Proceedings of the First Workshop on

Information Retrieval in Advertising

IRA 2008

July 24th, 2008
Singapore
Co-located with ACM SIGIR 2008

Ewa Dominowska, Eugene Agichtein, Evgeniy Gabrilovich,
and James G. Shanahan (Editors)

Preface

Advertising is a multi-billion dollar industry that has become a significant component of the Web browsing experience. Online advertising systems incorporate many information retrieval techniques by combining content analysis, user interaction models, and commercial constraints. Advances in online advertising have come from integrating several core research areas: information retrieval, data mining, machine learning, and user modeling.

The workshop will serve as an open forum for discussion of new ideas and current research related to information retrieval topics relevant to online advertising. The outcome will be a set of full and short papers covering a variety of topics. The short paper format will allow researchers new to the area to actively participate and explore novel themes. It will also enable researchers without access to extensive empirical data to propose ideas and experiments. We also expect the workshop to help develop a community of researchers interested in this area, and yield future collaboration and exchanges.

Despite its commercial significance, advertising is a rather young field of research. This workshop will help the emerging research community better organize and develop a common perspective. The workshop will serve as a forum for researchers and industry participants to exchange latest ideas and best practices while encouraging future breakthroughs. It will also aid in fostering collaboration between industry and academia.

Table of Contents

IRA 2005 Organization

WORKSHOP CO-CHAIRS

Ewa Dominowska (Microsoft)

Ewa Dominowska is a Technical Assistant for the Search, Portal and Ads Platform. Prior to that, she was a Senior Program Manager at Microsoft adCenter, where she created and led the Open Platform Research and Development Feature Team. Ewa has been working in the field of online advertising for the past three years, with a focus on ad ranking, ad targeting, and content analysis. She designed and architected adCenter's contextual advertising system. Ewa is the organizer and PC chair for the Internet Economics track of the Beyond Search Request for Proposals and the co-organizer and PC chair for TROA 2008. She earned her MEng and BS degrees from the Massachusetts Institute of Technology in 2002. Her research focused on natural language processing and machine learning. Ewa has authored over 25 patent applications in the area of online advertising.

Eugene Agichtein (Emory University)

Eugene Agichtein is an Assistant Professor in the Mathematics and Computer Science Department at Emory University, where he directs the Intelligent Information Access laboratory. Previously, Eugene was a Postdoctoral Researcher in the Text Mining, Search, and Navigation group at Microsoft Research, working on text and web mining for information retrieval. He received a Ph.D. in Computer Science from Columbia University in 2005, and a B.S. in Engineering from The Cooper Union in 1998. Eugene is a recipient of the "Best Student Paper" award at the ICDE 2003 conference, and the "Best Paper Award" at the SIGMOD 2006 conference. Dr. Agichtein works on information extraction, text mining, and user behavior modeling for more effective information access, management, and discovery from large text datasets such as the web, social media, and scientific literature.

Evgeniy Gabrilovich (Yahoo! Research)

Evgeniy Gabrilovich is a Senior Research Scientist at Yahoo! Research. His research interests include information retrieval, machine learning, and computational linguistics. He serves on the program committees of ACL-08:HLT, AAAI '08, JCDL '08, CIKM '08 and WWW '08, and in the past he served on the program committees of AAAI, EMNLP-CoNLL, COLING-ACL, served as a mentor at SIGIR '07, as well as reviewed papers for ACM TOIT, IP&M, JNLE, CACM, AAAI, AAMAS, WWW and CIKM. Evgeniy earned his MSc and PhD degrees in Computer Science from the Technion - Israel Institute of Technology.

James Shanahan (Independent Consultant)

Jimi has spent the last 20 years developing and researching cutting-edge information management systems to harness information retrieval, linguistics and machine learning. Prior to being an independent consultant, Jimi was Chief Scientist (and member of executive team) at Turn Inc. where he focused on the development and deployment of an online ad targeting system (CPA/CPC/CPM-based) in a principled and measured way that leveraged advanced statistical and machine learning techniques;These responsibilities included leveraging the entire reservoir of data assets in order to develop methods for identifying key optimizations, deploying relevant analytical tools and improving the user experience. Prior to joining Turn, Jimi was Principal Research Scientist at Clairvoyance Corporation where he led the Knowledge Discovery from Text Group. Before that he was a Research Scientist at Xerox Research Center Europe (XRCE), where, as a member of the Co-ordination Technologies Group, he developed Document Souls, a patented document-centric approach to information access. In the early 90s, he worked on the AI Team within the Mitsubishi Group in Tokyo. He has published six books and over 50 research publications in the area of machine learning and information processing. Jimi is General Chair for CIKM 2008. Jimi received his Ph.D. in engineering mathematics from the University of Bristol, U. K. and holds a bachelor of science degree in computer science from the University of Limerick, Ireland. He is a Marie Curie fellow and member of IEEE and ACM.

PROGRAM COMMITTEE MEMBERS

Misha Bilenko (Microsoft Research)
Andrei Broder (Yahoo! Research)
Max Chickering (Microsoft)
Brian Davison (Lehigh University)
Susan Dumais (Microsoft Research)
Ayman Farahat (AdMob.com)
Anindya Ghose (New York University)
Jason Hartline (Northwestern University)
Monika Henzinger (Google)
Vanja Josifovski (Yahoo! Research)
Oren Kurland (Technion)
Ping Li (Cornell)
Chris Meek (Microsoft)
Donald Metzler (Yahoo! Research)
Vanessa Murdock (Yahoo! Research Barcelona)
Jan Pedersen (Yahoo! Research)
Yan Qu (Advertising.com)
Matt Richardson (Microsoft Research)
Arun Surendran (Microsoft AdLabs)
Roumen Vragov (CUNY)
Ryen White (Microsoft Research)
Yi Zhang (UC Santa Cruz)

INVITED TALKS

Collaborative Engineering of the Knowledge Web, **Tarek Najm and A.C. Surendran,** *Microsoft Corporation*

Introduction to Mobile Advertising, **Andrei Broder,** *Yahoo! Research*

Contrasting and Blending Marketing and Computer Science Approaches to Online Advertising, **Ram Akella,** *University of California, Santa Cruz*

WORKSHOP ROGRAM

9:00-9:10 Opening Remarks

9:10-10:00 Keynote 1: *Collaborative Engineering of the Knowledge Web*, Tarek Najm and A.C. Surendran, *Microsoft*

10:00-10:30 Break

10:30-12:00 Research Paper Presentations

- ***Evaluating Vector-Space and Probabilistic Models for Query to Ad Matching,*** Hema Raghavan and Rukmini Iyer
- ***Characterizing Query Intent From Ad Clickthrough Data.*** Azin Ashkan, Charles Clarke, Eugene Agichtein and Qi Guo.
- ***Domain-Specific Query Augmentation using Folksonomy Tags: the Case of Contextual Advertising.*** Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich and Bo Pang.
- ***Understanding Abandoned Ads: Towards Personalized Commercial Intent Inference via Mouse Movement Analysis.*** Qi Guo, Eugene Agichtein, Charles Clarke and Azin Ashkan.

12:00-1:00 Lunch

1:00-1:45 Keynote 2: *Introduction to Mobile Advertising,* Andrei Broder, *Yahoo! Research*

1:45-2:30 Keynote 3: *Contrasting and Blending Marketing and Computer Science Approaches to Online Advertising,* Ram Akella, *University of California, Santa Cruz*

2:30-3:00 Break

3:00-3:30 Discussion: Enabling Research in the Area of Online Advertising
***Microsoft adCenter Challenge: Data to the People*, Misha Bilenko,** *Microsoft Research*

3:30-4:00 Open Discussion and Networking

Invited Talk 1

Collaborative Engineering of the Knowledge Web

Tarek Najm and A.C. Surendran
Microsoft Corporation

Abstract

Today's killer applications on the internet – search, email, social networks - are all constrained by their underlying data stores which have limited understanding of meaning in content, its social context, or user actions and intentions. There is an opportunity to re-architect this information into an intelligent platform of connected data. This platform will create structured knowledge bringing together the content graph, the social graph and user activity over time.

In this talk, we will present the vision of a web transformed using such an intelligent, connected data store. We envision that this platform can only be created by massive web ecosystem – developers using the integrated knowledge to create new compelling applications, which will bring more users to the system. As more data is added upstream, we imagine that this platform will add more intelligence downstream creating a virtuous cycle, thus building the ultimate collaborative knowledge web.

We will discuss how this platform can be a disruptive powerhouse, spurring the creation of a series of new killer applications, including the next-generation search and advertising engines.

Bio

Tarek Najm is a Technical Fellow working in Microsoft's Advertising and Business Intelligence Systems. Najm's group is responsible for the vision, architecture and direction of Microsoft's next generation adCenter advertising platform.

Tarek has been with Microsoft for 10 years during which he held jobs including: Senior Architect of large scale systems, Billing Group Manager, Director of Business Intelligence, Product Unit Manager of adCenter, General manager of all advertising systems including Display Advertising, Paid Search, Content Advertising and eMail Advertising platforms. Tarek is also Co-Founder and Co-Executive Sponsor of Microsoft AdCenter Incubation Labs, and AdLab.

Tarek holds degrees in Computer Science, Mathematics, and Statistical Analysis. He has over 17 years of experience in Software Engineering and Systems Architecture Design. He is an expert in the design of large-scale Transactional Processing systems, Data Modeling, Distributed Systems, Massive Parallel Processing, and VLDB Data Warehousing systems

Dr. A.C. Surendran is a Senior Applied Researcher at Microsoft adCenter Labs where he is responsible for solving advanced problems in online digital advertisement, especially in targeting and content analysis, using machine learning and data mining. Previously, he was a researcher, first at Bell Labs and then at Microsoft Research.

Dr. Surendran is co-organizing ADKDD – the KDD Workshop on data mining for advertising in 2008 (as he did in 2007) and he is the treasurer for KDD 2008. He has also served on the program committee of major workshops on online advertising like WWW TROA 2008 and SIGIR IRA 2008. He has published in a variety of topics including online advertising, signal processing, speech & speaker recognition and text processing, and had filed over 15 patents.

Dr. Surendran has a PhD in Electrical Engineering from Rutgers University. You can find more information about him on his home page: http://research.microsoft.com/users/acsuren

Invited Talk 2

Introduction to Mobile Advertising

Andrei Broder
Yahoo! Research

Abstract

[TBD]

Bio

Andrei Broder is a Yahoo! Research Fellow and Vice President for Computational Advertising. Previously he was an IBM Distinguished Engineer and the CTO of the Institute for Search and Text Analysis in IBM Research.

From 1999 until early 2002 he was Vice President for Research and Chief Scientist at the AltaVista Company. Before that he has been a senior member of the research staff at Compaq's Systems Research Center in Palo Alto.

He was graduated Summa cum Laude from Technion, the Israeli Institute of Technology, and obtained his M.Sc. and Ph.D. in Computer Science at Stanford University under Don Knuth.

Broder is co-winner of the Best Paper award at WWW6 (for his work on duplicate elimination of web pages) and at WWW9 (for his work on mapping the web).

He has published more than seventy papers and was awarded twenty patents. He is an IEEE fellow and served as chair of the IEEE Technical Committee on Mathematical Foundations of Computing.

Invited Talk 3

Contrasting and Blending Marketing and Computer Science Approaches to Online Adertising

Ram Akella
University of California, Santa Cruz

Abstract

[TBD]

Bio

Prof. Akella's current research interests include in knowledge management, process learning, quality, fab economic models, cost of ownership and financial justification for IT Management and equipment, production planning and control, and bio-informatics. His other interests are Enterprise Systems, IT and Software, Financial Engineering, High Tech and E-Business, and range from cell and factory level design and control to enterprise-wide coordination and logistics, including supply chain management and contracts, financial engineering and investment, demand management, E-Commerce and E-Business exchanges, and product and process portfolios for risk management and design capacity management.

Prof. Ram Akella is currently Professor and Director of Information Systems and Technology Management, at the University of California at Silicon Valley Center/ Santa Cruz, and was Founding Director, SUNY Center for Excellence in Global Enterprise Management. At Stanford, and also at Berkeley, and Carnegie Mellon, as a faculty member and Director, Prof. Akella has led major multi-million dollar interdisciplinary team efforts in High Tech and Semiconductors. He joined the faculty at Carnegie Mellon University in 1985 as an Associate Professor in the Graduate School of Industrial Administration and the School of Computer Science (Robotics Institute) at Carnegie Mellon University in Pittsburgh, Pennsylvania. His research and teaching at Stanford University have been in High Tech, IT, Knowledge Management, Semiconductors, Cost Competitiveness, Product Life Cycle Management, Supply Chain Management, Financial Engineering and Investment, Business Process Optimization and E-Business. At the University of California at Berkeley he has taught in Industrial Engineering and Operations Research, and conducted research on Semiconductor Process Learning. He has also been a Postdoctoral visitor at Harvard University and worked at M.I.T., Cambridge (EECS/LIDS and the Leaders for Manufacturing Program). Professor Akella completed his B.S. in Electronics at I.I.T. Madras, and a Ph.D. in Systems/EECS at I.I.Sc. Bangalore. His doctoral students have gone on to teach at major schools such as Northwestern, Michigan (Ann Arbor), NYU, USC, Dartmouth, and the London Business School, and to work at major corporations such as IBM, KLA-Tencor, TSMC, ABN/AMRO, and BCG, while masters students have gone on to become Vice Presidents of major corporations, such as AT Kearney.

He has received several awards, such as the IBM Faculty Award, the AMD Research Award, and the KLA Award, has been cited in Marquis' Who's Who, and has interacted extensively with industries, including those corporations such as AMD, TI, IBM, Digital, Hyundai, LSI Logic, HP, AT&T, KLA, Applied Materials, SRC, American Axle, Delphi Automotive, General Motors, and Rich Food Products, along with various Japanese and European companies. In leading the STPI-Stanford/SUNY study on IT Outsourcing, he has interacted with many of the US software companies and their Indian suppliers. He has also lectured extensively by invitation in Europe and the Pacific Rim including Japan, Taiwan, Korea, and Singapore. He is on the Technical Advisory Council of Yield Dynamics, and boards including E-Soft. He enjoys helping companies grow and become more profitable, and is delighted when executives give him stock as a token of appreciation.

Professor Akella has served as an Associate Editor for Operations Research and IEEE Transactions on Semiconductor Manufacturing, and has been on the Editorial Board of Technology and Operations Review. He has also served as Guest Editor for IEEE Robotics and Automation: Special Issue on Manufacturing Systems.

Evaluating Vector-Space and Probabilistic Models for Query to Ad Matching

Hema Raghavan
Yahoo! Inc
Great America Parkay
Santa Clara, CA, 95054
raghavan@yahoo-inc.com

Rukmini Iyer
Yahoo! Inc
Great America Parkay
Santa Clara, CA, 95054
riyer@yahoo-inc.com

ABSTRACT

In this work, we evaluate variants of several information retrieval models from the classic BM25 model to Language Modeling approaches for retrieving relevant textual advertisements for Sponsored Search. Within the language modeling framework, we explore implicit query expansion via translation tables derived from multiple sources and propose a novel method for directly estimating the probability that an advertisement is clicked for a given query. We also investigate explicit query expansion using regular web search results for sponsored search using the vector space framework. We find that web-based expansions result in significant improvement in Mean Average Precision.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Information Retrieval, Vector Space Models, Language Models, Translation Models

1. INTRODUCTION

The primary source of revenue for major search engines is through advertising. The online ad spend of advertisers has been growing significantly over the past few years [1]. In the popular auction model used by search engines, an advertisers bids on a keyword such as *used cars*. When a user types in the query *used cars*, this particular advertiser's ads will be among the candidate set of ads that can be displayed alongside the search results. If an advertiser opts in for "advance match", his ad may also be shown for queries such as *cheap cars* or *old cars*. Once the ads are part of the candidate set, they are then ranked by a product of relevance of the ad to the query and the bid. If a user clicks on the ad, the advertiser pays the search engine for the click. The cost paid is determined by the bid and relevance of the ad shown immediately below the given ad, following the framework of a generalized second price auction [11]. Since the advertiser pays only when a user clicks on the ad, this monetization model is often called "pay-per-click" marketing. If the relevance score is assumed to be a measure of the probability that an ad is clicked for a given query, or the estimated click-through-rate of the ad for that query, ranking by bid $\times$ relevance is optimal for maximizing revenue.

