ABSTRACT

INTENT-BASED USER SEGMENTATION WITH QUERY ENHANCEMENT

by
Wei Xiong

With the rapid advancement of the internet, accurate prediction of user's online intent underlying their search queries has received increasing attention from the online advertising community. As a rich source of information on web user's behavior, query logs have been leveraged by advertising companies to deliver personalized advertisements. However, a typical query usually contains very few terms, which only carry a small amount of information about a user's interest. The tendency of users to use short and ambiguous queries makes it difficult to fully describe and distinguish a user's intent. In addition, the query feature space is sparse, as only a small amount of queries appear very often while most queries appear only a few times. Users may use different search terms even if they have the same interests. For example, "Camera", "digital camera", "Sony" and "RX100" are all about cameras. This study aims to address these challenges with user queries in the context of behavioral targeting advertising by proposing a query enhancement mechanism that augments user's queries by leveraging a user query log.

Different from traditional user segmentation methods, which take little semantics of user behaviors into consideration, this study proposes a user segmentation strategy by incorporating the query enhancement mechanism with a topic model to explore the relationships between users and their behaviors in order to segment users in a semantic manner. This research also proposes, in the case that the dataset is sanitized, an alternative to define user's search intent for evaluation purposes. This approach

automatically labels users in a click graph, which are then used in training an intent-based user classifier. The empirical evaluation demonstrates that the proposed methodology for query enhancement (QE) achieves greater improvement than the baseline models in both intent-based user classification and user segmentation. Comparing with a classical clustering algorithm, K-means, the experimental results indicate that the proposed user segmentation strategy helps improve behavioral targeting effectiveness significantly. Particularly, the average PUR (Positive User Rate) improvement rates under “K-means + QE” strategy significantly increase over simple K-means strategy in different number of segments across all six domains. The PUR improvement rate can be as high as 136.6% by using the proposed user’s intent representation technique with the query enhancement mechanism under the LDA model. By further analysis, the proposed “LDA + QE” strategy significantly exceeds K-means and “K-means + QE”.

INTENT-BASED USER SEGMENTATION WITH QUERY ENHANCEMENT

by
Wei Xiong

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Information Systems

Department of Information Systems

August 2014

.

APPROVAL PAGE

INTENT-BASED USER SEGMENTATION WITH QUERY ENHANCEMENT

Wei Xiong

Dr. Y.F. Brook Wu, Dissertation Co-Advisor Date
Associate Professor of Information Systems, NJIT

Dr. Michael Recce, Dissertation Co-Advisor Date
Associate Professor of Information Systems, NJIT

Dr. Lian Duan, Committee Member Date
Assistant Professor of Information Systems, NJIT

Dr. Songhua Xu, Committee Member Date
Assistant Professor of Information Systems, NJIT

Dr. William Browne, Committee Member Date
Modeling Engineer, Quantcast Corporation, San Francisco, CA

BIOGRAPHICAL SKETCH

Author: Wei Xiong

Degree: Doctor of Philosophy

Date: May 2014

Undergraduate and Graduate Education:

- Doctor of Philosophy in Information Systems,
  New Jersey Institute of Technology, Newark, NJ, 2014

- Bachelor of Engineering in Software Engineering,
  Hubei University of Economics, Wuhan, P. R. China, 2007

- Bachelor of Management in Marketing,
  Hubei University of Economics, Wuhan, P. R. China, 2007

Major: Information Systems

Presentations and Publications:

Xiong, W., Recce, M., Wu, YB. (2014) "Intent-based User Segmentation with Query Enhancement", International Journal of Information Retrieval Research

Xiong, W., Song, M., Watrous-deVersterre, L. (2010) "A Quantitative Assessment of SENSATIONAL with an Exploration of its Applications", In Proceedings of the 23rd International FLAIRS Conference. Menlo Park, CA: AAAI Press

Xiong, W., Song, M., Watrous-deVersterre, L. (Book chapter). A Comparative Study of an Unsupervised Word Sense Disambiguation Approach, In P. McCarthy and C. Boonthum (Eds.), Applied Natural Language Processing and Content Analysis: Identification, investigation, and resolution. IGI Global

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents, Kemin and Yanping whose words of encouragement and push for tenacity ring in my ears.

ACKNOWLEDGMENT

With a deep sense of appreciation, I would like to express my sincere thanks to Dr. Y.F. Brook Wu and Dr. Michael Recce for their immense help in planning and executing this work. Their support, encouragement and reassurance are greatly acknowledged.

My sincere thanks to Dr. Lian Duan, Dr. Songhua Xu and Dr. William Browne for reviewing this work and providing me with important suggestions.

I would like to express profound gratitude to my family and friends for uncomplaining patience, perennial support, and enormous encouragement during my studies.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

CHAPTER 1

INTRODUCTION

1.1 Introduction

With the dramatic advancement of the World Wide Web, online advertising has been the fastest growing advertising medium in history. It started out as online banner ads back in 1994 and has turned into a multi-billion dollar market that continues growing.

Ad targeting has been receiving more and more attention in the online publishing world, where advertisers want their ads to be seen by the potential consumers at the right time. There have been studies on ad targeting technologies which try to understand characteristics of online users and deliver them ads based on their interests. For example, the most basic targeting approach is to show ads based on the geographic information of the users, such as the physical location of the user. This approach is effective for advertisers who want to target a specific location, such as countries, cities or a radius around a location. One of the main reasons one may use geographic targeting is simply because one only offers products or services within specific areas. Geographic targeting also offers advertisers the ability to target their ads to users based on other parameters such as user connection speed, Internet Service Provider (ISP), domain name, and so on. For example, advertisers can deliver a competitive ad based on a user's domain name.

Similarly, demographic targeting approach targets ads to people based on the demographic information of the users, such as gender, income, age and more. For example, if you are a skateboard advertiser and know that skateboard users tend to be young males, you can set your campaign to show mostly to that audience. One of the advantages of demographic targeting is that advertisers can select a small amount of users based on demographics rather than displaying ads to all the users. However, this approach could also miss out potential buyers who do not fall into a specific demographic category. For example, a grandmother can also be a skateboard buyer if she wants to give a skateboard to her grandson as a gift.

Another three commonly used targeting methods are contextual targeting, keywords targeting, and retargeting. Contextual targeting is an advertising model where advertisements are targeted to the content of a webpage. In this model, the advertisement in a webpage is usually relevant to the content of that webpage. For instance, if a user is viewing a webpage pertaining to travel and that webpage uses contextual advertising, the user may see banner or pop-up ads for travel-related companies, such as flights dealers, hotels, and so on. Google AdSense was a major contextual advertising network and a large part of Google's profit is from its share of the contextual advertisements displayed on the websites running the AdSense program that searches for the relevant ads using Google's search algorithm. Contextual ads will be displayed based on the keywords after a contextual advertising system scans the text of a webpage.

