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [138] R. W. White, P. N. Bennett, and S. T. Dumais. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1009–1018, 2010.
- [139] R. W. White and G. Buscher. Text selections as implicit relevance feedback. In *Proceedings of the 35th international ACM SIGIR conference on*

Research and development in information retrieval, SIGIR '12, pages 1151–1152, 2012.

- [140] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 21–30, 2007.
- [141] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 87–96, 2009.
- [142] R. W. White and J. Huang. Assessing the scenic route: measuring the value of search trails in web logs. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 587–594, 2010.
- [143] R. W. White, A. Kapoor, and S. T. Dumais. Modeling long-term search engine usage. In *Proceedings of the 18th international conference on User Modeling, Adaptation, and Personalization*, UMAP'10, pages 28–39, 2010.
- [144] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 297–306, 2006.
- [145] R. W. White and D. Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 255–262, 2007.
- [146] R. W. White, M. Richardson, M. Bilenko, and A. P. Heath. Enhancing web search by promoting multiple search engine use. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 43–50, 2008.

- [147] B. Xiang, D. Jiang, J. Pei, X. Sun, E. Chen, and H. Li. Context-aware ranking in web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 451–458, 2010.
- [148] L. Xiong and E. Agichtein. Towards Privacy-Preserving Query Log Publishing. In E. Amitay, C. G. Murray, and J. Teevan, editors, *Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007)*, May 2007.
- [149] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 512–519, 2005.
- [150] Z. Zheng, K. Chen, G. Sun, and H. Zha. A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 287–294, 2007.
- [151] F. Zhong, D. Wang, G. Wang, W. Chen, Y. Zhang, Z. Chen, and H. Wang. Incorporating post-click behaviors into a click model. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 355–362, 2010.
- [152] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 543–550, 2007.