Although sponsored search is a relatively new area of research, matching an ad to a query poses problems quite similar to those addressed by the information retrieval and web search community for many years. However, there are some key differences between web search and sponsored search. One of the primary differences is that the collection of web documents is significantly larger than the advertiser database, and retrieving candidate ads for tail queries using advanced match is a very important area of research for sponsored search. Another big difference from web search is the fact that the user model is different. Many queries do not have commercial intent; displaying ads on a query like "formula for mutual information" may hurt user experience and occupy real-estate on the search results page in a spot where a more relevant web-search result might exist. Therefore, in sponsored search, we prefer not to show any ads when the estimate of click-through-rate and/or relevance of the ad is low. Using the same user experience argument, for a navigational query [4] like "bestbuy.com", we would rather show only the most relevant exact ad if that ad existed in the advertiser database. We refer the reader to the study of Jansen and Resnick [15] for further details on user perceptions of sponsored search. In web-search, determining how many candidates to retrieve and display is not as much of an issue as the generally accepted user model is one where users read the page in sequence and exit the search session when their information need is satisfied. While all of the above problems are interesting areas of research, we restrict ourselves to the following scope.

Scope of this work:

In this paper, we are concerned with only those ads that have opted in for advanced match. We intend to compare various well known information retrieval methods for (advance) matching queries to ads. While in practice several features associated with the observed historical click-through-rate of an ad might be used in addition to word-overlap features (see eg. the work of Richardson et al [30]), in this paper we consider mainly the problem of using traditional information retrieval features to determine relevance of an ad

to a query. We seek to evaluate whether techniques that work well in classic information retrieval will work for the problem of advance matching queries to ads. In addition, we propose a novel variant of translation models in information retrieval, where the relevance score is an estimate of the click-through rate of an ad to a query. We discuss how we would incorporate our methods into more feature based methods in Section 7.

In the following section, we describe the structure of an advertisement and the related terminology. In section 3, we describe the various models that we apply to the problem of advertisement retrieval. Section 4 outlines our experimental setup and evaluation procedure. In section 5, we report initial results and analyze the differences across the various approaches. Next, in section 6, we place our work in the context of other work in sponsored search and information retrieval. Finally, we derive conclusions from our current set of experiments and explore future directions in section 7.

The main contributions of this work are: (1) A study of information retrieval techniques for sponsored search with insights into when and why certain types of techniques will and will not work (2) A proposal for a new click-based translation model for sponsored search.

2. AN ADVERTISEMENT

In this section, we describe the template of an advertisement using the specific example of Yahoo!'s Panama platform. The other search engines are similar. We use an illustrative but fictitious example of an advertiser who sells all kinds of shoes on the internet. This advertiser typically will have an **account** with the search engine, and would run many **campaigns**. A campaign can consist of many ad groups each of which in turn consist of a set of related keywords for a campaign. [1]

Bidded terms or keywords: For each adgroup, there is a set of keywords that the advertiser bids on, e.g., *sports shoes*, *stilletoes*, *canvas shoes* etc.

Creative: A creative is associated with an adgroup and is composed of a title, a description and a display url. Advertisers may chose to use a template for an ad. The template may have a title like *Buy* {*keyword:cheap shoes*}, an abstract - *Find* {*keyword:shoes of all styles*} *at low prices* and a display url *cheapshooz.com*. The portion between curly braces can be substituted by alternate text (henceforth called alt text) corresponding to a bidded term. So for a bidded tem *sports shoes*, if the advertiser has specified the alt text as *sneakers*, the title will be converted to *Buy Sneakers*. Similar is the case for the abstract. The default text in the template is used in case the ad exceeds a certain length after the template is filled out.

Matchtype: An advertiser can choose to use "standard" or "advanced" match for the keywords or adgroup. For example, if the advertiser choses to use only standard match for the keyword "sports shoes", his ad may be shown for that exact query. Whereas, if he enables the keyword to be advance matched, the search engine can show the same ad for the queries "running shoes" or "track shoes".

Bid: Associated with each keyword is a bid. The final ranking displayed on the search engine is a product of the *bid* and the *relevance* of the ad to the query. Relevance can be assumed to be a surrogate for the expected click through rate (CTR) of the ad. Hence ranking by the product of relevance and bid is an attempt to maximize revenue for the search engine.

Landing Page: Clicking on an ad will lead the user to the landing page of the advertiser which can also be very informative of the relevance of the ad to the query. Note that the landing page is typically the "document" used in web search - the title and creative for displaying a web result are typically auto-generated by a model at runtime.

[1] `http://help.yahoo.com/help/l/us/yahoo/ysm/sps/manage/mngca/import_spreadsheet.html`

3. MODELS

A document $\mathcal{D}$ in our index is a unique creative composed of 5 zones: the unfilled templates of the title and description, the display url, the bidded terms and the alt text. Let $z = 1...5$ represent an index into each of the above mentioned five zones respectively. In this work, a query $\mathcal{Q}$ is represented as a bag of words $q_1, q_2, ..., q_n$. For all our models, the similarity of a query to a document ($S(\mathcal{D}, \mathcal{Q})$) is a linear combination of the similarity of the query to the individual zone. In other words,

$$S(\mathcal{D}, \mathcal{Q}) = \sum_z w_z S(\mathcal{D}_z, \mathcal{Q})$$

where $\mathcal{D}_z$ represents a zone of a document $\mathcal{D}$ and w_z is the weight attributed to the zone. Henceforth, we use $\mathcal{D}$ to mean $\mathcal{D}_z$ for better readability.

3.1 BM25

We evaluate the classic BM25 model [31] which is an approximation of the 2-poisson model and is often considered the state-of-the-art for many information retrieval tasks [34]. The model, as we used it, is outlined below:

$$S(\mathcal{D}, \mathcal{Q}) = \sum_i IDF(q_i) \frac{tf_{q_i,\mathcal{D}}(k_1+1)}{tf_{q_i,\mathcal{D}} + k_1(1 - b + b\frac{L_\mathcal{D}}{avg\_dl})} \quad (1)$$

$$IDF = log\frac{N - n_i + 0.5}{n_i + 0.5} \quad (2)$$

$$n_i = \text{number of documents containing } q_i$$

$$N = \text{collection size}$$

$$tf_{q_i,\mathcal{D}} = \text{term frequency of } q_i \text{ in } \mathcal{D}$$

$$L_\mathcal{D} = \text{length of document } \mathcal{D}$$

The parameters k and b are tuned empirically.

3.2 Term Presence Absence Model (PA)

We also explored a simple model that ignored TF and considered only the IDF part of the above equation. We call this simple model the term presence absence model; this model is likely to place greater emphasis on a document that has all the terms in a query than the above BM25 model.

3.3 Combination Vector Space Model (PA+-BM25)

Experiments on our training data with this model showed increased precision and decreased recall due to the PA model. This lead us to try out a linear weighted combination of the BM25 and this model as well (PA+BM25). We discuss all these models and their advantages and disadvantages in greater detail in section 5.

3.4 Query Expansion using the Web as an external resource

External resources for query expansion have proven useful for several information retrieval tasks [9, 18, 7] as well as sponsored search [28]. Since web queries are very short and often only 2-3 words in length, expansion helps add context, as well as add synonyms to the original query.

We tried a simple query expansion approach using the web. We queried our native search engine for the top 10 results. We concatenated the bag-of-words of the query-biased summaries of the top 10 results, and retained the top 5 most frequent terms as query expansion terms. We expect that the addition of terms such as "bank of america" and "bankofamerica" to the query "boa". We use the expanded query to retrieve documents in the vector space framework outlined in sections 3.1 and 3.2 above.

3.5 Language Models

The language modeling framework makes the assumption that a user has an idea of what the "perfect" document for his or her information need will look like. The user samples from this perfect document to generate the query $\mathcal{Q}$. The task of the system then is to estimate the document closest to the ideal document for the query $\mathcal{Q}$.

$$\underset{\mathcal{D}}{\operatorname{argmax}}\, P(\mathcal{D}|\mathcal{Q}) = \underset{\mathcal{D}}{\operatorname{argmax}}\, \frac{P(\mathcal{Q}|\mathcal{D})P(\mathcal{D})}{P(\mathcal{Q})}$$

In the query likelihood model we rank documents by $P(\mathcal{D}|\mathcal{Q})$. Typically the terms $P(\mathcal{D})$ and $P(\mathcal{Q})$ are ignored leading to ranking by $P(\mathcal{Q}|\mathcal{D})$. Like the Ponte and Croft model [26], we model $P(\mathcal{Q}|\mathcal{D})$ as an i.i.d sampling of the query words from a document model as $P(\mathcal{Q}|\mathcal{D}) = \prod_{i=1}^{n} P(q_i|\mathcal{D})$ where,

$$\begin{aligned} P(q_i|\mathcal{D}) &= \lambda P_{mle}(q_i|\mathcal{D}) + (1-\lambda)P_B(q_i|\mathcal{D}) \\ P_{mle}(q_i|D) &= \frac{tf_{q_i,\mathcal{D}}}{|\mathcal{D}|} \\ P_B(q_i|C) &= \frac{\sum_j^N tf_{q_i,j}}{\sum_i^{|V|}\sum_j^N tf_{q_i,j}} \end{aligned}$$

Note that, if we use a context-dependent formulation, or n-grams where n > 1, we will inherit some of the qualities of the PA model. Using bigram or even trigram context has two distinct advantages: (a) documents which have more of the query terms in close proximity will be preferred, and (b) the back-off probability will help enforce phrases. In this paper, however, we only report results using unigram models.

3.6 Translation Models

Berger and Lafferty [3] proposed modeling the query as a translation of a document. As described in section 3.5, the user has a notion of an ideal document. In this model, the query formulation process can be viewed as a translation of the ideal document into a query through a noisy channel. The translation process accounts for deletion and substitution of terms from the ideal document in the query. While the basic language modeling framework described above does not allow for query expansion, this model does. More recently this model has shown success in several information retrieval tasks such as sentence retrieval and FAQ retrieval [18, 24] where the "lexical gap" between the query and document is high and the documents are short.

We add a twist to the original model by trying to estimate the probability that the document will be clicked for a query-ad pair as follows. If C is a binary random variable that takes the value 1 to indicate that a click is observed and 0 to indicate that it is not, in this model we aim to rank documents by the $P(C = 1|\mathcal{D}, \mathcal{Q})$:

$$P(C|\mathcal{D},\mathcal{Q}) = \frac{P(\mathcal{Q}|\mathcal{D},C)P(C|\mathcal{D})}{P(\mathcal{Q}|\mathcal{D})} \quad (3)$$

where

$$P(\mathcal{Q}|\mathcal{D},C) = \prod_{i=1}^{n} P(q_i|\mathcal{D},C) \quad (4)$$

$$(5)$$

$P(q_i|\mathcal{D}, C)$ can be estimated in a number of ways. We chose the following mixture model:

$$\begin{aligned} P(q_i|\mathcal{D},C) &= \lambda_1 P_{mle}(q_i|\mathcal{D}) \\ &\quad +\lambda_2 P_B(q_i|\mathcal{D}) + \lambda_3 P_{TM}(q_i|\mathcal{D},C) \end{aligned} \quad (6)$$

The first two components of $P(q_i|\mathcal{D})$ are as in section 3.5 and in estimating them we do not consider the conditional factor of the clicks. The third component, viz, P_{TM} or the translation model can be expanded as follows:

$$\begin{aligned} P_{TM}(q_i|\mathcal{D},C) &= \sum_j^{|\mathcal{D}|} P(q_i|t_j,C)P(t_j|\mathcal{D},C) \quad (7) \\ P(t_j|\mathcal{D}) &= \sum P_{mle}(t_j|\mathcal{D}) \end{aligned}$$

The key to the translation model is in estimating the translation tables which associates a probability $p(q_i|t_j, C)$ for a word pair q_i, t_j where q_i may correspond to the token "shoes" and t_j correspond to the token "sneakers". Note that self translations are also modeled, i.e. we can have $p(q_i|t_j, C)$ where $q_i = t_j$. In this way, the model assigns a non-zero probability mass to those ads for which "translations" or synonyms (t_j) of the query term q_i occur in the ad. There are several sources for deriving the translation tables: from clicked query-ad pairs, web search results, wikipedia, user sessions, etc. Smoothing the translation probability across multiple sources provides robustness and diversity of translations. As described in section 3.5, using n-gram probabilities, where n > 1, enforce term proximity and automatically capture multi-word phrases.

To implement Equation 3, we need to model two additional components: $P(C|\mathcal{D})$ and $P(\mathcal{Q}|\mathcal{D})$. $P(C|\mathcal{D})$ can be considered to be a quality score for an ad independent of the query, which can be estimated from syntactic and semantic features and the prior historical click-through-rate of the ad. $P(\mathcal{Q}|\mathcal{D})$ plays the role of IDF in the vector space approach; we estimate the statistics for this component from all ads displayed for all queries, not just the clicked query-ad pairs. The denominator in Equation 3 can be used to discriminate the clicked ads from the non-clicked ads given a query.

Note that in this paper, we only focus on deriving unigram translation tables from clicked ads. We leave the estimation of the complete $P(C|\mathcal{D}, \mathcal{Q})$ and the use of n-gram probabilities for future work.

4. EXPERIMENTAL SETUP

In this section, we describe the tools, training and test data and the evaluation.

$p(q_i = \text{yoga}\|t_j)$		$p(q_i = \text{cyst}\|t_j)$		$p(q_i = \text{acetaminophen}\|t_j)$	
ilchi	0.500	dermoid	0.466	antipyret	0.250
dahn	0.453	pilonidal	0.465	paracetemol	0.153
iyengar	0.439	bartholin	0.440	overdose	0.068
ashtanga	0.400	epidermoid	0.416	analgesic	0.037
astanga	0.384	ganglion	0.361	acetylcysteine	0.033
kriya	0.355	epiderm	0.273	pathophysiology	0.016
asana	0.354	sebaceous	0.242	caplet	0.010
hatha	0.320	popliteal	0.158	hydrocodon	0.007

Table 1: Example translation tables (from Web Search). The table shows the top terms sorted by $p(q_i|t_j)$ for 3 different t_j.

4.1 Tools

We indexed only those ads for which the advertiser had opted in for advanced match. We use a similar infrastructure to the work of Broder et al [7], i.e., we use Hadoop grid computing infrastructure to preprocess the ads and build an inverted index [13] and use the WAND algorithm to retrieve ads [6]. Stemming was done using the Porter stemmer[27]. Since our methods score each zone differently, we maintained a separate postings list for each zone. Hence, TF, IDF and other statistics can be computed for each zone. URL segmentation was done using a simple unigram model whose vocabulary was trained on a web document collection and a decoder that used a dynamic programing algorithm to retrieve the best segmentation for a given URL.

4.2 Data

Our training and test data comprised of query-ad pairs that had been judged by trained editors on a 5 point scale - *Perfect, Excellent, Good, Fair* and *Bad*. The editors only looked at the creative and not the landing page while making their judgment. The editors were trained for the task and were instructed to reserve the judgment "Perfect" to those query-ad pairs where the query has an unambiguous target (eg. "abc.com") and the ad's display url corresponds exactly to that target. This is typically true only for navigational queries. The judgment "Fair" was reserved to those query-ad pairs for which it was not completely obvious that the user would be able to find what he or she was looking for after a click, but there was a reasonable chance of doing so. The judgment criterion was quite similar to the work of Metzler and Dumais [22].

We had a set of about 47000 query-ad pairs for about 1000 unique queries that had been judged for a different system that we used as our development training set. We tuned some of the parameters for the different models on this training set. We indexed all the ads that had opted in for advanced match and that were present in our advertiser database on one day in April. We retrieved ads using BM25, PA, LM, TM, TM(Web) and the PA+BM25 models on 317 queries sampled from a query log of the first 2 weeks of April from a major search engine. We used the median score of a method as a threshold to filter out query-ad pairs so as to decrease the effort of the editors. In all we had about 10000 query ad pairs to be judged. In our final evaluation data set the number of Perfect, Excellent, Good, Fair and Bad judgments were 5, 21, 335, 1648 and 7735 respectively. 87 queries had no relevant documents (examples are *how to make a pinata* and *human body system*) and only 76 queries had greater than 10 relevant documents. In our final evaluation, we considered un-judged documents retrieved by an algorithm as non-relevant.

4.2.1 Estimating the translation tables

In this work we estimated the translation tables ($p(q_i|t_j, C)$ in Equation 7) from two sources of data.