On the other hand, keywords-targeted advertisements are displayed on the search results pages based on the keywords in the queries issued in search engines. Google AdWords is one of the most well-known forms of keywords targeting, where Google displays search ads based on the word(s) typed into its search box. One of the most widely used strategies is to bid on keywords by geography, allowing advertisers to maximize click-through-rate (CTR). For instance, one could adjust bids by geographic areas to get more exposure in areas that perform well. Furthermore, the keyword targeted campaigns are usually charged on a cost-per-click (CPC) basis, where advertisers are only charged when a user clicks on their ad and is taken to their landing page. The final CPC rate is calculated based on the advertiser's maximum CPC bid as well as the search engine's internal system of scoring keyword ads. Therefore, it is crucial to select accurate and appropriate keywords relevant to the product or service in the ad and set the maximum CPC bid (the most the advertiser is willing to pay per click).

Retargeting works by keeping track of users who visit a company's website and displaying ads from that company encouraging them to buy its products while they are visiting other sites online. The idea behind retargeting is that, only a small amount of users will convert on the first visit to a website. Retargeting was introduced in an effort to help advertisers allocate their advertising budget efficiently to their targeted audience and hence increase the effectiveness of online advertising. Yahoo! Retargeting, for example, is an online advertising platform that tracks users who have browsed a publisher's website before and tries to bring them back by displaying the ads the next time the user is on a Yahoo network. As a powerful and effective targeting strategy, retargeting focuses the advertising spending on users who are already familiar with the product or have recently shown interest. By displaying ads to the users multiple times after they leave the website, retargeting increases the chances that they will come back again.

However, with the rapidly expanding breadth of Internet usage data collected by marketers, behavioral targeting makes online advertising more effective. To some extent, behavioral targeting is another application of machine learning methods to online advertising. Unlike contextual targeting and keywords targeting, behavioral targeting does not primarily rely on the contextual information. Instead, behavioral targeting helps advertisers reach the most relevant users by learning from user's online behavior, such as user’s search queries and web browsing history. This research introduces a user intent representation strategy and a query enhancement mechanism to tackle the problem of user classification and use segmentation from a behavioral targeting perspective for online advertising.

1.2 Background and Motivation

As a rich source of information on web user’s behavior, query logs have been leveraged by advertising companies to deliver personalized advertisements. These log files typically consist of a unique identifier for the user, the query string submitted by the user, a timestamp, and URLs clicked for that query. To carry out research on behavioral targeting, it is always desirable to have benchmark datasets available, which contain both query logs and ad click information. This type of dataset can be used to train and test a model that predicts user’s ad click behavior. Yet, they are rarely available in the academic community, which makes conducting research in this area difficult. The publicly available query logs are small, dated, and sanitized, as search engine companies tend to be reluctant to release complete query log data. One of the objectives of this research is to propose an alternative to define user’s search intents for evaluation purposes, in the case that the dataset is sanitized. The desired approach should be able to automatically label user’s online intents, which then can be used in training and testing the proposed models.

The volume of queries has grown at an unprecedented pace during the past decade. However, the length of queries always tends to be short. A typical query usually contains very few terms, which only carry a small amount of information about a user’s interest. The tendency of users to use short and ambiguous queries makes it difficult to fully describe and distinguish a user’s intent. For instance, the user intent behind query “Steve Jobs” will be represented as two terms in the BOW model: “Steve” and “Jobs”, along with their weights in the feature space, which could describe an intent of a user who is either interested in the person “Steve Jobs” or looking for a job. In addition, the number of queries issued by different users over a period of time greatly varies. Hence, even less information can be captured from the users who issue only a couple of search queries in a given period of time, which makes the problem even more challenging.

On the other hand, the query feature space is sparse, as only a small amount of queries appear very often while most queries appear only a few times. Users may use different search terms even they have the same interests. For example, “Camera”, “digital camera”, “Sony” and “RX100” are all about cameras. However, “RX100” is a more specific query with much fewer occurrences. Without knowing “RX100” is a camera model, this query would not lead to more focused advertisements.

One of the crucial problems in Behavioral Targeting is user segmentation with the purpose of grouping users into user segments with similar behaviors. Under the traditional Bag of Words model, users who have similar online intent but use different query terms can be very hard to be grouped into the same segment. For example, a user who issued query “cheap flight” and another who issued query a “discount airfare” may have the exact same intent of purchasing a flight, even though the queries issued by them are totally different.

Overall, the behavioral targeting advertising research problem involves the following three challenges:

- Lack of golden standard datasets on Behavioral Targeting in academia.

- Short and ambiguous queries making it difficult to describe and distinguish a user's intent.
- Sparseness of query space.

This research aims to address the above challenges with user queries in the context of behavioral targeting advertising by proposing a user intent representation strategy and a query enhancement mechanism. This dissertation focuses on investigating the intent based user classification performance and the effectiveness of user segmentation under a topic model that helps explore semantic relation between user queries in behavioral targeting.

1.3 Research Questions

Assume a user who issued queries like "best carry-on luggage" and "foreign transaction fee". From the observation of this user's queries, it can be inferred that this user is probably planning an oversea trip and may have an intent to purchase a flight. Thus, it is the opportunity not only for advertisers to deliver flight advertisements, but also for other online service providers to offer travel related service.

This study is focused on capturing relevant users based on their online intents. To perform such a study, three major research questions need to be investigated: First question is how to represent a user's online intent. Since user's offline activities cannot be easily captured online, a user's online intent should be modeled based on the user's online behavior, such as the search queries issued by the user and the search results clicked. Also, for a certain online intent, a user can be classified as either having this intent or not having this intent. Therefore, a good intent representation strategy should be able to effectively differentiate users based on their online intents. Furthermore, it would be also interesting to investigate how much intent-based user clustering could help behavioral targeting by grouping similar users into segments according to their online intent. More specifically, the following primary research questions are to be answered:

Question No 1:

How to represent a user's online intent?

Question No 2:

How well can users be classified based on their intents?

Question No 3:

Does the intent-based user segmentation improve the performance of behavioral targeting significantly?

1.4 Methodology and System Framework

This research first reviews the background of behavioral targeting advertising and related work in user segmentation as well as query log exploitation, and then presents the query enhancement solution for user intent representation. The proposed query enhancement mechanism augments the query by leveraging a user query log, which provides more information about the user's interests and hence reduces the ambiguity in the user's intent for better user classification and behavioral targeting effectiveness.

Traditional user segmentation is based on the Bag of Words model and does not take the sematic relation among user queries into consideration. This study proposes to project user's queries to a topic level which represents the semantics underlying user's queries. The proposed approach is motivated by the use of topic models in the field of information retrieval and adopts Latent Dirichlet Allocation (LDA) to present user's online intents on a topic level in order to investigate the impact of intent-based user segmentation on the performance of behavioral targeting.