1. Using Sponsored Search Click Data: We used a month's worth of sponsored search click data that was available to us from November of last year. The data consists of tuples of the form <query, ad, click>, where *click* indicated whether a click was observed (click = 1) for that query and ad pair or not. Clicks were spam filtered. We considered all lines where a click was observed and estimated the probability of $P(q_i|t_j)$ to be $P(q_i|t_j, click = 1)$ as follows.

$$P(q_i|t_j) = \frac{\#(q_i \in \text{query \& } t_j \in \text{ad \& click=1})}{\#(t_j \in \text{ad \& click=1})}$$

   A query and ad pair had to be clicked a minimum number of times before it was used in the computation.

2. Using Web Search Logs: Web search engines typically show a ranked listing of results for a query. The listings usually comprise of a title, an abstract of the page and a url (similar to an ad's title-creative-url). We used the Yahoo! API and obtained the summaries for the top 10 results for the top 200K unique queries on our search engine. We estimated $P(q_i|t_j, C)$ in a manner similar to what is described above, except that we used information from the web-page abstract shown on the "organic" or web-search results page and ignored the click information (we did not have the click information for the search engine and we believe it is reasonable to assume that the search engine did quite well for popular queries). In other words $P(q_i|t_j, C) \sim P(q_i|t_j)$.

Table 1 shows example translation probabilities learned using the second method listed above. We deliberately chose non-commercial examples since we did not want to label specific brands or product names and model numbers, which often show up as the top translations for many commercial queries. As can be seen from the examples, many of the expansions explore facets of a query.

The probabilities $P(q_i|t_j, C)$ can also be estimated from other sources of click data: eg., reformulations of a query where the reformulated query had a web result clicked on and so on. If these additional sources of information are available, we believe they can prove to be quite useful.

4.3 Evaluation Measures

We evaluate our work using standard measures from information retrieval and web-search such as precision at different ranks and recall. The precision at rank r, P_r is defined as the fraction of relevant documents in ranks 1 to r that are considered relevant. Recall is defined as the fraction of documents that are relevant to the query that are retrieved. Average precision of a query is computed by averaging the precision at different points of recall. In other words:

$$AP_q = \frac{\sum_{r=1}^{M} P_r \times R(r)}{\text{no. of rel docs for q}} \quad (8)$$

where M is the number of retrieved documents and $R(r)$ is a binary variable which takes on the value 1 if there is a relevant document at rank r and 0 otherwise. Mean Average Precision is the average of the average precision scores computed for all queries. Since MAP requires binary relevance judgments, we report results for two cases: (1) where every document judged "Good" or better was judged relevant and (2) where every document judged "Fair" or better was judged relevant . Anything un-judged was marked as "Bad".

Since our judgments are obtained on a five point scale, we can also compute metrics based on discounted cumulative gain (DCG) [17], a measure which aims at measuring user experience. The DCG of a query is given by:

$$DCG = \frac{\sum_{i}^{M} \frac{\text{score}(label_i)}{log_2(i+1)}}{M} \quad (9)$$

where M is the number of retrieved documents for that query and $label_i$ is the judgment of a document at rank i. The scores assigned to Perfect, Excellent, Good, Fair and Bad documents are 10, 7, 3 and 0.5 respectively. The normalized DCG (nDCG) metric for a query normalizes the DCG by the maximum value of DCG that is possible for that particular query. In addition to the average DCG of the retrieved results we report DCG at ranks 1 and 5.

5. RESULTS AND DISCUSSION

Results from our experiments are given in Table 2. Of the first two models, BM25 and PA, the BM25 model retrieves many more ads that are "Good" or better (higher recall), but the PA model does better in ranking the good ones at the top of the ranked list. Many ads can contain several 100s of bidded terms for the same creative for eg., "running shoes", "jogging shoes", "walking shoes", "comfortable walking shoes" and so on. Thus a word like "shoes" can get repeated a few 100 times inflating the TF component of the above model quite significantly. For creatives such as these the PA model does much better, but for longer queries the BM25 model appeared to do better. The model with the weighted combination of these two (PA+BM25) has a weight of 0.333 for the PA model and 0.666 for the BM25 model. Of the vector space models without expansion this model seems to perform the best.

Of the language modeling methods, the translation models are expected to decrease precision and increase recall. The translation models learned from querying the search engine [TM(web)] significantly increases the recall as expected. However, the translation models that learn from click data (TM) did not improve recall as much. The best performing model is the model that performs expansion based on the abstracts of the web-results (last column in Table 2).

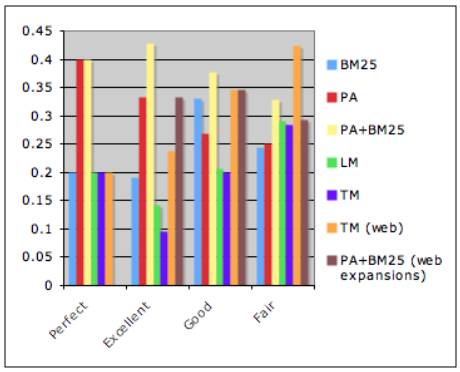


Figure 1: Fraction of "Perfect", "Excellent", "Good" and "Fair" results retrieved by each method. Note that there are only 5 "Perfect" results, 21 "Excellent" ones, and 335 and 1648 "Good" and "Fair" ones respectively.

Although there is some non significant decrease in precision and nDCG, the overall improvement in mean average precision is statistically significant. The loss in DCG comes mainly from the top ranks where the decrease is statistically significant.

The numbers in Table 2 are all micro-averaged. If we consider macro-averaged recall as shown in figure 1, we notice that the translation model using web based expansions has high recall, especially for "Fair" documents. This means that there are certain queries for which the translation models perform significantly better. An example is *amtrak and schedule* to which the addition of the terms *rail,vacation* and *adventure* and the url token *amtrak.com* results in significant increase in recall.

The translation model learned from click data did not perform as well as expected. Upon analysis, we found that many of the translations learned were from a query-word to a segment of the display url. Since the data used to train the translation probabilities was a few months older than the index on which retrieval was performed, many of the the ads containing the same tokens in the display url did not exist in the current database. Few expansions actually triggered and hence the TM model is quite similar to LM in performance. We believe that cleaner dictionaries that "expire" less easily may be constructed by considering only the query and ad-title and/or query and bidded term. Another reason for the TM model not performing well may lie in the fact that we did not do any rank normalization when we learn the translation probabilities from clicked query-ad pairs. Since exact matches typically show up higher in the display, it is likely that we are learning more self translation probabilities than synonym translations. On the other hand, with the web data, we had 10 results per query to expand with and the dictionaries learned were less noisy and had more synonym translations.

The translation model that learned from web-result abstracts does not perform as well as the PA+BM25 model that expands on the web abstracts. The principle difference here is that the translation dictionaries in the language modeling framework were not query specific. This led to much quicker "query drift". For example the query *baja* occurs in many queries that are in the context of *baja*

	BM25	PA	LM	TM	TM (Web)	PA+ BM25 (unigrams)	PA+ BM25 (web expansions)
DCG_1	0.494	**0.591**	0.452	0.448	0.398	0.522	0.334
DCG_5	1.074	**1.134**	1.049	0.987	0.940	1.095	0.667
DCG	1.456	1.406	1.494	1.438	1.415	**1.570**	1.154
NDCG	0.357	0.331	0.362	0.351	0.361	**0.384**	0.324
Good or Better is relevant							
P_1	0.218	0.276	0.206	0.206	0.178	0.237	**0.288**
P_5	0.147	0.136	0.146	0.130	0.126	0.138	**0.169**
Recall	0.318	0.241	0.296	0.268	0.301	0.338	**0.407**
MAP	0.156	0.153	0.137	0.130	0.116	0.163	**0.236**
Fair or Better is relevant:							
P_1	**0.459**	0.391	0.412	0.401	0.367	0.457	0.386
P_5	0.296	0.291	**0.312**	0.303	0.283	0.309	0.300
Recall	0.289	0.273	0.335	0.321	0.361	0.320	**0.381**
MAP	0.198	0.178	0.210	0.202	0.199	0.213	**0.246**

Table 2: Results of 7 different models on the task of retrieving ads relevant to a set of queries. Bolded values indicating the best performing system for a given metric and underlined values indicate statistical significance (at the 95% level of confidence) as compared to the baseline (BM25)

motorsports, leading to a translation dictionary that corresponds to this theme. However expanding on this theme for a query like *baja fresh* will change the context of the original query significantly. There are many solutions to work around this problem: one way is to construct query specific translation dictionaries using content of the landing page of the clicked url in the same way that the web-abstracts were used. This approach can be expensive from a storage perspective and may not work for truly tail queries. Another approach would be to embed context into the translations, explicitly via query segmentation into phrases for phrase to phrase translation tables [2], or implicitly via use of n-gram contexts in the translation probabilities as proposed earlier in this paper. Also, although we filtered out low frequency terms (t_j) in computing $P(q_i|t_j)$, the filtering was probably insufficient, leading to some incorrect high probability translations. Modeling deletions and insertions via more advanced translation models like IBM Model-3 can also be beneficial.

There are also certain classes of queries for which unigram models do not perform well. Particularly notable in our analysis were queries with geographic intent and people names. It is obvious that showing ads for *pizza in springfield, illinois* is probably not acceptable for a user located in Springfield, MA. Significant number of web-search queries have geographic intent and they should probably be handled seperately [20, 12]. Likewise showing ads for *jeniffer lopez* for the query *Jeniffer Howard* (a politician) is not acceptable. Some of these errors may be controlled by using phrase based models or using entirely different retrieval models for certain classes of queries [16].

We also played around with alternate forms of smoothing like Dirichlet smoothing that overcomes some of the document length normalization issues (note that the zone corresponding to the bidded terms can have significant variance in length). We did not see significant improvement in performance as compared to the Jelinek Mercer Smoothing method whose results we reported.

6. RELATED WORK

The information retrieval community has studied the problem of matching queries to relevant document documents for several years [34]. Queries can be long like in the TREC adhoc retrieval tasks unambiguous natural language questions or web-queries. Likewise retrieval on several types of collections have been studied. Perhaps the most relevant areas within information retrieval for this work are the following areas: the classic ad-hoc retrieval task, web document retrieval, retrieval of small snippets of text and online advertising, a fairly nascent field.

Work in online advertising focuses on two main areas: contextual advertising and sponsored search. Contextual advertising mainly concerns itself with the placement of ads on publisher pages. Since publisher pages are rich in content, a rich set of features can typically be extracted from the webpage and used to find relevant ads [5, 33]. The Sponsored Search problem on the other hand suffers from the same problem as web-search – that the queries are short and have little context. Exacerbating the problem is the fact that the document is short with little context as well. One way of overcoming this problem is though "query rewriting" techniques. The transformed query is then used for retrieval. Models to predict query rewriting techniques may be learned from query logs [19, 28]. Alternately some techniques, including ones explored in this paper, expand the query using the *organic search* or web-search results [7]. A third source of data that we do not use in this paper, but has been proven useful for sponsored search is the historical click-through-rate (CTR) of a query-ad pair in predicting its relevance to a query. We envision our system to be a two-stage one where the first stage relies less on history and the second "re-ordering" stage may use historical CTR in addition to word-overlap features [8]. This allows new advertisers who have never been shown for a given query to have a chance to be show up on the search results page. CTR information can also be incorporated into the term $P(C|\mathcal{D})$ in Eq 3.

Query expansion is generally accepted as beneficial for information retrieval. Expansion may be through pseudo rele-

vance feedback [36, 21] or interactive techniques [14]. Using the web as an external resource has proven beneficial for sponsored search and other information retrieval tasks [7, 35, 9, 10, 32] and particularly when the snippets of text that need to be matched are short [22, 32]. Ribeiro-Neto et al [29] found expanding the content of publisher pages to be useful to the problem of contextual advertising. In the field of machine learning the addition of unlabeled data to improve classification accuracy has received special attention through the fields of unsupervised and semi-supervised learning [23].

Jeon[18] and Murdock [24] recently used the translation models of Berger and Lafferty [3] for query expansion for two new tasks and found considerable improvement. While Jeon was attempting a Q&A retrieval task, Murdock was attempting a sentence retrieval task. Given the short length of sentences Murdock naturally found expansion beneficial. Murdock et al [25] also applied the translation model approach to contextual advertising. They computed translation models between publisher pages and landing pages by using a parallel corpus determined by human judgments. We presented a variation of the translation models in which the translation dictionaries use click information. Our dictionaries can easily be computed from search engine logs and therefore our method is more robust to seasonal variations in the vocabulary of commercial terms.

The work of Zhou et al [38] attempts to model $P(C|\mathcal{Q},\mathcal{D})$ in by factoring out the query into units. In their method, the query $\mathcal{Q}$ is broken into units or phrases for which high CTR for the document $\mathcal{D}$ is seen. The probability $P(C|\mathcal{Q},\mathcal{D})$ is then computed as the product of the observed CTR of the sub-phrases when issued as individual queries. However, their method does not incorporate query expansion.

7. CONCLUSIONS AND FUTURE WORK

In this paper we explored several models of information retrieval from the classic BM25 model to the translation models. The BM25F model has shown to be effective for retrieval on certain semi-structured documents like HTML pages. As a next step, we would also like to train the BM25F model that uses a separate b and f parameter for each zone and an additional saturating parameter [37]. Since the different zones in the ad creative have different properties such as length and the way in which terms repeat, we expect that tuning parameters per zone would improve performance.

The translation models in general improved recall. Expanding the original query with web search results significantly improves performance; using sponsored search clicked query-ad pairs showed less benefit. However, we believe that there is plenty of room for improvement of the translation model. We used very simple co-occurence probabilities for our translation dictionaries. The next step would be to use more sophisticated translation models including phrase based and/or ones with n-gram contexts. The translation models also lend themselves naturally to using a mixture modeling framework, where the mixtures can be over different corpora from which the translation tables are derived, or different query slices that determine semantic categorization of queries. We have not looked into using query reformulation data for learning translations, or into using non-Yahoo sources of data. Combining translation tables from different sources will improve smoothing (reinforce translations coming from multiple sources) and increase the coverage of the translation models over a larger fraction of the tail queries. Slicing queries based on query intent and/or based on query clusters may also improve targeting the translation models to a specific query ("baja fresh" and "baja motorsports" fall in different semantic categories).

We proposed a new variant of the translation model which aims at capturing the probability that a given ad will be clicked for a query. However, in this paper, we have not fully explored this model. We would like the use historical CTR of an ad and the query, combined with textual features as a prior in equation 3. Rank-normalizing the impact of clicks on ads may also play an important role in learning more synonym translations rather than self translations.

Finally, we did not explore feature-based models for retrieval in this paper. For instance, we can use the scores from our information retrieval models in models that learn from click logs like those of Richardson et al [30] and Ciaramita et al [8] either as a prior or as a feature in a machine learning model. These models are can be trained and evaluated on clickthrough-logs. We however, believe such models that are trained and optimized to perform well on historical data should be used in a re-ranking step as opposed to the initial retrieval since such models are heavily biased towards what has been seen, resulting in fewer opportunities for new advertisers to be shown on the page. Using a model that relies on historical clicks of ads for re-ranking, but not for initial retrieval would provide new advertisers the opportunity to be shown on the page, and improve their ranking if clicks are observed.

While the problem of sponsored search is relatively new and offers several new areas of exploration, we believe that several tools and techniques developed for several information retrieval tasks may be directly applied with small modifications and enhancements for the new task.

Acknowledgments

We would like to thank our colleagues in the Sponsored Search team at Yahoo! and the reviewers of this paper for useful feedback.

8. REFERENCES

[1] `www.emarketer.com/Article.aspx?id=1006319`.

[2] L. Ballesteros and W. B. Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. *SIGIR Forum*, 31(SI):84–91.

[3] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 222–229, New York, NY, USA, 1999. ACM.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 559–566, New York, NY, USA, 2007. ACM.

[6] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *CIKM '03: Proceedings of the twelfth international conference on Information and*

knowledge management, pages 426–434, New York, NY, USA, 2003. ACM.

[7] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using web knowledge. In *SIGIR '07*, pages 231–238, New York, NY, USA, 2007. ACM.

[8] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *WWW '08: Proceedings of the 16th international conference on World Wide Web*, 2008.

[9] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *SIGIR '06*, pages 154–161, New York, NY, USA, 2006. ACM.

[10] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. Web question answering: is more always better? In *SIGIR '02*, pages 291–298, New York, NY, USA, 2002. ACM.

[11] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, March 2007.

[12] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel. Analysis of geographic queries in a search engine log. In *LOCWEB '08: Proceedings of the first international workshop on Location and the web*, pages 49–56, New York, NY, USA, 2008. ACM.

[13] Hadoop. `http://hadoop.apache.org/core/`.

[14] D. Harman. Towards interactive query expansion. In *SIGIR*, pages 321–331, 1988.

[15] B. Jansen and M. Resnick. Examining searcher perceptions of and interactions with sponsored results. In *Workshop on Sponsored Search Auctions*, 2005.

[16] B. J. Jansen, D. L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1149–1150, New York, NY, USA, 2007. ACM.

[17] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR '00*, pages 41–48, New York, NY, USA, 2000. ACM.

[18] J. Jeon. *Searching Question and Answer Archives*. PhD thesis, University of Massachusetts, Amherst, 2007.

[19] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 387–396, New York, NY, USA, 2006. ACM.

[20] R. Jones, W. V. Zhang, B. Rey, P. Jhala, and E. Stipp. Geographic intention and modification in web search. *International Journal of Geographical Information Science*, 22(3):229–246, 2008.

[21] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, New York, NY, USA, 2001. ACM.

[22] D. Metzler, S. Dumais, and C. Meek. Similarity measures for short segments of text. *Advances in Information Retrieval*, pages 16–27, 2007.

[23] T. M. Mitchell. *Machine Learning*. McGraw-Hill Higher Education, 1997.

[24] V. Murdock. *Aspects of Sentence Retrieval*. PhD thesis, University of Massachusetts, Amherst, 2007.

[25] V. Murdock, M. Ciaramita, and V. Plachouras. A noisy-channel approach to contextual advertising. In *ADKDD '07: Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, pages 21–27, New York, NY, USA, 2007. ACM.

[26] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281, New York, NY, USA, 1998. ACM.

[27] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130?137, 1980.

[28] F. Radlinski, *A. Broder*, P. Ciccolo, E. Gabrilovich, V. Josifovski, and L. Riedel. Optimizing relevance and revenue in ad search:a query substitution approach. In *SIGIR '08*, 2008.

[29] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. de Moura. Impedance coupling in content-targeted advertising. In *SIGIR '05*, pages 496–503, New York, NY, USA, 2005. ACM.

[30] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 521–530, New York, NY, USA, 2007. ACM.

[31] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94*, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[32] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 377–386, New York, NY, USA, 2006. ACM.

[33] W. tau Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 213–222, New York, NY, USA, 2006. ACM.

[34] TREC. `http://trec.nist.gov`.

[35] E. Voorhees. Overview of the trec 2004 robust retrieval track. In *Proceedings of the 14th Text REtrieval Conference*, 2005.

[36] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.

[37] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson. Microsoft cambridge at trec-13: Web and hard tracks. In *TREC-2004*, 2004.

[38] D. Zhou, L. Bolelli, J. Li, C. L. Giles, and H. Zha. Learning user clicks in web search. In *IJCAI*, pages 1162–1167, 2007.

Characterizing Query Intent From Sponsored Search Clickthrough Data

Azin Ashkan, Charles L.A. Clarke
University of Waterloo, Canada
{aashkan, claclark}@cs.uwaterloo.ca

Eugene Agichtein, Qi Guo
Emory University, United States
{eugene, qguo3}@mathcs.emory.edu

ABSTRACT

Understanding the intention underlying users' queries may help personalize search results and therefore improve user satisfaction. If a commercial intent exists, and if an ad is related to the user's information need, the user may click on that ad. In this paper, we develop a methodology for using ad clickthrough logs from a commercial search engine to study characteristics of commercial intent. The findings of our study suggest that ad clickthrough features, such as deliberation time, are effective in detecting query intent. We also study the effect of query type and the number of displayed ads on ad clickthrough behavior.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Query Intent, Sponsored Search

Keywords

Ad Targeting, Query Log, Clickthrough

1. INTRODUCTION

Intent detection is one of the crucial long-standing goals of information access. Understanding the intent underlying user queries may help personalize search results and therefore improve user satisfaction. User intent may correspond to any of the standard categories of Web query [2]: *navigational*, *informational*, and *transactional*. On the other hand, in the context of sponsored search, information providers may also wish to know whether a user has the intention to purchase or utilize a commercial service, or what is called *online commercial intention* [3]. Sponsored search has evolved to satisfy the needs of users for relevant search results and the desires of advertisers for increased traffic to their Websites. It is now considered to be among the most effective marketing vehicles available [5]. It basically operates by matching ads to queries as they are received by a search engine. These ads are displayed to the user, along with organic search results. The most common model is pay per click, where advertisers are charged based on user clicks (if any) on the displayed ads [1]. Ideally advertisers wish to bid on multiple low-cost, highly targeted keywords that will generate high clickthrough rates for their ads.

In order to identify search query intention, implicit feedback techniques take advantage of user behavior to understand their interests and preferences. Amongst the implicit feedback techniques, clickthrough-based analysis considers the history of user-submitted queries and user-selected documents on the corresponding search result pages. Bringing the query intent detection into the context of sponsored search can help advertisers to automatically create more appropriate and relevant ad content, develop better ranking ads by matching the content of the ads with the users query intent, as well as contribute to the general understanding of user intent inference and web search behavior modeling. In this regard, we divide our motivations of this work into three parts: i) detecting the intentions of queries based on ad clickthrough features, ii) estimating the average ad clickthrough rate for each query type, and iii) studying the ad clickthrough behavior of newly arriving queries in different categories of query intent.

In the first part, we define features and train a decision tree classifier to categorize queries in two dimensions: commercial-noncommercial and navigational-informational. We define a *commercial* query as a query with the underlying intention to make an immediate or future purchase of a specific product or service, while anything else falls into the *noncommercial* category. Furthermore, a *navigational* query is defined as a query with the underlying intention to locate a specific Web site or page, while an *informational* query is everything else. Rose and Levinson [14] conducted a study, developing a hierarchy of query goals with three top-level categories: informational, navigational and resource. Under their taxonomy, a transactional query as defined by Broder [2] might fall under either of their three categories, depending on details of the desired transaction. In this paper, a transactional or resource query would be subsumed under one of the two categories of either navigational or informational, as appropriate.

We aim to distinguish between different query types according to their ad clickthrough behavior. For this reason, in

SIGIR-IRA, July 2008, Singapore.

the second part of the work, the average clickthrough rate is estimated for different query types. For each query type, this estimation is performed separately according to the number of displayed ads. The results of these estimates can be used as evidence to indicate how much the number of displayed ads determines the number of ad clicks for each query type. Finally, in the last part of our work, we use the obtained average clickthrough ratio for each type of query as a means for calculating the number of ad clicks for previously unseen queries.

The remainder of the paper is organized as follows: Section 2 discusses related work. Section 3 presents a general picture of the data set and provides some details on the post-processing of the data in order to prepare it for analysis. In Section 4, we study properties of queries with respect to deliberation time and the number of displayed ads. These properties are used in our classifiers and in our prediction model. Section 5 presents our decision tree based classifiers used to identify query intent based on ad clickthrough features. Section 6 describes the proposed ad clickthrough prediction model based on query intent and on ad numbers. Finally, we conclude the findings and discuss further research possibilities in Section 7.

2. RELATED WORK

Regelson et al. [12] estimate the clickthrough rate of new ads by using the clickthrough rates of existing ads with the same bid terms or topic clusters. Similar work by Richardson et al. [13] incorporated features that depend on more than just the bid terms, including information about the ad itself, such as the length of the ad, the page the ad points to, and statistics concerning related ads. On the other hand, in recent work by Debmbsczynski et al. [4] the authors did not have access to the ad contents and keywords. They approximated the title and the body of each ad by combining all queries for which a given ad was displayed. They also use features based on the search result page (the rank of the ad and result page number) and on the ad's target URL. They used these extracted features to build a prediction model based on decision rules that they used to generate recommendations on how to improve the quality of ads. All three of these works focus on ad-based features in order to predict the clickthrough rate of new ads that would help to predict the quality of these new ads. We study the average ad clickthrough rate for queries in terms of their underlying intent using the query-based features and ad clickthrough statistics.

In the area of sponsored search, Dai et al. [3] propose a commercial query detector. They train machine learning models from two types of data sources for a given query: content of the search result page(s) and contents of the top pages returned by the search engine. Their findings indicate that frequent queries are more likely to have commercial intent. In the general context of query intention based on clickthrough data, Lee et al. [11] predict user query goals in terms of navigational and informational intents. They show that the prediction can be performed on the basis of two types of feature sets: past user-click behavior and anchor-link distribution. In this paper, we consider two dimensions of query intent, commercial/noncommercial and navigational/informational, utilizing features extracted from the ad clickthrough data for search queries.

In [6], Ghose et al. study the effect of sponsored search at a keyword level on the ad clickthrough rate. Their results indicate that while retailer-specific ads (based on navigational queries) increase clickthrough rates, brand-specific ads (based on transactional queries) decrease clickthrough rates. However, they focus on data from one advertiser only. Our work studies differences amongst queries (with different underlying intents) in terms of average ad clickthrough rates by pooling data on ads from multiple advertisements. We show that, on average, commercial-navigational queries receive more ad clicks than commercial-informational queries.

Jansen et. al. [10] study the factors influencing the ad clicks by searchers. They report that searchers have a bias against sponsored links (ad results) as compared to non-sponsored links (organic results). In other work, Jansen [9] studies the relevance of sponsored results and non-sponsored results by submitting a set of previously collected commercial queries to three major search engines. Jansen concludes that average relevance ratings for sponsored and non-sponsored links are practically the same, although the sponsored links relevance ratings are statistically higher.

3. DATA SET

The results reported in this paper are based on a data set obtained from Microsoft adCenter search and ad click logs sampled over a few months. Personally identifying information was removed from this data set. The data includes a sample of roughly 100 million search impressions, where an impression is defined as a single search result page. There is also a set of ad clicks (about 8 million) that are associated with the impression data.

Each impression and each click is described by a set of attributes. The set of attributes from the impression data used in this paper are as follows: *date and time of the impression*, *user query*, *number of ads displayed in results of the impression*, *user session ID*, and *impression ID*. The set of attributes used from the clickthrough data is as follows: *date and time of the click*, *user query*, *the target host for the clicked ad*, *user session ID*, and *impression ID*.

3.1 Data Post-processing

Queries are assumed to be in the English language. We removed any extra space at the beginning and end of the queries, and between words of the queries for both the impression and the clickthrough files. We then case-normalized the queries. We found about 27 million queries occurring only once in the impression file, mostly with no ads. Such queries were removed from the impression data. Impressions with a duplicate combination of impression id and user session id were removed in order to filter out repeated queries from the same user. Consequently, we ended up with about 25 million unique queries in the impression data set (about 75 million unique impressions) and about 2.4 million unique queries for which there was at least one ad click recorded in the click data.

In order to prevent train-test contamination, we split the impression and clickthrough data into three equal-sized sets (train, test, validation) on a query-level. In other words, all the impressions and click data for a given query went into the same set. This process was achieved by randomly assigning each query (with all its impression and click information) into one of the three sets. All the three sets

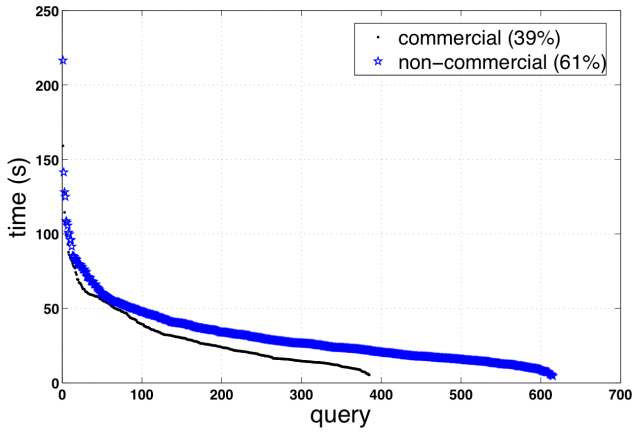


Figure 1: Deliberation Time between Entering a Query and Clicking on an Ad for that Query (for Queries Manually Labeled as Commercial/Noncommercial)

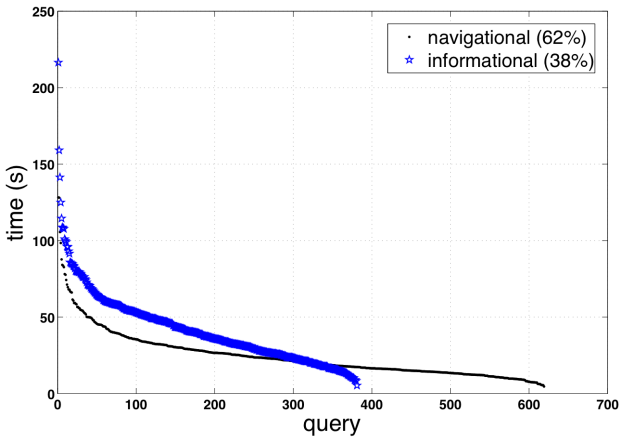


Figure 2: Deliberation Time between Entering a Query and Clicking on an Ad for that Query (for Queries Manually Labeled as Navigational/Informational)

contain approximately the same number of queries (about 8.3 million). As mentioned before, there are many queries with very small number of ad clicks. Similar to Richardson et al. [13], since our analysis deals with empirical ad clickthrough of queries, it may be wildly different from the true clickthrough rate for queries with few number of ads, leading to noise in the training and testing processes. Hence, we further filtered the three sets to include only those queries that have at least four ad clicks. After the filtering, we ended up with 44,941, 45,032, and 44,909 queries in the test, train, and validation sets respectively (134,882 queries in total).

In the remainder of the paper, we will refer to the case/ space/ user normalized impression data and its corresponding clickthrough information as the *original* data. Otherwise, by impression or clickthrough data, we mean one of the three sets (i.e. train, test, and validation) of impressions and their corresponding clickthrough data created from the original one.

3.2 Labeling Process

The original impression data was sorted based on the time of the impression. Starting from an arbitrary point in the file (approximately 1/5 of the length of the file from the beginning), 1000 queries were selected for which: i) the query was contained in the training data, and ii) the ad click frequency of the query was greater than 10. Each selected query was then manually labeled as commercial/noncommercial and navigational/informational by one of the authors.

If the assumed purpose of a query was to locate a specific Web site or page, the query was labeled as "navigational". Everything else was considered as "informational". We ended up with 62% of the queries labeled as navigational and 38% labeled as informational. The person who labeled the data was responsible for judging the assumed commercial intent of the search queries from the perspective of a user as well. If the assumed purpose of submitting a query was to make an immediate or future purchase of a product or service, the query was labeled as "commercial". Otherwise, if the purpose of the query was assumed to have little to do with commercial activity, it was labeled as "noncommercial". Since we focus on queries with at least a few number of ad clicks (11 for the labeled data and 4 in general for the three sets), one could consider that the data set construction favors including only commercial queries. In that case, commercial and noncommercial could be considered as "strongly commercial" and "slightly commercial" respectively. As a result of the labeling, 39% of the queries were labeled as commercial and 61% were labeled as noncommercial. Moreover, the focus of this paper is on the ad clickthrough-based features, where they should not be that meaningful for queries with no or relatively few displayed ads leading to noises in the classification. Therefore, we stick to classifying our filtered set of queries from the commercial/noncommercial perspective. It is also worth mentioning that the labeling result (specially for commercial/noncommercial) is subjective. In order to have a more confident result, a further exploration of this work could use multiple annotators in order to assign the final labels based on the maximum agreement among the annotators.

4. INITIAL ANALYSIS OF THE DATA

In this section, we study some properties of the data set for use in further experiments. One property is deliberation time, the average time between a query and an ad click. The other one is the average clickthrough rate for all impressions for which a particular number of ads are displayed.