With the lack of benchmark datasets, this research also proposes alternatives to define user's search intents. The proposed approach automatically labels a large amount of users in a click graph, which are then used in training an intent-based user classifier. The evaluation focuses on the performance of the proposed user classification method and the effectiveness of the proposed behavioral targeting model. The performance of the user classification is measured by the positive precision, since advertisers always want to deliver ads to those who have a high probability of having an intent related to the product. The effectiveness of the proposed behavioral targeting model is measured by the positive user rate (PUR) improvement in the user segment.

Figure 1.1 shows an overview of the system framework. The system includes two components. The first component performs user classification and query enhancement, which includes user labeling, query enhancement mechanism and user classification. The system first takes a query log and the external dataset Delicious to label the users and build a click graph which is then used to augment user's search query in the query enhancement mechanism. The user's intents are then presented in the BOW model and a classifier is trained. The performance of the proposed user classification is evaluated after feeding a set of testing dataset into the classifier. The first component will be described in detail in Chapter 3. The second component of the system performs user segmentation, which presents user's intent on a topic level and users are clustered into different segments under an LDA model. The datasets used in the second component are processed in the same way as in the first component. Detailed discussion on the second component can be found in Chapter 4.

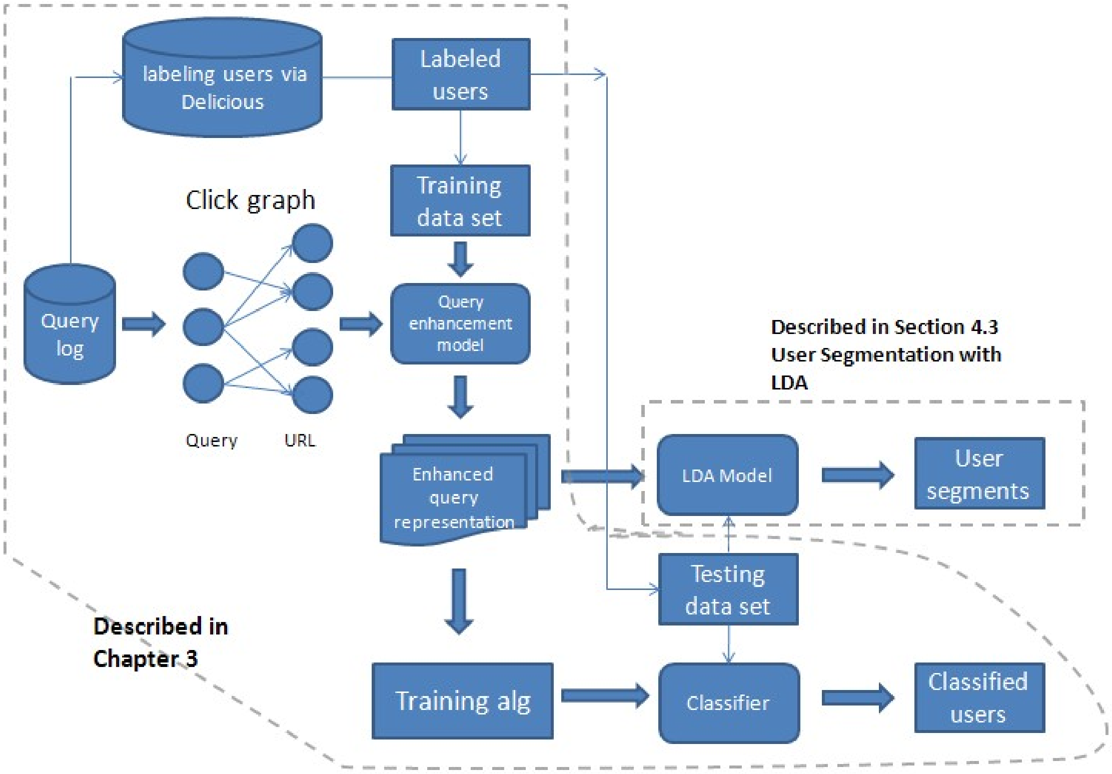


Figure 1.1 Proposed system framework. The first component performs user classification and query enhancement, which includes user labeling, query enhancement mechanism and user classification. The second component of the system performs user segmentation, which presents user's intent on a topic level and users are clustered into different segments under an LDA model.

1.5 Organization of This Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 provides an overview of online advertising. Chapter 3 provides a review of the literature related to this study. It presents the background of behavioral targeting and an overview of applications of query logs. It also discusses query representation techniques and insights into user online behavior. Chapter 4 introduces a user intent representation strategy and proposes a query enhancement mechanism to address the sparseness issues with search queries. It focuses on the problem of binary user classification based on user's online intents and provides the description of the datasets along with evaluation of the proposed method. It also proposes, in the case that the dataset is sanitized, an alternative to define user’s search intents for the evaluation purpose. Chapter 5 proposes an LDA-based user segmentation approach and examines the effectiveness of user segmentation in behavioral targeting. Chapter 6 summarizes the dissertation and discusses the contribution of this research as well as limitations.

CHAPTER 2

ONLINE ADVERTISING

2.1 Introduction

Online advertising has been a market where websites sell space on their webpages to advertisers, who pay for this space to display their ads to the website's audience. There are several different categories of ads: display/banner ads, search ads, and video ads. Banner ads are one of the earliest forms of online advertising and generate a big part of the revenue for many web sites, which appear somewhere on the page and led to the advertiser's site when clicked. Search ads are targeted to match search terms entered on search engines and appear on web pages that show results from search engine queries. Video adverting is a relatively new form of advertising and it is served before, after or during a video content.

A recent report by eMarketer indicates that, search ads spending is about half of all online advertising spending. Table 2.1 illustrates the evaluation of ads spending over six major media: newspapers, radio, TV, magazines, internet and outdoor. The presented numbers show the share of each medium as a percentage. According to Table 2.1, advertisers spent 5.2% of their advertising budgets on internet ads in 2001 and it is predicted that this share will grow to 26.4% by 2015. It is worth noting that, although online advertising spending has been increasing at an unprecedented pace over the past decade, it has not surpassed the amount spent on TV advertising.