4.1 Time Analysis on the Labeled Data

For each hand-labeled query, the average deliberation time for that query was calculated. The plots for each of the dimension of query type are depicted in Figures 1 and 2, where the queries are sorted by decreasing deliberation time. According to Figure 1, the deliberation time for a commercial query is generally less than for a noncommercial one. We explain this observation according to the intuition that ads basically target commercial queries more than noncommercial queries. In other words, it is more likely to find a related

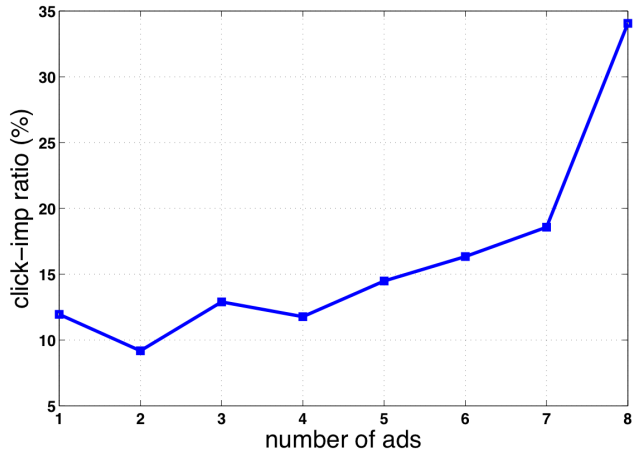


Figure 3: Average Click to Impression Ratio for Impressions with a Particular Number of Ads (lines do not imply interpolation)

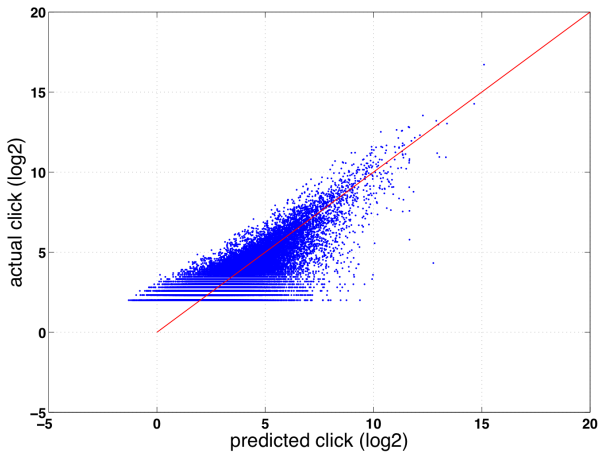


Figure 4: The Actual Number of Clicks vs. the Estimated Number of Clicks

answer among the ads for commercial queries comparing to noncommercial ones. Therefore, finding the related answers to commercial queries and thus clicking on them should take less time on average than the case for noncommercial queries. On the other hand, when it comes to navigational versus informational queries (as depicted in Figure 2), users would spend less time for navigational queries in comparison to informational ones. Because, there is usually one answer for a navigational query, often presented at the top of the result list. Users would click on their desired target as soon as they find it among the displayed ads.

Generally speaking, it appears that deliberation time can be considered as an important feature for distinguishing commercial queries from noncommercial ones, and navigational queries from informational ones. Therefore, we provide it as a feature to the classifiers described later in the paper.

4.2 Click to Impression Ratio for Varying Number of Ads

The average number of clicks per impression (clickthrough rate) for queries with a particular number of ads was calculated for the training set. In order to do that, the impressions are sorted according to the number of ads displayed for each. The number of ads in the impression file varies from one to eight. Thus, impressions are divided into eight groups, each denoted as set A_i, where i is the number of displayed ads for the impressions in that set. The value $|A_i|$ indicates the number of impressions with i ads displayed. We use the unique id number for each impression (impression id) to find out whether it resulted in an ad click. Repeating this process for all impressions in the eight groups, we can calculate the total number of ad clicks resulting from the impressions in each group.

Let $id_i^j \in A_i$ denote the unique id for the j^{th} impression in A_i. We define c_i^j to represent whether there was an ad click resulting from such an impression. In other words, $c_i^j = 1$, if there is an ad click associated with id_i^j in the clickthrough data, and $c_i^j = 0$ otherwise. Hence, the average number of ad clicks per impression (clickthrough rate), CTR_i, for queries with a particular number of ads i is obtained as follows:

$$CTR_i = \frac{\sum_{j=1}^{|A_i|} c_i^j}{|A_i|} \qquad 1 \leq i \leq 8 \tag{1}$$

We calculated the average clickthrough rate for the eight ad-based groups of the train set which resulted in the plot depicted in Figure 3. For clarity of presentation, we connect the points for each particular number of ads, while the lines do not imply the interpolation of points.

We use the obtained rates for the training set in order to estimate the number of clicks for each query in all the three sets (train, test, and validation). Such an estimation is performed based on the number of ads displayed for each query (thus, the average clickthrough rate corresponding to that ad#) and the number of unique impressions in which the query appears.

For a given query q in each set (train, test, or validation), let imp_q^i denote the number of times query q appears in the impressions with i number of ads. In Equation 1, we estimated the average ad clickthrough rate for such a query as CTR_i. Thus, the estimated number of ad clicks for such a query is calculated as follows:

$$click_q = \sum_{i=1}^{8} CTR_i \times imp_q^i \tag{2}$$

We simply pass through all the impressions for a query and multiply the number of impressions with a particular number of ads by the value of the obtained average clickthrough rate corresponding to that particular number of ads. As the ads number varies from one to eight, we add the eight different values for each query together in order to estimate the total click number for the query.

Figure 4 depicts the actual number of clicks versus the estimated number of clicks in the *test* set. The plot is presented in log-log basis. As can be seen in the plot, 55% of the queries appear above the line $y = x$ in the plot, meaning that their actual number of clicks is greater than that which was estimated for them. The plots for the other two sets (train and validation) follow the same pattern as the test

set, so we do not present them here. As is shown in the figure, the predicted number of clicks and the actual number of clicks are correlated, particularly as the number of clicks increases. This observation could indicate that the number of ads actually determines the number of clicks, at least in part. We further study this issue later in the paper, when we discuss query intent.

5. CLASSIFYING QUERY INTENT

As mentioned previously, two dimensions of query intent are studied in this work: commercial/noncommercial and navigational/informational. We utilize decision trees for our intent classification process and stick to a set of features extracted from our large sets of impression and ad clickthrough data. In similar work with sponsored search data by Richardson et. al [13], some features of the ad itself, such as the structure of the landing page and bid terms, were used in order to estimate the clickthrough rate for a given ad. Since we are concerned with the query intent, we limited our feature set to those related to queries and their ad clickthrough information.

As Lee et. al in [11] suggest, predicting user query goals can be performed based on the two type of feature sets, past user-click behavior and anchor-link distribution. We considered similar types of features based on what we had available in the given ad clickthrough data. Some of the features were first normalized and then fed to the classifiers for both types of query intents. The set of features used for each query are as follows:

- The query length in terms of the number of characters in a query.
- A feature, namely *URL-element*, is set to 1 if the query has any URL element, such as .com, .org, .ca, and etc, otherwise it is set to 0.
- Number of target hosts is calculated as the number of different ad links that were clicked for the query.
- Average number of clicks per target host (namely *avg*, which is equal to the number of ad clicks for the query divided by the number of different hosts clicked for that query).
- Significance of the query's most frequent target host (the number of times a click happens on the most frequent target host as a result of the query, divided by *avg*).
- The level of decrease in clicks between the top two frequent target hosts for the query (the number of times click happens on the most frequent target host as a result of the query divided by the number of times the top second frequent target host receives click as a result of the same query).
- Click rate defined as the ratio of the total number of ad clicks resulting from the query against the total number of impressions in which the query appeared.
- Number of target hosts of which the query string is a substring divided by the total number of different hosts clicked for that query.

Table 1: Prediction Accuracy

Classifier	Query Intent	Precision	Recall	Accuracy
A	Commercial	0.77	0.61	72.5%
	Noncommercial	0.70	0.82	
B	Navigational	0.82	0.87	83.27%
	Informational	0.85	0.79	

- Total number of clicks on target hosts of which the query string is a substring divided by the total number of ad clicks for that query.
- The difference between the query impression time and its ad click time on average (the deliberation time).

The above features have been extracted for the 45,032 queries of the test set, and the 44,909 queries of the validation set, and also the 1000 labeled queries selected from the train set. The 1000 labeled queries along their features were first fed to a C4.5 decision tree (using the WEKA tool [7]) in order to train the classifier (separately for each dimension of query intent). We applied the 10-fold cross validation method on the labeled queries to measure the accuracy of our classifiers. A report of the prediction accuracy for the commercial/noncommercial classifier (A) and the navigational/informational classifier (B) is presented in Table 1. Afterwards, the test and validation queries (total of 89,941 unlabeled queries) were passed to each classifier in order to predict their types.

As a result of this intention classification, each query will fall into one of the following four intention categories: i) commercial and navigational, ii) commercial and informational, iii) noncommercial and navigational, and iv) noncommercial and informational. We will consider the first two of these 2-dimensional intentions for each query in our future analyses in order to characterize different queries commercial intent according to their ad clickthrough information.

6. CLICK PREDICTION BY NUMBER OF ADS AND QUERY INTENTS

The average clickthrough rate for particular number of ads should be different for various types of queries. In this section, we study this issue and use it as a basis for predicting the number of ad clicks for a given query.

6.1 Click to Impression Ratio by Query Intent

At this point, we follow a similar approach to what we did in calculating the average click to impression ratio for all the impressions with particular number of ads in the training set. However, this time, we consider only the impressions for which their associated queries are of a particular type. Note that we calculate the ratio values for the queries in the *training* set, and later we will use these values associated with particular number of ads and query intent in order to estimate the number of ad clicks for queries in the *test* set.

The average clickthrough rates for the four general types of queries (i.e. commercial, noncommercial, navigational, and informational) with the particular number of ads are plotted in Figure 5. The same analysis is performed for two pairs of query types (queries that are either commercial-navigational or those that are commercial-informational),

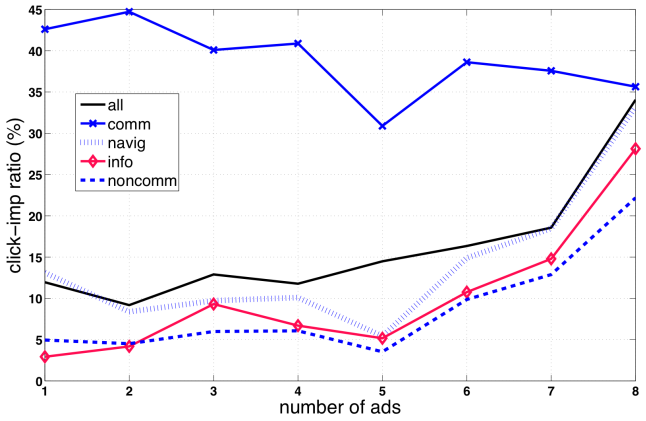


Figure 5: Average Click to Impression Ratio for Commercial, Noncommercial, Navigational, Informational, and All Types of Queries at a Particular Number of Ads (lines do not imply interpolation)

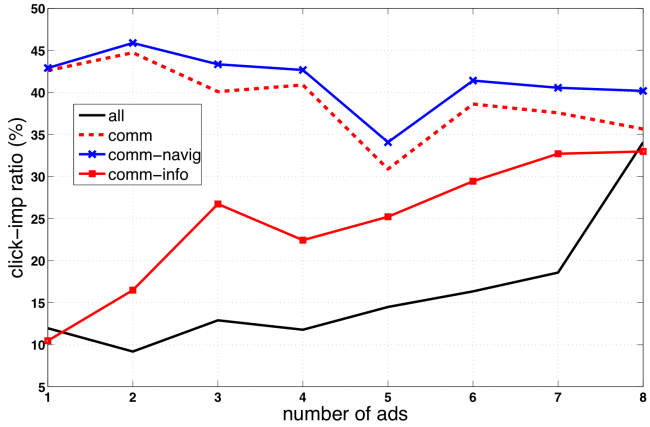


Figure 6: Average Click to Impression Ratio for Commercial, Commercial-Navigational, Commercial-Informational, and All Types of Queries at a Particular Number of Ads (lines do not imply interpolation)

which results in the two plots depicted in Figure 6 along with the plot for general commercial queries. The plot from Figure 3 is also placed in Figures 5 and 6 to provide a baseline for comparison. The plots for noncommercial-navigational and noncommercial-informational queries are not presented, because we are interested in studying the behavior of commercial queries and the effects of navigational and informational intent on user behavior for these commercial queries.

According to both figures, for most of the query types, the more ads are displayed as results of a query, the more clicks they receive. Moreover, note that the commercial query type is the leader (Figure 5) in terms of click frequency for all number of displayed ads. This frequency is larger (Figure 6) when the commercial query is also navigational rather than informational.

There are some peaks and valleys in the plots that could be because of the location of different ads (top or side of the result page) for which the clicks are recorded. According to Jansen [8], top-listed ads are assumed to be more relevant than organic results and side-listed ads. This could affect the frequency of clicks for ads at different locations and could be the cause of bumps at some points of the plots. The location of the ads is not available to us, however we believe it should be the subject for further study on this issue.

As depicted in Figure 5, navigational queries receive more ad clicks than informational queries on average. Similarly, Figure 6 conveys that commercial-navigational queries receive more ad clicks than commercial-informational queries on average. Our intuition for explaining both observations would be the fact that a query is navigational restricts the top result links (either ad or organic results) to a particular website. Therefore, the top results for a navigational query would more likely match with what user is seeking. This could result in more ad clicks for navigational queries in comparison to the informational queries. The difference is even larger when the user's intent is also commercial, because the target of click for a commercial query (in this case, commercial-navigational) is most likely for an ad. This could make a larger difference in number of ad clicks between the commercial-navigational queries and the commercial-informational queries as compared to the navigational queries and the informational queries in general.

A good example for illustrating this difference is "American airlines" as a commercial-navigational query against "airline tickets" as a commercial-informational query. The chance that user would find a related ad for the former query is greater than the later one, because the former query is restricted by the airline name. As Jansen suggests [8], searchers approach e-commerce searching from two major perspectives, one to look for a specific product or service, and the other to detect information. We believe that the commercial-navigational queries fall mostly into the former category, while the commercial-informational queries fall into the latter one.

Figure 6 shows similar clickthrough rates for commercial and commercial-navigational queries with one displayed ad (and also for two ads), while the rate for the commercial-informational queries is much smaller. We looked into this issue and found that out of 160,040 impressions for the commercial queries with one displayed ads in the training set, only 1,462 belong to commercial-informational queries while the rest (158,578 impressions) belong to the commercial-navigational queries. Moreover, 153 clicks were recorded for 1,462 impressions with one displayed ads for the commercial-informational queries (i.e. 10.5 % clickthrough rate). These numbers are 68,025 out of 158,578 (i.e. 42.9 % clickthrough rate) for the commercial-navigational queries which are relatively high.

We hypothesize that entering a navigational query (in this case, commercial-navigational query) results in a specific highly-related page. Hence, if only one ad is supposed to be displayed for such a query, it will most likely be the same as (or highly related to) that single page. Therefore, the impressions for which only one ad is listed correspond most closely to the commercial-navigational queries rather than commercial-informational queries. In other words, comparing to the commercial-navigational queries, the commercial-informational queries provide more chance so that various ads will be displayed as the result of the queries.

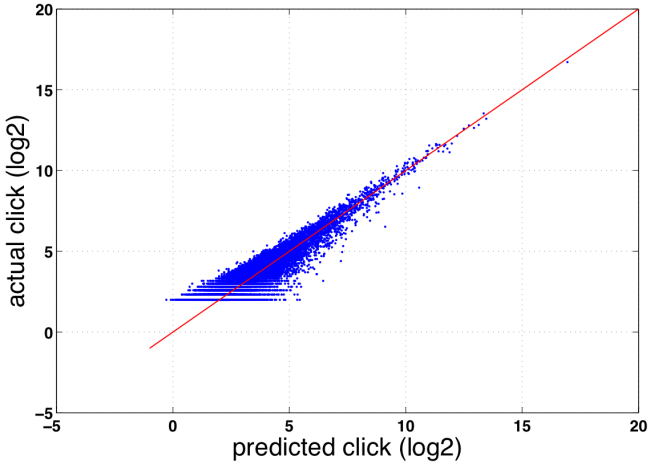


Figure 7: The Actual Number of Clicks vs. the Estimated Number of Clicks for the Commercial Queries

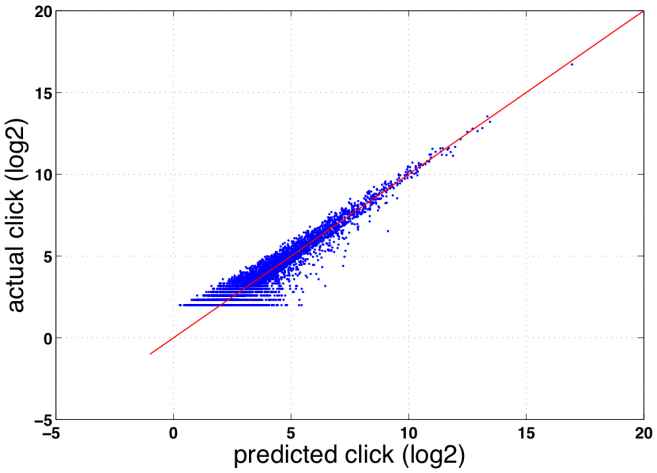


Figure 8: The Actual Number of Clicks vs. the Estimated Number of Clicks for the Commercial-Navigational Queries

6.2 Click Prediction

Using the average clickthrough rate obtained for the two dimensions of query types from the training set, we now focus on the test set. Recall that all the queries in this set have 4 or more clicks recorded for them in the click data. Also note that we are generally interested in behavior of commercial queries in the domain of sponsored search (commercial-navigational and commercial-informational queries, more specifically).