Table 2.1 US Major Media Ads Spending Share

Year	newspapers	radio	TV	magazines	internet	outdoor
2001	32.5	13.0	37.9	7.8	5.2	3.7
2002	30.7	13.2	40.7	7.7	4.2	3.6
2003	30.0	12.8	40.6	8.2	4.8	3.6
2004	31.7	14.2	33.8	8.5	6.6	5.1
2005	30.6	13.9	34.4	8.2	7.9	5.0
2006	26.6	10.8	37.6	12.8	8.8	3.5
2007	25.1	10.2	36.9	12.9	11.0	3.9
2008	19.9	9.9	39.4	13.1	13.5	4.2
2009	16.7	9.5	42.3	11.9	15.5	4.1
2010	14.6	9.8	43.9	11.2	16.6	3.9
2011	14.6	10.7	41.3	9.5	19.5	4.4
2012	13.4	10.6	41.8	8.6	21.1	4.4
2013	12.8	10.6	41.2	8.0	22.8	4.5
2014	12.2	10.4	40.8	7.4	24.7	4.5
2015	11.7	10.2	40.3	6.9	26.4	4.5

Source: Internet Advertising Bureau (2001-2010) and eMarketer (2011-2015).

The data indicates that advertisers are gradually shifting their budgets from other media to internet advertising. This is not only due to its increasing number of users but also due to its inherent advantages over other media. For instance, internet is the only medium that allows truly international access. An ad shown in a web page can be viewed and clicked by any user in the world who accesses that web page. In addition, online advertising has brought a real revolution in term of targeting potential customers. For example, an advertiser that sells sport products would show his ads in a sport channel, since sport channels attract overwhelmingly male audience and the buyers of sport products are mostly male, too. Although only a subset of these viewers are interested in buying sport products, it is extremely difficult for the advertiser to identify them and target his ad only to these users. However, internet brings about much richer information about a user and his intent which allow more effective targeting. For instance, in keywords targeting advertising, an advertiser that sells laptops can show the ad only to users who issue queries such as “best laptop”. Online advertising not only helps the advertisers target the really interested users but it also helps users receive less irrelevant ads for a better online experience.

John Wanamaker, who is known as the “Father of modern advertising”, declares that “Half the money I spend on advertising is wasted; the trouble is I don’t know which half”. It illustrates how difficult it is to reach potential customers and how difficult it is to truly measure the impact and effectiveness of an advertising campaign. Online advertising, however, has begun to reduce those uncertainties, where advertisers can monitor the user interaction with the ad and have a clear picture of the impact of their advertising campaign. For instance, an advertiser can see how many users have clicked an ad and in some cases, whether the user who views the ad ends up making a purchase or signing up a service. This helps advertisers better estimate the effectiveness of their online ads versus ads on other media.

In general, there are two types of advertising: branding and direct response. Brand advertisements aim to build the awareness of their brand in a large audience without expecting to elicit a purchase right away. Direct response advertisements, on the other hand, urge a prospective customer to respond immediately: for example, “click here to get a free quote”. There is no clear boundary between branding and direct response. For instance, a video advertisement can tell a story about a brand but also invite the audience to click on the embedded link to make a purchase.

2.2 Advertisers, Ad Agencies and Publishers

Traditionally, ad agencies provide advertisers with a variety of services, ranging from ad designing to media buying. Over the past decade, ad agencies have been transitioning from relatively small organizations to a small number of holding companies that control almost all of major agencies. The consolidation in the agency business, however, allows individual agencies to retain their own identity.

As online advertising becomes increasingly data-driven, both ad agencies and technology companies, such as Google, have been in the business of analyzing user's behavioral data with the purpose of increasing ad effectiveness. The relationship between ad agencies and technology companies also becomes complex and ad agencies may be feeling pressure from technology companies. In an annual report, WPP [31] reviewed the complex relationship it has with Google, and provided these comments:

"*All in all, Google is opening up the attack on many fronts. Perhaps too many, particularly when you consider the other theatres it is fighting in, such as book publishing and robots to the moon. One gets the impression it is throwing a lot of mud against the wall to see if any sticks – maybe sticking to mobile search would be best. Yahoo! has a different approach, working through its agency partners and believing in the power of people, rather than Google's greater focus and belief in technology. Certainly, even now, a combination of Microsoft and Yahoo! in any way will bring* *greater balance to the markets. Our clients and our agencies will favour a duopoly rather than a monopoly*."

On the other hand, publishers are companies or individuals that develop and maintain websites. To some extent, any part that sells ad inventory within the online advertising industry can be referred as a "publisher", such as an online retailer. Traditional publishers focus on the production of the content and are dependent on advertising revenue, such as magazines and blogs. Since different advertisers are interested in different groups of audiences, it is critical for publishers to understand their audiences and show that their audiences have value for the advertisers.

There are typically two categories of inventory: "premium" and "remnant" [69]. Premium inventory could be the ad spots on the home page of a web site which may be seen by millions of people every day, and they can be sold directly to advertisers at a higher price. Remnant inventory, on the other hand, is the inventory that cannot be sold directly to advertisers due to the fact that they are obscure pages or do not have relevant contents that interest advertisers. However, there is no hard distinction between premium inventory and remnant inventory. In some cases, remnant inventory could turn into premium inventory if packaged in the right way to advertisers.

2.3 Interactions among Parties in Online Advertising

There are there parties involved in online advertising: the advertisers, the users, and the advertising media which includes search engines in sponsored search and the web site publishers in display advertising. Generally, the advertiser wants to deliver a message to users of interest, which usually prompts the users to perform actions that benefit the advertiser, such as make a purchase from the advertiser's online store. The advertiser reaches the users via the advertising media, such as the search engines or the web site publishers. In particular, the advertiser provides the media with its ad and preferences of audiences, such as young male audience or users who are interested in a camera. The media deliver the ads to users based on the advertiser's budget and audience preferences. The advertiser pays for the media and expects returns from online advertising campaign, which comes from the user's action as a response to the ads. The interactions among the parties can be summarized into three steps [50] :

1. *Bidding*: This is a step that happens in the interaction between the advertiser and the media. The advertiser provides the media with its ad messages, its references of audience and the price that it is willing to pay. In sponsored search, the audience preferences are characterized by keywords. An advertiser may want its ad only to be displayed to the users whose search query matches one of the provided keywords. Those keywords are usually closely related to the products or services the advertisers provide. In display advertising, the advertiser selects a set of web pages and a timeframe. The advertiser wants its ad to be displayed when a user views one of the selected pages during the specified timeframe. The advertiser can also express its audience preferences by selecting the demographic characteristics, such as gender and income, of the users who view its ads. Regarding the pricing and payments in the sponsored search, the advertiser pays the search engine for every click on its ads. Advertisers typically bid on the keywords relevant to their product or services, and the amount being charged per click depends in part on the maximum cost-per-click bid, which is also called "max CPC" bid. This indicates the highest amount that the advertiser is willing to pay for a click on its ad. Similarly, in display advertising, the web site usually charges the advertiser for every impression of its ads and the selection of advertisements to show on a given page during a specific time frame can be also chosen based on price, using an auction in a similar way to sponsored search.
2. *Delivery*: The step of delivery happens between the media and the users. When a user visits the web site of a medium, the medium needs to decide which ad to show to the user. Ad selection is challenging and important. The ad to be displayed should not only conform to the advertiser's references, but also it should optimize the use of the inventory from the medium's perspective. The process of ad selection through an auction among thousands of ads happens in a real-time setting, usually within a few hundred milliseconds. Publishers simply want to make the most advertising revenue from the web sites, without irritating users by overwhelming them with ads. Particularly, advertisers bidding on the same keyword in sponsored search repetitively take part in all of the auctions for this keyword. Therefore, advertisers are not allowed to change their bids in the auctions to prevent advertisers from affecting the price that they need to pay to

win. Otherwise, lots of resources of the media would be wasted due to the bid fluctuations. The delivery step ends with the display of the selected ad to the user.