Let CTR_i^{cn} and CTR_i^{ci} be the average clickthrough rates for impressions with i number of ads that belong to commercial-navigational and commercial-informational queries respectively. Similarly, let CTR_i^{nn} and CTR_i^{ni} be the average clickthrough rates for impressions with i number of ads that belong to noncommercial-navigational and noncommercial-informational queries respectively. For a given query $q \in Q$, where Q is the set of all queries, we define function t as $t : Q \rightarrow T$. T is the set of pairs of query intents which are among the followings: commercial-navigational, commercial-informational, noncommercial-navigational, and noncommercial-informational. According to our previous notation, we consider $T = \{\text{cn, ci, nn, ni }\}$. Based on what we just defined, the proposed prediction strategy obtains the query intents using the function t. It then uses the average rate corresponding to that category of intents in order to calculate the estimated number of clicks by going through all the impressions (similar to the Equation 2) of the query:

$$click_q^{int} = \sum_{i=1}^{8} CTR_i^{t(q)} \times imp_q^i \qquad (3)$$

where $click_q^{int}$ is the estimated number of clicks based on the proposed prediction model which considers the average clickthrough rate for different query intents.

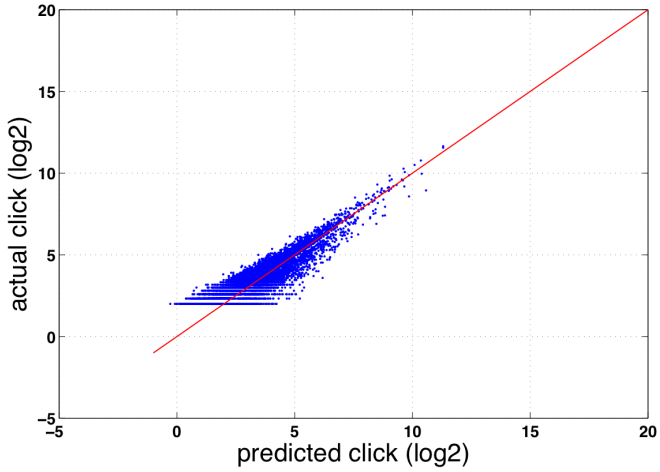


Figure 9: The Actual Number of Clicks vs. the Estimated Number of Clicks for the Commercial-Informational Queries

The plots for the commercial queries are presented in Figure 7. As is shown in the figure, the predicted number of clicks and the actual number of clicks are more correlated than the baseline depicted in Figure 4. We measured the correlation for each plot by calculating the covariance of the two data sets (the predicted clicks versus the actual clicks), where a perfect prediction with all the points on the line $y = x$ would result in correlation equal to 1. The correlation for the plots in Figure 4 is calculated as 0.786, while the one for the commercial queries (Figure 7) is 0.927. This may indicate that the number of ads represents a major factor in determining the number of clicks for commercial queries.

To further study the effectiveness of the number of ads in such an intention-based prediction, we plotted the actual number of clicks versus the predicted clicks for commercial-navigational and commercial-informational queries in Figures 8 and 9 respectively. These two plots confirm our previous statement that commercial-navigational queries receive on average more ad clicks than commercial-informational queries. Moreover, as depicted in Figure 8, the actual number of ad clicks are greater than the number predicted by our model for most of the commercial-navigational queries. This could indicate that the number of ads determine the number

of ad clicks for commercial-informational queries more effectively than queries that are commercial and navigational. However, the covariance measure reports a slightly lower correlation for the former one compared to the later one (0.903 versus 0.950). Investigating the reason behind these observations is a direction for future work.

7. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we develop a methodology to use ad clickthrough logs from Microsoft adCenter Search ad click logs in order to study characteristics of different query intents. The findings of our study suggest that ad clickthrough features, such as the deliberation time between entering a query and clicking on an ad for that query, are effective in detecting different query intents. Utilizing this feature, along with other query-based and ad clickthrough features, we trained a decision tree to classify queries in two dimensions: commercial/noncommercial and navigational/informational. The average clickthrough rate is then estimated for different query types. The obtained rates were used to predict clickthrough rate for a given query with particular intentions and various number of ads (one to eight) displayed as the result of the query. All in all we can list our findings as follows:

- Users spend more time on average for noncommercial queries than commercial queries to find the related ad (if any) to click on.
- Users spend more time on average for informational queries than navigational queries to find a related ad to click on.
- According to the accuracy of our decision tree based classifiers (Table 1), ad clickthrough-based features correlate with query categories: commercial/noncommercial and navigational/informational.
- For most of the query types, the more ads displayed as results of a query, the more clicks they receive. However, commercial-navigational queries receive more ad clicks for all number of displayed ads compared to other types of queries.
- The number of displayed ads affects the number of ad clicks for each category of query intents differently. It seems this factor is more effective in predicting the ad clickthrough rate for commercial queries, especially the commercial-informational ones, compared to the others.

A possible future direction for this work is studying the organic clickthrough behavior as to whether it follows the behavior we have reported in this paper. As mentioned before, the average click to impression ratio for some query types has some bumps at particular number of ads (usually five). It is worth looking into this issue, to determine if the location of the clicked ads has anything to do with this observation. Another possible direction for this work would be to study the possibility of whether the clickthrough data can be exploited in the labeling process. Finally, the reasons behind these behaviors can be further explained.

8. ACKNOWLEDGMENTS

We would like to thank Microsoft Research and Microsoft adCenter for the Beyond Search data sets, and for partially supporting this work.

9. REFERENCES

[1] R. Briggs and N. Hollis. Advertising on the web: Is there response before clickthrough. *Journal of Advertising Research*, 37(2), 1997.

[2] A. Broder. A taxonomy of Web search. *ACM SIGIR Forum*, 36(2):3–10, 2002.

[3] H. Dai, L. Zhao, Z. Nie, J. Wen, L. Wang, and Y. Li. Detecting Online Commercial Intention (OCI). *Proceedings of the* 15^{th} *International Conference on World Wide Web*, pages 829–837, 2006.

[4] K. Debmbsczynski, W. Kotlowski, and D. Weiss. Predicting ads clickthrough rate with decision rules. *Workshop on Target and Ranking for Online Advertising*, WWW 2008.

[5] D. C. Fain and J. O. Pedersen. Sponsored search: A brief history. *Bulletin of the American Society for Information Science and Technology*, 32(2):12–13, 2006.

[6] A. Ghose and S. Yang. An empirical analysis of sponsored search performance in search engine advertising. *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 241–250, 2008.

[7] G. Holmes, A. Donkin, and I. Witten. WEKA: A machine learning workbench. *Proceedings of the* 2^{nd} *Australian and New Zealand Conference on Intelligent Information Systems*, pages 357–361, 1994.

[8] B. Jansen. The comparative effectiveness of sponsored and nonsponsored links for Web e-commerce queries. *ACM Transactions on the Web*, 1(1), 2007.

[9] B. Jansen, D. Booth, and A. Spink. Determining the user intent of Web search engine queries. *Proceedings of the* 16^{th} *International Conference on World Wide Web*, pages 1149–1150, 2007.

[10] B. Jansen, A. Brown, and M. Resnick. Factors relating to the decision to click on a sponsored link. *Decision Support Systems*, 44(1):46–59, 2007.

[11] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in Web search. *Proceedings of the* 14^{th} *International Conference on World Wide Web*, pages 391–400, 2005.

[12] M. Regelson and D. Fain. Predicting clickthrough rate using keyword clusters. *Proceedings of the* 2^{nd} *Workshop on Sponsored Search Auctions*, 2006.

[13] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the clickthrough rate for new ads. *Proceedings of the* 16^{th} *International Conference on World Wide Web*, pages 521–530, 2007.

[14] D. Rose and D. Levinson. Understanding user goals in Web search. *Proceedings of the* 13^{th} *International Conference on World Wide Web*, pages 13–19, 2004.

Domain-Specific Query Augmentation using Folksonomy Tags: the Case of Contextual Advertising

Andrei Broder Peter Ciccolo Evgeniy Gabrilovich Bo Pang
Yahoo! Research, 701 First Ave, Sunnyvale, CA 94089.
{abroder,gabr,ciccolo,bopang}@yahoo-inc.com

ABSTRACT

Folksonomies allow users to collaboratively tag a variety of textual and multimedia objects with sets of labels. The largest folksonomy projects, such as FLICKR and DEL.ICIO.US, contain millions of multi-labeled objects, and embed significant amounts of human knowledge. We propose a method for automatically using this knowledge to augment traditional IR systems, using contextual advertising as an application domain. Given a query, we first identify a set of relevant tags, and then use tags that cooccur with them to augment the query. Importantly, our method performs domain-specific query disambiguation, and can actually learn that a query "menu" is likely to have food connotation on FLICKR but user interface connotation on DEL.ICIO.US.

1. INTRODUCTION

Folksonomy is a method for assigning user-defined labels to objects stored in public repositories of textual or multimedia content. Examples of popular folksonomies include FLICKR (a photo collection), DEL.ICIO.US (a bookmark sharing project) and YOUTUBE (a video sharing system). Typically users can add tags to any object, whether they "own" it or not. Folksonomies facilitate interaction between Web users and promote knowledge sharing by integrating the user-defined tags in searching and browsing activities. In a sense, folksonomies comprise a competing approach to restricted lexicons, as the numerous labels potentially allow users to achieve higher recall. When the original content creator might not have thought of all the applicable tags, users who subsequently encounter the object are likely to add tags they deem relevant.

Some tags are automatically assigned (e.g., a FLICKR picture can be automatically labeled with the camera model and geographical location of the pictured scene), but the majority of tags are assigned manually by Web surfers. For example, a Flickr photo of an elephant could be labeled with tags such as *Thailand, Asia, colorful* and *sit on elephant back*. While some tags are only meaningful to their creator, many are useful to other users. Consequently, folksonomies encode a cornucopia of human knowledge, and in this paper we propose a method for leveraging this knowledge to achieve better focus in information retrieval.

In particular, we use co-tagging (i.e., tagging of the same object with different tags) to infer tag relatedness *in the context* of individual folksonomies. Prior studies in contextual information retrieval mainly defined context as a fragment of natural language text surrounding the object in question. We propose an alternative definition of context as a collection of tags assigned to or related to an object. Such contexts can be quite different between folksonomies, and can serve for word sense disambiguation. For instance, studying tag cooccurrence reveals that on FLICKR the word "menu" mostly refers to food and restaurants, but on DEL.ICIO.US it often describes elements of graphical user interface.

In this paper, we use sponsored search advertising as our application domain, where our aim is to match search queries in different folksonomies to most relevant textual ads. Besides the obvious commercial incentive in placing more relevant ads, judging the relevance of textual ads to a textual query is simpler than judging the relevance of, say, pictures or movies, and thus the relevance of ads provides a convenient means of validating our approach.

It is also of great interest to study the effect of returning site-specific ads for a given query. A query submitted to FLICKR most likely conveys a different intent of the user than the same query submitted at DEL.ICIO.US. That is, knowing at which site the query is submitted can help identify the search intent of user. Treating the content of the site as the context for queries and matching ads accordingly can potentially improve user experience. In the previous example, FLICKR ads for the query "menu" should ideally include offers from restaurants rather than services of UI experts, which would be more appropriate on DEL.ICIO.US.

Matching ads to short queries is challenging, and in mainstream information retrieval query expansion techniques are often used to augment queries with additional terms or concepts based on some form of relevance feedback [6, 14], dictionary lookup [13], ontological classification [3], or electronic encyclopedias [2]. However, to the best of our knowledge, no prior studies examined the use of folksonomies as an alternative source for query augmentation or explored the effect of using the content of a vertical site as the context for queries submitted to that site.

We propose a way to use tag cooccurrence statistics for site-specific query augmentation. Specifically, we use relevant tags to expand the bag of words for the query, as well as classify those tags to create new taxonomy-based features. Representing queries in this rich feature space results in more relevant ad matches, so that the ads displayed on different folksonomy sites better reflect the intent of their users. We present the results of an initial evaluation of the proposed method. The performance of our method is competitive with that of query expansion based on Web search results, and is superior to it at low recall (i.e., in the high precision region). We also analyze the difficulties of such evaluation, when judges needed to adopt the mindset of users of different folk-

sonomies (e.g., FLICKR and DEL.ICIO.US).

2. BACKGROUND

Folksonomies.

Tagging systems allow users to annotate a variety of resources with textual labels, or tags, which could be individual words or phrases [5, 4]. The term "folksonomy" is a *portmanteau* of "folk" and "taxonomy" and is due to Thomas Vander Wal [11]. Folksonomies provide a scalable way to collect metadata about objects; in fact, one of the first tagging projects, the ESP Game [12], was designed to collect tags to facilitate retrieval of images. Many folksonomies double as social networks, where users are grouped either explicitly by interests or explicitly by their tagging behavior.

Online textual advertising.

A large part of the Web advertising market consists of *textual ads*. There are two main channels for distributing such ads. *Sponsored search* places ads on the result pages of a Web search engine, where ads are selected to be relevant to the search query. *Content match* places ads on third-party Web pages, which range from individual bloggers and small niche communities to large publishers such as major newspapers.

In this work we focus on sponsored search, where a few carefully-selected paid textual ads are displayed alongside algorithmic search results. Identifying relevant ads is challenging because a typical search query is short and because users often choose terms to optimize Web search results rather than ads. There is a fine but important line between placing ads relevant to the query and placing unrelated ads. Users often find the former to be beneficial as an additional source of information or Web navigation, while the latter annoy the searchers and hurt the user experience.

Sponsored search is an interplay of three entities. The **advertiser** provides the supply of ads; as in traditional advertising, the goal of the advertisers is to promote products or services. The **search engine** provides "real estate" for placing ads (i.e., allocates space on search results pages), and selects ads that are relevant to the user's query. **Users** visit the Web pages and interact with the ads.

Search engines select ads based on their expected revenue, computed as a probability of a click times the advertiser's bid. However, in this paper we focus on ad textual relevance only. Several prior studies examined the textual aspects of relevance in sponsored search. For instance, people have looked into predicting click through rate based on keywords in queries as well as content of ads [8, 9, 7]. To the best of our knowledge, there has not been previous work that considers the site-specific nature of ads placement.

3. METHODOLOGY

We now present our methodology for using folksonomies for site-specific query augmentation. The input to our system is a search query, and the output is a set of ads that are relevant to this query. Processing the input query involves two main phases. First, given a query, we identify a set of relevant tags, and then identify tags that cooccur with them. We then pool these tags together in a *context vector*, i.e., a vector of tags whose individual entries are weighted by cooccurrence frequency. Second, we use the context vector to construct an augmented *ad* query, to be executed against a corpus of ads. The features of the ad query include an augmented bag of words and a set of taxonomy classes. We now describe these two phases in detail.

3.1 Building context vectors

Tags used to label the same object (an image in FLICKR, or a Web page in DEL.ICIO.US) are often semantically related words or phrases, as they represent different aspects or characteristics of the same object. Tag cooccurrence information aggregated over all the objects in a folksonomy reflects site-specific relatedness as defined (and shared) by its users. In the preprocessing phase, we try to capture this information by analyzing the set of objects in a folksonomy $\mathcal{F}$, and build a tag cooccurrence matrix M, where $M(i,j)$ is the number of objects co-tagged with tags t_i and t_j. To reduce noise, we ignore all cells such that $M(i,j) < 2$.

To construct the context vector for an input query, we tokenize the query into words, and then map the words into relevant tags. For each tag t_i, we look up its cooccurrence vector, namely, a row $M(i)$, and finally sum the retrieved vectors to obtain a single *context vector* V for the query. We decimate the vector entries by retaining only the n most frequently cooccurring tags ($n = 10 \ldots 100$). The values of individual vector entries are assigned using the TFIDF scheme [10], with logarithmic term frequency and IDF computed over the ad corpus.