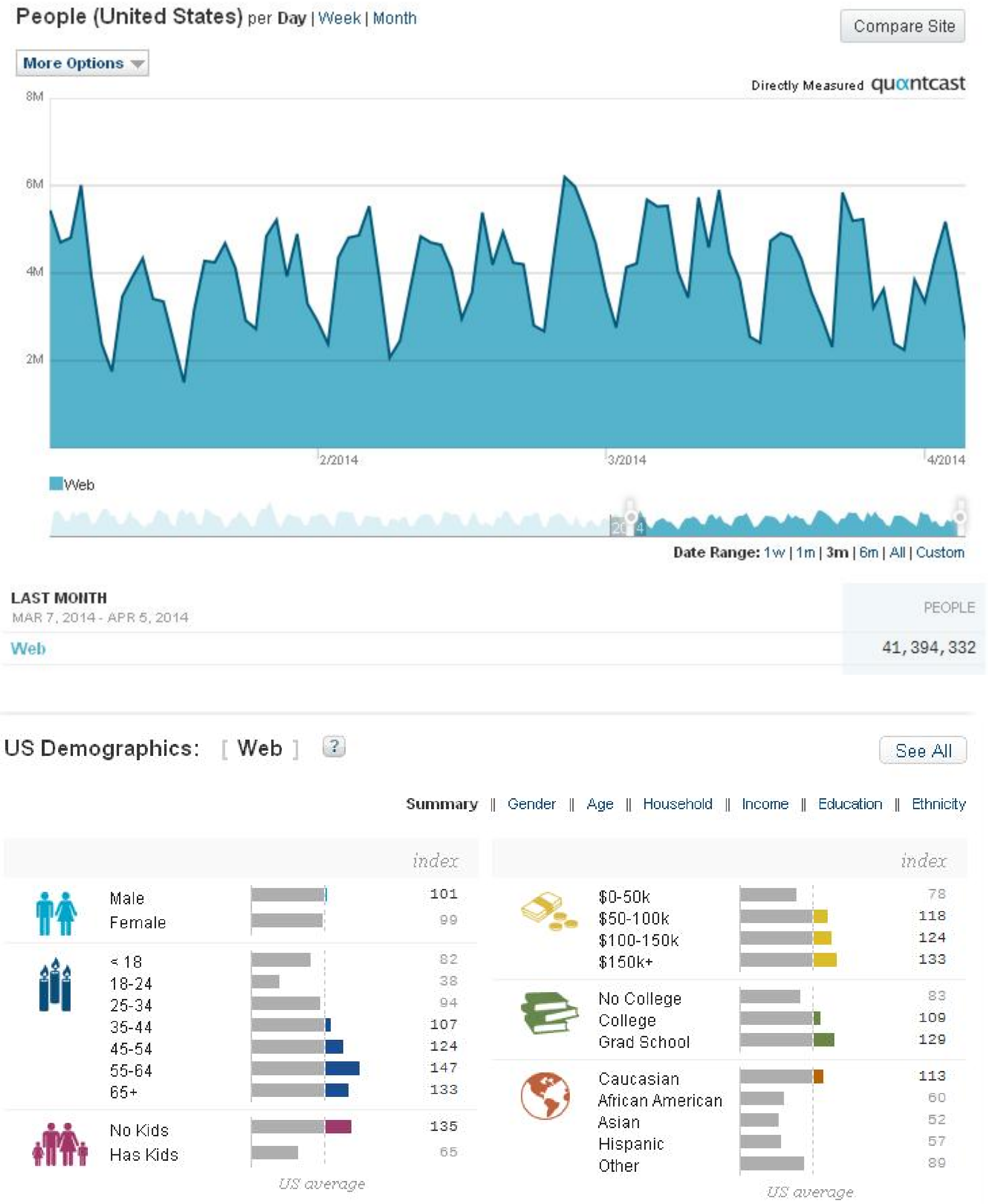
3. *Response*: This step refers to interaction between the users and the advertiser. A user can either ignore the ad or click on it to proceed to an action after viewing an ad. For example, the user could click on the ad which leads to the advertiser's online store. The purpose of advertising is to make the user a customer of the advertiser's business, such as make the user purchase from the advertiser's store or sign up the advertiser's service. However, it is hard to track the fulfillment of the purpose. For example, a user may purchase the product from the advertiser's online store several days later after he viewed the ad, and he could also make the purchase in the advertiser's local store. Therefore, the clicks on the ad have been widely used to measure user's response in online advertising industry. Other pricing models used in online advertising are also discussed in Section 2.5.

2.4 Online Audience Measurements

In order to better plan online advertising campaigns, marketers need to have an overview of the audience of a given website. General audience measurements typically include the number of unique visitors to a website and the demographics of the visitors, such as age, income, education level. The audience measurement companies usually use survey panels that collect data from a large number of users who have agreed to install software on their computers that records their online activities which include their browsing activities and shares it with the survey company. They also agree to report their age, gender and other demographic information so that the survey company can produce statistics about the audiences of different websites.

Nielsen and comScore are the biggest names in online audience measurement, but there are other players, such as Quantcast and Google's Display Planner (previously Google Ad Planner tool). All of these companies can produce statistics about the audience demographics of a given website, for example, the percentage of a given website's users that are female between 45 and 60, and have an annual income above

$80k. Figure 2.1 is a screenshot of the audience statistics for *nbcnews.com* website from Quantcast free analytics service. The trend graph shows unique number of visitors coming from the U.S. each day, over the past several months. In terms of the demographics, Quantcast reports a fairly even gender distribution. It also reports most of the audiences have no kids.


People (United States) per Day | Week | Month
Compare Site
More Options
Directly Measured quantcast
8M
6M
4M
2M
2/2014
3/2014
4/2014
Web
Date Range: 1w | 1m | 3m | 6m | All | Custom
See All
US average
US average

LAST MONTH
MAR 7, 2014 - APR 5, 2014

	PEOPLE
Web	41,394,332

US Demographics: [Web]

Summary || Gender || Age || Household || Income || Education || Ethnicity

	index
Male	101
Female	99
< 18	82
18-24	38
25-34	94
35-44	107
45-54	124
55-64	147
65+	133
No Kids	135
Has Kids	65

	index
$0-50k	78
$50-100k	118
$100-150k	124
$150k+	133
No College	83
College	109
Grad School	129
Caucasian	113
African American	60
Asian	52
Hispanic	57
Other	89

Figure 2.1 Screenshot of Quantcast audience data for “nbcnews.com”.

2.5 Online Advertising Pricing

There are several pricing models in online advertising industry. One common model is CPM or cost-per-mille impressions (mille means thousand in Latin), where an "impression" is counted each time the ad is shown. In other words, the payment in CPM is based on the number of times the ad is shown. It is calculated by dividing the cost of an advertising placement by the number of impressions (expressed in thousands) that it generates. For instance, if a publisher charges $5 CPM and an advertiser agrees to run a campaign on the publisher's website for 100,000 impressions, the advertiser would make a payment of $500. This model is widely used for "branding campaigns" where the main goal is to build the awareness of a product or a service. Publishers get paid for every impression and risk nothing on the ads performance, regardless of whether or not the ad leads to a click or other action. This results in a relatively predictable stream of earnings for publishers, which means if a publisher can predict his website traffic, he can predict his revenue.

Unlike CPM model where ad clicks do not affect the price, CPC model or cost-per-click, is a "performance-based" metric, where the advertiser only needs to pay the publisher only when a user clicks on an ad, regardless of the number of impressions served. It is preferred by advertisers, especially for those who are running "direct response" campaigns. For example, the same publisher and advertiser from the above example agree to use a CPC pricing model where the advertiser pays $3 for each ad click and the publisher generates 100 clicks by serving 100,000 impressions. In this case the advertiser needs to pay the publisher $300. From a publisher's perspective, there is a pretty big risk when running CPC campaigns: if the ads served do not lead to any clicks, the publisher could end up with zero compensation, even for serving a large amount of impressions on its websites. On the other hand, CPC campaigns are low risk for advertisers as they only need to pay for the ads that lead to clicks.

There is another pricing model called CPA (Cost-Per-Action), where the advertiser compensates the publisher only for ad clicks that subsequently result in a sale or conversion against advertiser's campaign goal, such as a purchase of a product or sign up for a credit card. It is also low risk for the advertisers because they only need to pay when the ads generate their desired outcome.

From a publisher's perspective, the CPM model gives the lowest risk as the publisher is guaranteed to receive the compensation as long as the ads are displayed. On the other hand, CPA has the highest risk for publishers, because the payment from the advertisers depends on whether or not the user performs an action that favors the advertiser after viewing the ad. Even if a user views the ad, clicks on it, but does not convert, publishers will not get any compensation under the CPA model. The risk level of a CPC model sits in the middle of CPM model and CPA model, and it has been widely used in online advertising industry.

CHAPTER 3

BEHAVIORAL TARGETING

3.1 Introduction

As a rich source of information on web searchers' behavior, query logs have been utilized by advertising companies to deliver personalized advertisements and leveraged by researchers to tackle other application problems, such as query suggestion. To carry out research on behavioral targeting, it is desirable to have golden standard data sets available, which contain both query logs and ad click information. This type of data sets is used by advertising companies to train and test a model that predicts user's ad click behavior. However, they are not available in academic community, which makes conducting research in this area difficult. The publicly available query logs are small, dated, and sanitized, since search engine companies are reluctant to release complete query log data. It is understandable considering that query logs can reveal private information and they cannot be thoroughly sanitized. This is because query logs potentially contain a great amount of sensitive personal information and it is possible to analyze the query log to identify individual users. Therefore, this study also attempts at finding alternative ways to define user's search interests.

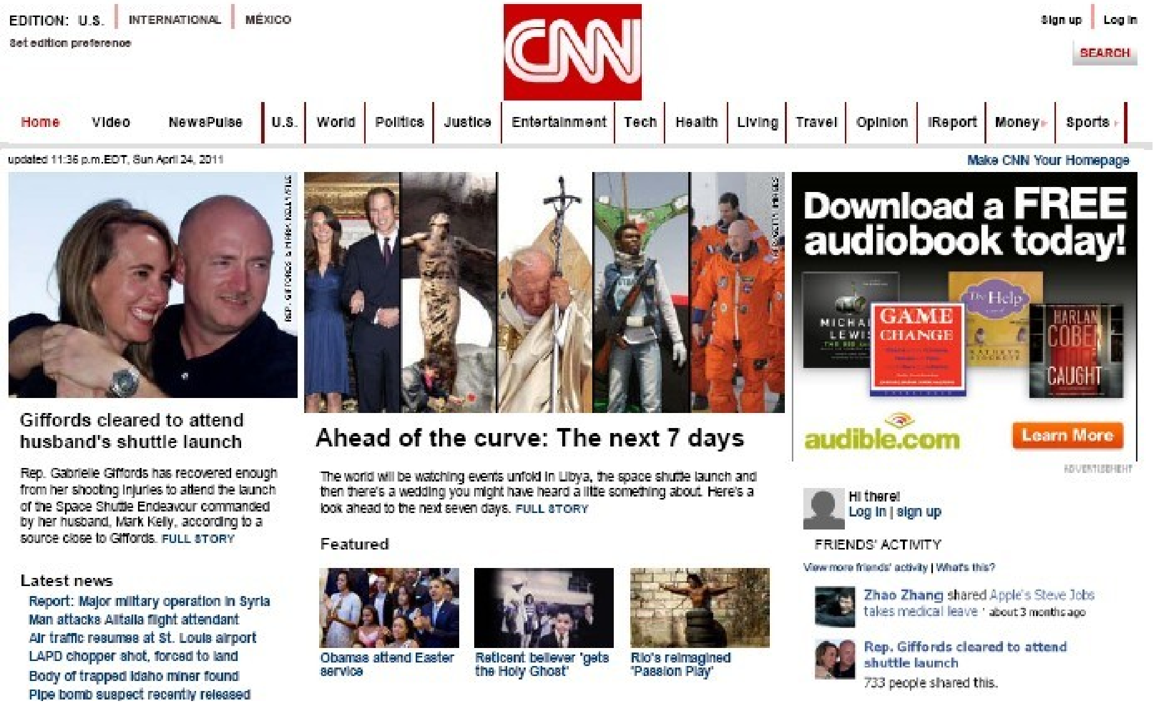
This chapter provides background information on behavioral targeting, overview of applications of query logs, query representation techniques, and insight into user online behavior.