We now address two research questions involved in this process, namely, how to handle multi-word tags and queries.

Mapping the tag space into the word space.

Many tags contain several words (e.g., "sanfrancisco" or "ToRead"). This does not pose problems for building the tag cooccurrence matrix M as this type of concatenation is a convention of the tagging system (indeed, some folksonomies automatically remove white spaces in phrases for each individual tag). However, it is problematic to use such multi-word tags for query augmentation since such concatenations are not common in the ad corpus, and as a result they are unlikely to improve the ad matching process. To this end, we use a dynamic programming algorithm (based on a unigram language model trained on the ad corpus) to break tags into individual words, and update the counts in V accordingly.

If a tag t_j is segmented into k tokens $t_{j,1}, ..., t_{j,k}$, we need to decide how to distribute the counts aggregated for t_j among these tokens. We considered two different options: each token receives the same count as the original tag, or only a portion thereof. More specifically, we compute a count $c(j,p)$ for token $t_{j,p}$ based on $M(i,j)$. If we consider each of the segmented tokens as a tag in itself, then each of them would have cooccurred with t_i $M(i,j)$ times, which suggests setting $c(j,p) = M(i,j)$. On the other hand, if we consider each tag to have the same importance for a given object, then each of the tokens on its own would not have cooccurred with t_i with the same likelihood, and one way to approximate this is to set $c(j,p) = M(i,j)/k$. Based on examining context vectors in the development set, we implemented the second option in our system.

Handling multi-word queries.

Building context vectors for multi-word queries is challenging, because some word combinations have meanings that are different from a simple composition of the meanings of constituent words. One possibility is to map each word into the closest tag and consider different ways to combine the context vectors retrieved for these individual tags. If we consider all the words in a query as context for each other, which can be employed to achieve further disambiguation, we should take the intersection of the vectors retrieved to represent the "common" context vector. Alternatively, if we consider each word as enrichment to other words in the query, we can take the sum over all the context vectors retrieved. The dataset we used in this work only contained a few multi-word queries, hence for simplicity we mapped each multi-word query into a single tag by taking out the white spaces. In future work, we are interested in exploring the effect of different strategies of combining context vectors where each constituent words in the query will be mapped into the closest tag instead of being concatenated into one single tag.

3.2 Retrieving ads

We now discuss how to use the context vector to construct an augmented *ad* query to be executed against a corpus of ads. Ad queries are represented with two kinds of features. We use feature selection to identify most salient words in the context vector V, and use the selected features to augment the bag of words representation of the original (short) query (with stop words removed). We also consider the context vector as a pseudo-document, and automatically classify it with respect to a large commercial taxonomy of over 6000 nodes. Previous work found it beneficial to include class information in ad retrieval [1, 7], as generalizing from individual words to classes allows one to match related queries and ads even though they might use different vocabularies. Furthermore, classifying the query context with respect to an external taxonomy introduces yet another valuable source of external knowledge. We adopted the taxonomy used in [1]; further details on the taxonomy are available therein. The 5 most relevant class nodes for each query, along with their ancestors, comprise a second kind of features. Our experiments confirmed previous work and found class information to be useful in our site-specific setting as well.

We analyze the ad text and construct the same two types of features as for queries, namely, words and classes. In an online advertising system, the number of ads can easily reach hundreds of millions, hence we use an inverted index to facilitate fast ad retrieval. Finding relevant ads for the query amounts to evaluating the scores of candidate ads, and then retrieving the desired number of highest-scoring ads. We compute query-ad scores as a linear combination of cosine similarity scores over the two feature sets.

4. EVALUATION

We implemented the above methodology for site-specific query augmentation in a software system called Alexandrite[1].

4.1 Editorial evaluation

Dataset.

We evaluated Alexandrite on two actual folksonomies, FLICKR and DEL.ICIO.US, while our hypothesis was that taking site-specific tagging patterns into account would allow us to match queries on each site to more relevant ads. We constructed the dataset by taking a set of most frequent queries from each site, as well as a set of queries with most different meaning (as judged by comparing their context vectors V defined in Section 3). After removing duplicates and adult queries, we ended up with 492 queries, of which about 10% contained more than one word. We held out 92 queries as a validation set to tune parameters, and the remaining 400 queries formed the test set.

Reference systems.

We compared Alexandrite with two other systems. The first one was a baseline system that did not use any site-specific information and implemented a generic Sponsored Search (SS) algorithm, which expanded queries with general purpose Web search results (Citation anonymized). Naturally, this baseline returns the same set of ads for both sites.

We also compared Alexandrite to a "site-aware" system, which used site-specific search results instead of those from general Web search. This approach is akin to so-called Content Match (CM) advertising scenario, where ads are matched to Web pages instead of queries. This system (referred to as CM in the sequel), used the input query to conduct a regular search on either FLICKR or DEL.ICIO.US, and then used the results page to build a rich ad query. Similarly to Alexandrite, both SS and CM systems represented ad queries and ads in the space of words and classes.

We implemented Alexandrite with the following parameters. Two parameters control the relative importance of words vs. classes in the augmented vector. We considered emphasizing only classes or words individually, as well as placing equal importance on both types of features. Another parameter controls the number of cooccurring tags to include in the context vector. We augmented queries with up to n most frequently cooccurring tags from M, and considered $n = 10, 20, 50, 100$.

Judging with FLICKR *or* DEL.ICIO.US *mindset.*

For each system, we matched each query to up to 3 ads for each of the two sites. We obtained human judgments for each query-ad pair on the following numeric scale: Perfect (0), Certainly Attractive (1), Probably Attractive (2), Somewhat Attractive (3), Probably Not Attractive (4), and Certainly Not Attractive (5). To compute the standard metrics of precision and recall, we converted the above judgments to binary by considering the first four as relevant, and the rest as irrelevant.

The query-ad pairs were judged by editors who are trained in conducting relevancy evaluations. They were not aware of the algorithmic details, and all the query-ad pairs were presented to them in random order. In order to evaluate how well our system can capture the site-specific context, we asked the editors to adopt the mindset of a typical FLICKR or DEL.ICIO.US user. Search result pages on each site were provided to help the editors better understand the scope of each site. The editors were also instructed to use Web search if they required additional information about the meaning of the query or about the products and services described in the ads.

4.2 Pilot study

One potential concern about the validity of our approach is that its utility may be limited by the available ad inventory. Even if our technique does model the site-specific context reasonably well, and the context vector does a good job of capturing site-specific user intent, if the ad inventory does not contain ads that reflect such differences, we will not be able to distinguish between the results produced by the different systems.

To assess the importance of this concern, we first conducted a pilot study with a set of single-word queries that exemplified different user intent in the two sites. Our goal was to verify whether there are any differences in the top ads returned for such queries in the two sites, and if so, whether the differences are consistent with an intuitive interpretation of the intentions of typical FLICKR or DEL.ICIO.US users. Table 1 presents a subset of queries with sample ads retrieved by Alexandrite. Indeed, the sample ads seem to be consistent with our intuition about FLICKR as a fairly general site and DEL.ICIO.US as a geek-oriented site with more technical content.

4.3 Results

Table 2 summarizes the average numeric scores for the different systems we evaluated (lower values correspond to more relevant ads and are better). For Alexandrite, the editorial judgment confirmed our expectation that the best performance is achieved by using both types of features (namely, words and classes), and taking the 50 most frequent tags for the context vectors.

Site \ Method	SS	CM	Alexandrite
FLICKR	3.88	3.95	4.09
DEL.ICIO.US	3.495	3.50	3.485

Table 2: Average system scores (at maximum recall)

Based on these preliminary results, Alexandrite performance is competitive with that of the two reference systems. Importantly,

[1]Alexandrite is a semi-precious stone that changes its color under different lighting conditions.

Query	Ads for FLICKR	Ads for DEL.ICIO.US
menu (table)	Online Restaurant Menu / Food Service Consultant	Quickly Learn HTML Web Site Design
sun	Sun 'n Sea Sunset Waters Beach Resort	Solaris (Sun) Training java
fly	Fly Fishing Shop	Blue sky air / airplane
mouse	Disney Mickey Mouse Items	Wireless Keyboard

Table 1: Sample Alexandrite output

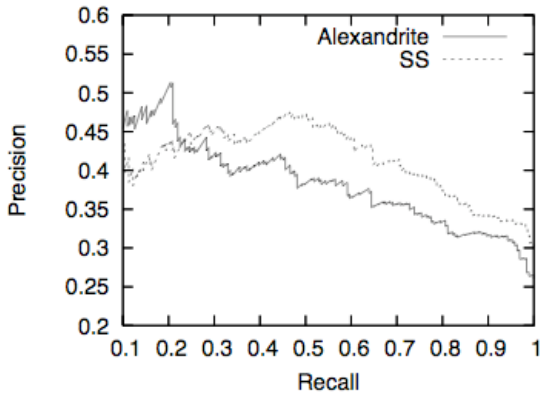


Figure 1: Alexandrite vs. SS

Alexandrite performs most tag cooccurrence analysis in the preprocessing phase, and is thus more efficient than both SS and CM, which involve a query-dependent search on the Web or the folksonomy.

Table 2 compares the systems at maximum recall. On a precision-recall graph produced by thresholding the ad retrieval scores (Figure 1), we observed that in the low recall (= high precision) range, the precision of our algorithm is superior to that of SS. We also experimented with different lengths of the context vector ($n = 10 \ldots 100$ tags), and predictably found $n = 50$ to yield optimal results. Lower values of n under-utilized available context, and higher values resulted in using less reliable tags owing to noise (we omit the graph for lack of space).

It is essential to note that the editors reported the task of adopting the mindset of FLICKR and DEL.ICIO.US users to be quite difficult, which partly explains why in our preliminary evaluation Alexandrite did not definitively outperform the baselines. For instance, for the query "Antarctica" on DEL.ICIO.US, Alexandrite returned a Web design ad, which was judged as Certainly Not Attractive. However, this particular ad offered the services of Antarctica Media company, which specializes in Web design, and hence should have arguably been scored much better. While FLICKR content is quite general, DEL.ICIO.US caters to the tech-savvy geek community, hence adopting the mindset of DEL.ICIO.US users was particularly difficult.

Also noteworthy is the disparity of scores for the SS system on the two sites (see Table 2). This system expanded queries using general Web search results (without any site-specific information), and hence we would expect its output ads to be more relevant for the more general FLICKR site. However, its ads have been judged more relevant (= lower score) for DEL.ICIO.US, which again reinforces our concern about the difficulty of judgment by adopting a particular mindset. Furthermore, some queries are indeed hard to judge for non-expert users. In our future work, we plan to improve our judgment procedure, and also evaluate the system by conducting an experiment with actual users, measuring the relevance of ads by actual click-through rates.

5. DISCUSSION

We proposed a methodology for using folksonomy tags for query augmentation in one IR task (sponsored search advertising). Our approach leverages co-tagging data to capture site-specific query intent and to disambiguate polysemous queries. Although we focused on sites with rich tagging information, the methodology proposed could also be applied to other sites by modeling site-specific distribution of words. Our initial evaluation confirmed that the proposed method is competitive with another system that performs query augmentation based on site-specific search results (CM). We also discussed inherent judging difficulties when editors are asked to adopt mindsets of typical users of particular Web sites. In our future work, we plan to further refine our method and to revise the editorial evaluation, as well as to perform a real-life evaluation of Alexandrite with actual folksonomy users and evalute the system with user-generated click data.

6. ACKNOWLEDGMENTS

We thank Vanja Josifovski, Ravi Kumar, and Malcolm Slaney for fruitful discussions, technical assistance, and pointers.

7. REFERENCES

[1] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. A semantic approach to contextual advertising. In *SIGIR'07*, pages 559–566. ACM Press, 2007.

[2] Ofer Egozi, Evgeniy Gabrilovich, and Shaul Markovitch. Concept-based feature generation and selection for information retrieval. In *AAAI'08*, July 2008.

[3] Evgeniy Gabrilovich and Shaul Markovitch. Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *JMLR*, 8:2297–2345, October 2007.

[4] Scott Golder and Bernardo Huberman. The structure of collaborative tagging systems. Technical report, HP Labs, 2005.

[5] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. In *HT'06*, 2006.

[6] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *SIGIR'98*, 1998.

[7] Filip Radlinski, Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. Optimizing relevance and revenue in ad search: A query substitution approach. In *SIGIR'08*, 2008.

[8] M. Regelson and D. Fain. Predicting click-through rate using keyword clusters. In *Second Workshop on Sponsored Search Auctions*, 2006.

[9] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *WWW'07*. ACM Press, 2007.

[10] Gerard Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[11] Thomas Vander Wal. Folksonomy definition and Wikipedia. `http://www.vanderwal.net/random/entrysel.php?blog=1750`, 2005.

[12] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *CHI'04*, 2006.

[13] Ellen M. Voorhees. Using wordnet for text retrieval. In Christiane Fellbaum, editor, *WordNet, an Electronic Lexical Database*. The MIT Press, 1998.

[14] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM TOIS*, 18(1):79–112, 2000.

Understanding "Abandoned" Ads: Towards Personalized Commercial Intent Inference via Mouse Movement Analysis

Qi Guo, Eugene Agichtein
Emory University, United States
{qguo3, eugene}@mathcs.emory.edu

Charles L. A. Clarke, Azin Ashkan
University of Waterloo, Canada
{claclark, aashkan}@cs.uwaterloo.ca

ABSTRACT

Clickthrough on ads and search results have been successfully used to infer user interest and preferences, but these indicators are typically most effective for modeling the "dominant" or most popular intent for a query. In this paper we begin to explore rich client-side instrumentation for inferring *personalized* commercial intent of users. In particular, we investigate whether mouse movement over search results can provide clues into the users' intent. As one practical application, we attempt to understand the causes of "abandoned" — unclicked — ads: that is, to automatically distinguish whether a user's search had no commercial intent at all, or if the commercial intent was present, but the ads were not sufficiently relevant. Our preliminary results indicate that in some cases mouse movement analysis can help distinguish these cases, providing information about user intent not previously available.

1. INTRODUCTION

As online advertising continues to fuel the internet, detecting and categorizing commercial intent behind user online activities becomes increasingly important. Web search engines in particular are supported largely by search advertising, where the ability to predict whether the searcher's intent is relevant to commerce is crucial. However, inferring user intent is challenging, since search queries are often ambiguous and may reflect diverse intents and information needs. That is, different users may have different needs expressed via the same search query. Furthermore, even the same user may issue the same query with different goals. For example, as searchers increasingly use queries as bookmarks [14], the same user may issue the same (previously informational intent) query with a navigational/re-finding goal.

Clickthrough on organic search results and search ads have been successfully used to infer user interest and preferences. However, a user's intent could be commercial even if an ad was not clicked (e.g., if a user examined the ads but did not find any that were relevant). Hence, to infer the *individual* commercial intent, clickthrough information is insufficient. Furthermore, existing server-side (i.e., clickthrough-based methods) tend to assign a single "dominant" intent to a query. Therefore, clickthrough information could be misleading when a user issues a rare query, or their intent differs from the majority. For example, the query "Nintendo Wii" could be commercial or non-commercial. People may search just to know more about Nintendo Wii, or perhaps the user goal is to find the best place to buy the game console. Ideally, we would like to categorize and understand the intent behind each *query instance*: that is, the particular search done by the user. Interestingly, exploiting personal user models directly may not solve this problem, as user goals may vary between search sessions.

Our approach is to use rich client-side instrumentation in order to obtain insights into user intent behind each query instance. In this paper, we focus on modeling mouse movements over the search results, which, we believe, could be used to help identify regions of the results of particular interest to the user. In particular, we attempt to predict whether a user had commercial intent (that is, exhibited interest in a commercial activity) for a given query instance, even in the cases where an ad was not clicked. Our preliminary experiments confirm that in some cases we can distinguish between commercial and non-commercial intent for the different instances of the same query string.

In summary, our contributions include: 1) a feasibility study of adapting lightweight client instrumentation for commercial intent detection; 2) a preliminary exploration of the effectiveness of mouse movement trajectory features for commercial intent inference; and 3) result analysis, focusing on the difficult and ambiguous cases that deserve further study.

2. 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 [3]. With the explosion of the popularity of the web, and with increasing availability of large amounts of user data, the area of modeling users, user intent, and in general web usage mining has become an active area of research in the information retrieval and data mining communities. In particular, inferring user intent in web search has been studied extensively, including references such as [13, 9, 1, 15]. Taxonomies of web search and user goals have been relatively stable since Broder's classic paper classifying intent into navigational, transactional and informational [4]. Recently, topical commercial query classification was presented in [13].