3.2 Behavioral Targeting

Online advertising spending has been increasing at an unprecedented pace over the past decade. In order to increase advertiser's revenue, models are built based on user's web activities, such as search queries, to personalize advertisements. There are hundreds of companies and many different approaches, e.g., context, social, cookie-based, etc., for precisely targeting advertising. The largest internet companies, such as Google, Facebook, and Yahoo, are all advertising companies. Data from search activities, web surfing and social connections are all mined to optimize advertising revenue.

3.2.1 Overview

There are two major types of online advertising: search ads and display ads. Search ads are the advertisements links on the search result page when users look for information online, while display ads are shown on a page after the page navigation. In display ads, every time a user loads a page with a spot for advertising, an auction is held for advertisers to bid for the opportunity to display their ads to this user. Advertisers make their bid decisions by predicting the user's interest. This process is very fast as the communication between advertisers and publisher takes place in only milliseconds while the page is loading.



EDITION: U.S. | INTERNATIONAL | MÉXICO

Set edition preference

Sign up | Log in

SEARCH

Home Video NewsPulse | U.S. | World | Politics | Justice | Entertainment | Tech | Health | Living | Travel | Opinion | iReport | Money | Sports |

updated 11:36 p.m.EDT, Sun April 24, 2011

Make CNN Your Homepage

Giffords cleared to attend husband's shuttle launch

Rep. Gabrielle Giffords has recovered enough from her shooting injuries to attend the launch of the Space Shuttle Endeavour commanded by her husband, Mark Kelly, according to a source close to Giffords. FULL STORY

Latest news

Report: Major military operation in Syria
Man attacks Alitalia flight attendant
Air traffic resumes at St. Louis airport
LAPD chopper shot, forced to land
Body of trapped Idaho miner found
Pipe bomb suspect recently released

Ahead of the curve: The next 7 days

The world will be watching events unfold in Libya, the space shuttle launch and then there's a wedding you might have heard a little something about. Here's a look ahead to the next seven days. FULL STORY

Featured

Obamas attend Easter service

Reticent believer 'gets the Holy Ghost'

Rio's reimagined 'Passion Play'

Download a FREE audiobook today!

audible.com

Learn More

ADVERTISEMENT

Hi there!
Log in | sign up

FRIENDS' ACTIVITY

View more friends' activity | What's this?

Zhao Zhang shared Apple's Steve Jobs takes medical leave · about 3 months ago

Rep. Giffords cleared to attend shuttle launch
733 people shared this.

Figure 3.1 An example of targeted advertisement. A targeted advertisement is displayed on cnn.com in the upper right corner.

Figure 3.1 shows a targeted advertisement displayed on cnn.com in the upper right corner of the page. During this process, there are two important datasets used to predict a user’s interest, and a third dataset for the advertising bid request. The first of these datasets is the accumulated data about each user from their online search activities. This data includes cookie id, user’s search term, clicked link, date and time, IP address and so on. The second data stream indicates date and time of user purchase (called conversion) activities. The third dataset, the bid request, contains data to allow many different companies to bid on an ad on an individual user’s page view. This includes the topic of the page, the cookie id, the local time of day, the web location (url), and the size, type and location of the ad space.

3.2.2 User Segmentation

One of the crucial steps in Behavioral-targeting (BT) is to segment users according to their online interests or preferences. As a popular clustering algorithm, K-means [38] has been widely used to perform user segmentation in recent studies due to its quickness, good scalability and high efficiency in handling large datasets. Zheng et al. [78] applies K-means to cluster users by analyzing the characteristics of Web service and user's interests. The experimental results in their study indicate that they can effectively recommend web services to users by clustering users and establishing a recommendation service library. An empirical study conducted by Yan et al. [77] studies how BT can truly help online advertising in search engines. They use K-means for user segmentation and find that the user search behavior can be used to produce much better prediction accuracy than user browsing behavior, when used as user representation strategies for BT. A study presented in [72] also points out that ads need to be relevant to user's interests in order to increase the probability of ad clicks.

K-means based user segmentation also has been used to improve online recommendation systems by clustering users based on their historical data. Bouras et al. [13] incorporates an external knowledge source with K-means algorithm to cluster user's preferences and demonstrate its effectiveness on a recommendation engine. A similar work is found in [76] where a K-means based algorithm for mining user clusters is presented. In addition, K-means has also been applied in several studies on market segmentation [41][60].

Although K-means has been widely applied in user segmentation, most previous studies fail to take semantics of user behaviors into consideration, which makes it very hard to correctly segment users who have the similar interest but no common queries. In order to meet this challenge, this study proposes a topic based user segmentation by projecting user's queries to a topic level which allows mining of the semantics underlying user's behaviors.

In addition to traditional clustering approaches, Tyler et al. [71] consider user segmentation problems as a ranked retrieval task over an index of known users based on language modeling and vector space modeling. The experimental results show that both vector space and language models are able to perform well for the audience selection problem.

3.2.3 Demand-driven Taxonomy in BT

Currently, BT advertising inventory comes in the form of some kind of demand-driven taxonomy, which consists of BT categories designed to capture a broad set of user interests. Chen et al. [19] propose a Poisson model to estimate the click probability of a user, when shown a display advertisement in a BT category. In their work, ad clicks, page views and search queries are considered as three types of entities and a simple frequency-based feature selection method is adopted. Publicly available ontologies are also used to represent a user's interest. Wang et al. [73] build a hierarchical and efficient topic space based on Open Directory Project (ODP) ontology to match a user's photo tags with ads. The ads are represented in a topic space, and their topic distributions are matched with the target user interest.

However, the topics covered in the demand-driven taxonomy are not always comprehensive and need manual update over time. A taxonomy that works in one advertising system might not work in another, which makes the usage of behavioral targeting categories very limited across different domains.

3.2.4 Machine Learning Techniques in BT

Machine learning techniques have been leveraged in several prior works. Ranking SVM is applied in [45] to rank users according to their probability of interest in an advertisement. User's search query history and click history are used to create user profiles. Similarly, Ratnaparkhi et al. [58] propose a model that attempts to estimate the probability that a user will click a given ad shown on a page. In this work, the feature space is extracted by combining user search queries, the ad, and the page on which this ad is shown. Lacerda et al. [42] also propose a framework for associating ads with web pages based on Genetic Programming (GP). Their experimental results indicated that GP was able to discover effective ranking functions for placing ads in relevant web pages.

Recently, researchers have been looking at the ad targeting system from a high level: how to build a predictive model that can automatically handle hundreds of different and concurrent display ad targeting campaigns. Raeder et al. [57] propose four design principles for large-scale autonomous data mining systems and demonstrates the application of these principles within an automated ad targeting system. A challenge for the system is that each campaign may have a different performance criterion, and system needs to learn models automatically for each new campaign with minimal human intervention. These problems have also been described in detail previously in [52, 55].