Previous research on user behavior modeling for web search focused on aggregated behavior of users to improve web search or to study other general aspects of behavior [6]. However, it has been shown that user goals and experience vary widely (e.g., [15]) and have significant effect on user behavior. Recently, eye tracking has started to emerge as a

SIGIR-IRA, July 2008, Singapore.

useful technology for understanding some of the mechanisms behind user behavior (e.g., [8, 5]).

In this paper we explore using *mouse movement* to attempt to infer *commercial query intent*. There have been indications that mouse activity (e.g., page scrolling) correlate with user interest [7], and could be used for better implicit feedback. Recent work showed a correlation between eye movement and mouse activity (e.g., [12, 11]). In other work, researchers have shown the value of mouse movement tracking for usability analysis [10] and [2] and activity tracking. However, we are not aware of previous work on using mouse movements to *automatically* infer query intent, that is, to automatically classify query instances into classes such as commercial vs. non-commercial, as we describe next.

3. CSII: CLIENT-SIDE INTENT INFERENCE

We now describe our CSII system for client-side instrumentation. Our goal is to capture as much information as possible about the user interactions with relatively lightweight and portable implementation. First we describe the architecture, and then report the details of our feature representation and classification methods.

3.1 Client-side instrumentation using LibX

For our research, we developed a minor modification of the Firefox version of the OpenSource LibX toolbar [1]. The instrumented Firefox browsers were installed on the public-use shared terminals in a major university library. The usage information was tracked only for users who have opted in to participate in our study, and no identifiable user information was stored to protect the privacy of the participants.

Specifically, we used simple JavaScript code to sample events such as the mouse movements on the pre-specified web search result pages. The events are encoded in a string and when the buffer of the events is filled are sent to the server. We then represent the interactions as *feature vectors* and then apply standard machine learning/classificaiton methods to classify query instances according to user intent. Next we describe the specific events captured, and the features used to represent the events.

3.2 Representing user interactions

In this study, we primarily focus on the mouse movements and the corresponding mouse move trajectories. First, we consider a coarse features such as the length, vertical range, and horizontal range of trajectories. For example, commercial queries are likely to require wider horizontal range of mouse movements to hover over the ads.

As the naive representation above is not rich enough to capture the possible information hidden in the mouse movements, we also capture more precise physiological characteristics, following the work of [11]. In particular, we attempt to capture properties such as *speed*, *acceleration*, *rotation* and other precise characteristics of the mouse movements.

To distinguish the patterns in different stages of the user interactions with the search results, we split each mouse trajectory into five *segments*: initial, early, middle, late, and end. Each of the five segments contains 20% of the sample points of the trajectories. Then, for each segment of the mouse trajectory we compute the average speed, average acceleration, slope and the rotation angle between the current segment and the segment connecting the beginning and the end (the click position) of the trajectories. The list of feature types is reported in Table 1.

[1]Available at `www.libx.org`

Note that, although this modeling can capture a user's commercial interest in a fine-grained manner(e.g., a user moves the mouse towards her interested ads without hovering the mouse over them), more explicit features such as "mouse over ads" or "mouse over organic results" would also be helpful which we plan to explore in the future.

Feature	*Specification*
TrajectoryLength	Trajectory length
VerticalRange	Vertical range
HorizontalRange	Horizontal range
Seg. AvgSpeed	Time elapsed between endpoints
Seg. AvgAcceleration	Velocity change from previous to current segment
Seg. Slope	Vertical range / horizontal range
Seg. RotationAngle	The angle between previous and current segment vectors

Table 1: CSII Mouse Movement Features

3.3 Learning to classify commercial intent

For our initial exploration we use standard supervised machine learning classification techniques. In particular, we used the Weka[2] implementation of the Support Vector Machines (SMO). We find that even out-of-the-box classifiers are able to demonstrate the feasibility and the benefits of using client-side instrumentation for commercial intent inference. We describe our empirical study next.

4. EXPERIMENTAL EVALUATION

4.1 User Interactions Dataset and Labeling

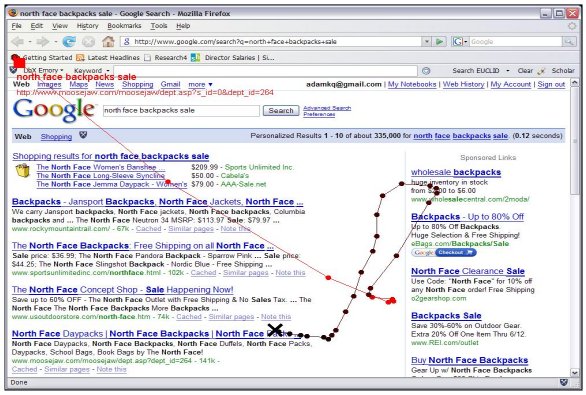
The data was gathered from mid-January through mid-May 2008 from public-use machines at a major university library. Our dataset contained searches for nearly 2,000 unique users, 32,000 search sessions, and 59,000 queries.

In order to focus on potentially commercial queries, we identified 235 query instances (178 unique queries) with clicks on Google ads as most easily available examples of potentially commercial queries. There were 736 instances of these queries in our logs, and we randomly selected 200 of these for manual labeling. In the sample, 87 (43.5%) query instances resulted in an ad clickthrough, and of the rest, 26 (13%) were labeled as Commercial, 87 (43.5%) were labeled as Non-commercial. Recall, that we focus on the "abandoned" potentially commercial queries (that is, to classify intent in the cases where an ad was not clicked) and hence we focus on the Commercial and Non-Commercial classes.

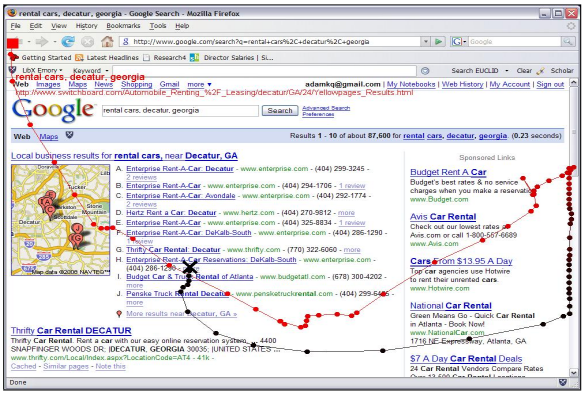
To manually label whether the query intent was commercial, we used intuition based on clues such as query terms and the next URL (often the URL of a clicked result). We also "replayed" the pre-click user interactions with the results for each query instance, drawing the corresponding mouse trajectory on a rendered snapshot of the result page. To illustrate the input we used to label the instances manually, Figure 1 shows examples of two commercial query instances; Figure 2 reports a sample of two non-commercial instances.

Finally, as the dataset described above turned out to be too small for effectively training a classifier, we *extended* the training dataset as follows. To provide additional commercial intent examples, we included the 235 instances with clicks on ads as training. To provide additional non-commercial examples, we manually labeled additional 300 instances with clicks on organic results, with clear informational or navigational (non-commercial) intent. This extended dataset was used only for training; the test set was still the original

[2]At `http://www.cs.waikato.ac.nz/ml/weka/`.

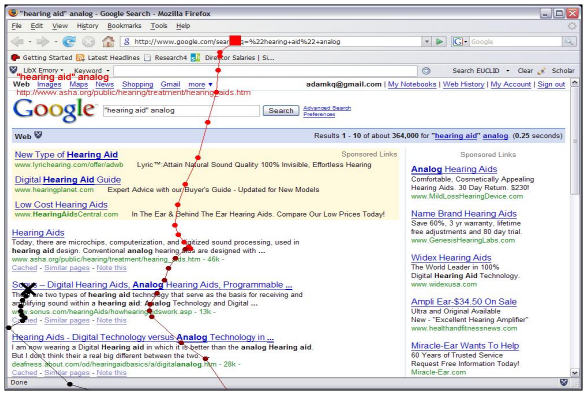


Query: "northface backpacks sale"

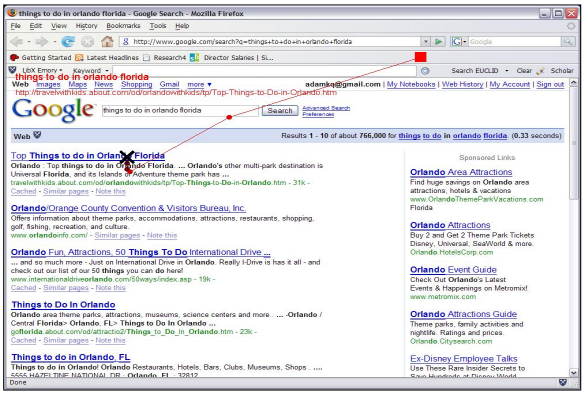


Query: "rental cars, decatur, georgia"

Figure 1: Two examples of commercial intent query instances



Query: "hearing aid analog"



Query: "things to do in orlando florida"

Figure 2: Two examples of non-commercial intent query instances

200 search instances originally labeled for "Commercial" or "Non-Commercial" intent.

4.2 Metrics

We use standard IR and classification metrics:

- **Precision (P)**: For a given class, of predicted instances that were correctly labeled.
- **Recall (R)**: For a given class, fraction of all true instances that were correctly identified.
- **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 as we have here. The F1 measure for each class is computed as $2 \cdot PR/(P+R)$.

4.3 Methods Compared

- **Naive Baseline**: always guess the majority class (*Non-Commercial*).
- **CSII**: SVM, trained on the extended training set described above.
- **CSII (Tuned)**: Same as above, but additionally tune the parameters of the Weka SMO implementation[3].

4.4 Results

We now report the preliminary experimental results in Table 2. Clearly, always guessing "non-commercial" results in sub-optimal performance, and is not useful. However, by including the extended training data (obtained as described above) results in a substantial improvement of 31% relative to the Naive baseline. Further tuning the parameters of the classifier adds another 6%, resulting in macro-averaged F1 value of 0.598. These results are promising, but leave much room for improvement, as we begin to explore in the rest of the paper.

To better understand the contribution of the different features we report the information gain of each feature (computed for the extended training set) in Table 3. As we can see, the most important features include trajectory length, vertical and horizontal range, and different aspects of mouse trajectories (e.g., rotation, slope, speed) in the initial and end stages.

Information Gain	*Feature*
0.305	RotationAngle (segment 4)
0.2273	Slope (segment 4)
0.1996	Slope (segment 0)
0.196	TrajectoryLength
0.1848	RotationAngle (segment 0)
0.1601	VerticalRange
0.1436	HorizontalRange
0.1037	AvgSpeed (segment 4)
0.0708	AvgSpeed (segment 0)
0.0678	RotationAngle (segment 1)

Table 3: Salient CSII features by Information Gain

5. DISCUSSION

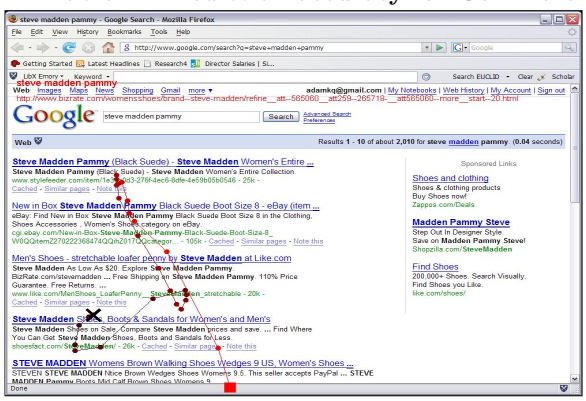
We examined the classifier errors and identified three main problems with our implementation:

Commercial organic results: Our method is in part based on the belief that a user's intent is commercial when she examines an ad. However, we can not distinguish a user's

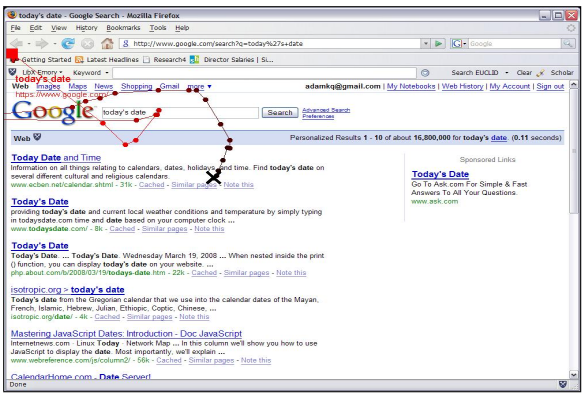
[3]All default Weka.SMO parameters, except for the following three: "NormalizedPolyKernel -C 250007 -E 2.0"

Method	*Commercial*			*Non-commercial*			*Macro Average F1*
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*	
Naive Baseline	0	0	0	0.77	**1**	**0.87**	0.435
CSII	0.326	0.538	0.406	0.829	0.667	0.739	0.573 (+31%)
CSII (Tuned)	**0.354**	**0.654**	**0.459**	**0.862**	0.644	0.737	**0.598 (+37%)**

Table 2: Prediction accuracy for Commercial vs. Non-commercial intent of CSII variants



Query: "steve madden pammy"



Query: "today's date"

Figure 3: Error examples: (a): Commercial search incorrectly classified as "Non-commercial" and (b): Non-commercial search incorrectly classified as "Commercial"

intent when she directly clicks on an organic result - a user's intent could still be commercial, as long as the organic result contains commercial information/service. In other words, if the organic results are strongly commercial and have large overlap with the ads, our method will not work. The first example in Figure 3 demonstrates such cases where our classifiers (incorrectly) predicts intent to be non-commercial.

Not using a mouse as a reading aid: Our detection of users' examination of ads depends on a user's mouse movement towards or over ads. However, if a user does not use the mouse as a reading aid, we can not detect commercial intent even if she has looked at an ad without hovering over it.

Similar mouse movements of examining organic results with examining ads: Sometimes the mouse movement pattern (e.g., rotation, slope, speed) of examining organic results is similar to examining ads. Or a user moves mouse seemingly towards the ads region for some reason (eg. no interesting organic results exist and some ad seems relevant at first glance even if her intent is non-commercial). The second example in Figure 3 demonstrates such cases where our classifiers (incorrectly) predicts intent to be commercial.

In summary, we presented preliminary exploration of using mouse movement analysis to automatically infer commercial search intent for search results with un-clicked ads. As we have shown, for some query instances, mouse movements can be successfully used to identify commercial or non-commercial intent. In particular, we can successfully detect a user's commercial intent even if a user does not click on an ad. We have also identified three key areas of improvement with our current implementation, that we believe can make our approach significantly more effective.

6. REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proc. of SIGIR*, pages 3–10, 2006.

[2] R. Atterer, M. Wnuk, and A. Schmidt. Knowing the users every move: user activity tracking for website usability evaluation and implicit interaction. In *Proc. of WWW*, pages 203–212, 2006.

[3] N. J. Belkin. User modeling in information retrieval. *Tutorial presented at the Sixth International Conference on User Modelling (UM97)*, 1997.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 2002.

[5] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proc. of CHI*, pages 407–416, 2007.

[6] D. Downey, S. T. Dumais, and E. Horvitz. Models of searching and browsing: Languages, studies, and application. In *Proc. of IJCAI*, pages 2740–2747, 2007.

[7] 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):147–168, 2005.

[8] 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.

[9] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proc. of WWW*, pages 391–400, 2005.

[10] F. Mueller and A. Lockerd. Cheese: tracking mouse movement activity on websites, a tool for user modeling. In *Proc. of CHI*, pages 279–280, 2001.

[11] J. G. Phillips and T. J. Triggs. Characteristics of cursor trajectories controlled by the computer mouse. *Ergonomics*, 44(5):527–536, 2001.

[12] 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.

[13] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proc. of WWW*, pages 13–19, 2004.

[14] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: repeat queries in yahoo's logs. In *Proc. of SIGIR*, pages 151–158, 2007.

[15] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proc. of WWW*, 2007.

Discussion: Enabling Research in the Area of Online Advertising

Microsoft adCenter Challenge: Data to the People

Mikhail Bilenko
Microsoft Research

Abstract

The dearth of open, large-scale datasets is a key problem for empirical research in computational advertising. We describe Microsoft adCenter Challenge, a search advertising dataset collected over real search engine traffic, that will be shortly available to the public. The dataset is unique in that every query impression is accompanied by three ads shown in random order, which allows removing positional bias effects. We define the ad selection task on this dataset, which maps to the key problem of pay-per-click advertising, clickthrough prediction, and provide some preliminary analysis related to the dataset and the task.