3.3 Query Log Exploitation

With the creation of ever increasing volumes of digital data, the web search engines have become the most widely used tools for people to seek online information or service. Log files of the interaction between users and search engines are usually kept by web search engine companies and Internet service providers. These log files typically consist of a unique identifier for the user, the query string submitted by the user, a timestamp, and URLs clicked (if any) for that query. The earlier studies on query logs date back to late 1990s mainly focused on investigating important details of user's queries, such as query length distribution and number of clicked URLs [35, 65]. These studies provide important details of user's search behavior and have served as the foundation of later works on search query. Its related applications including query suggestion [11, 29, 74, 75], and search results re-ranking [26, 37, 67, 79].

However, the publicly available query log resources are fairly limited and dated. There are only a few query logs that can be used by researchers working outside search engine companies, such as query logs released by Excite [63], AlltheWeb [66], and AltaVista [34] from 1997 to 2002. The most recent publicly available query log for the academic community was released by AOL in 2006, which contains more than 30 million queries sampled in three months from over 650,000 users [51].

Most of the work on the exploitation of query logs tackles the problem of query similarities in order to expand query, provide query suggestion, or cluster queries for other applications. Cui et al. [25] point out that a document can be considered as relevant to a query, if the user clicks that document. They perform query expansion based on this idea. on click-through data, assuming that, terms which appear both in the queries and the clicked documents are somewhat related. Similarly, Huang et al. [33] propose a log-based approach to relevant term extraction and term suggestion, where they suggest the relevant terms for a user's query using those that co-occur in similar query sessions from search logs.

Mei et al. [49] describe a query suggestion algorithm which takes the hitting time on a large scale bipartite graph into consideration. Their method is able to control the sematic consistency of the suggested queries to the original query based on the computation of hitting time on large scale bipartite graphs. A similar work was developed by Liu et al. [44], where correlation among query log time series is applied to help identify semantically coherent clusters. They report that combining time-series and session similarity could lead to the best results for identifying semantically related queries.

Query logs have also been exploited in other applications. For instance, spelling correction problem is addressed by utilizing search query log [17, 24]. A technique to refine the ranking of search results for any given query by constructing the query context from search query logs is proposed by Zhuang et al. [79]. The analysis of query logs is also used to address the problem of query caching in order to reduce the computing and I/O requirements needed in [48], and a similar idea is also implemented by Qasim et al. [56] in recommender systems. Last, but not least, Chuang and Chieu [21] use query logs to facilitate the engineering process of constructing Web taxonomies based on a query-categorization approach.

3.4 Query Representation

Query representation has received increasing attention in recent years, in which a click graph, a bipartite graph, is the common model for describing the relationship between queries and clicked URLs. The edges in click graph connect a query with the URLs clicked by users, with two types of nodes: queries and URLs. The edges of a click graph capture certain semantic relations between the objects they represent [53]. For instance, two queries connected with the same URL, are more likely to be similar than two connected with different URLs. Craswell and Szummer [23] weight the edge by computing the total number of clicks from all users and applied Markov random walk to a large query log. For a give query, a probabilistic ranking of document is produced. Unlike Craswell and Szummer [23] who use the raw click frequency from a query to a URL, normalized click frequency is introduced in [49, 53] based on transition probability from clicks of many users.

The disadvantage with click graph is that, the information in query logs is sparse: given that there can be a huge number of URLs available for each query, it may not be trivial that a URL clicked for a query must appear in the list of results returned for that query. Another inherent disadvantage with click graph is the bias in the ranking of results returned by search engines, since users tend to click more on higher ranked URLs. Also, some malicious clicks could make the information in query logs very noisy.

There are several approaches that have been developed to avoid the sparsity issue in modeling the representation of queries on the click graph. Baeza-Yates et al. [8] propose a term-weight vector model for a query using the content of the clicked web pages. The weight for each term corresponds to the query frequency and the number of clicks on the web pages where that term appears. Thus, the similarity of two queries can be computed as the similarity of their vector representations. The assumption behind this idea is that semantically similar queries may not share query terms but they may share terms in the web pagess or their snippets that are clicked by users. In a later work [10], the authors introduce another way to represent queries in a natural vector space where queries are treated as points in a high dimensional space. Each unique URL is considered as a dimension and the weight associated with each dimension is assigned by the number of clicks on that URL. In this way, a query is based on all the different URLs in its URL cover. In addition, Poblete et al. [54] create a new query-set model based on frequent query patterns which outperform the traditional vector space model used for clustering and labeling documents. Instead of using text of the documents, the authors select a bag of query-sets as features, which is also a novel method to deal with the problem of document representation.

Since different users may have completely different search tasks underlying the same query, there are also several prior attempts on modeling queries for personalization [27, 68]. In [68], both the returned results of a query and a user's interaction history with the query are used to characterize queries. These features are also used to build predictive models to identify the queries that will benefit most from personalization. Similarly, Dou et al. [27] define click entropy of queries to indicate the variation in query clicks. They experimental results demonstrate the impact of different click entropy distribution on the click results.

3.5 Mining User Behavior

The online environment has changed significantly in the past decade, with dramatic growth in the capabilities users expect. Both the search results returned by search engine and content displayed in a web page are crucial for user's satisfaction with their online experience. User behavior contains valuable information which is usually described as a set of features in the user behavior "space" in both search and web browsing activities.

3.5.1 User Interaction with Search Engines

Accurate modeling of user interaction with search engines has important applications to ranking search results [2], personalization search [67], among others. Providing relevant search results to users has been a fundamental problem in information retrieval (IR). Traditional approaches mainly focus on the similarity of a search query and web pages [9, 20]. Nevertheless, user's implicit feedbacks have also been utilized to improve the rankings. For instance, Agichtein and Zheng [4] present an approach of leveraging user interactions with search engines to predict the "best bet" top results preferred by the users who have searched similar queries before. A background component (such as a user's query) and a relevance component (such as query-specific behavior indicative of the relevance of a result to a query) are represented as features. Then these features are correlated with the explicit user judgments for a set of training queries in order to learn to interpret the observed user behavior.

Hassan et al. [32] report that user behavior alone can give an accurate picture of the success of the user's web search goals, even without knowing the relevance of the returned results. The baseline methods used to compare with their approach include a set of static features and query-url relevance. A rich representation of user behavior is introduced by Fox et al. [28]. The features used to represent user search interactions included query-text features, clickthrough features, as well as browsing features. A similar representation that is used to estimate user preferences is described in [3]. More recently, Joachims et al. [37] perform eye tracking studies as an empirical assessment of interpreting click through evidence.

3.5.2 Web Browsing Activity

Web browsing activity has been extensively studied in recent years. Bucklin and Sismeiro [15] develop and estimate a model of the browsing behavior of users based on two basic aspects: the user's decisions to continue browsing or to exit the site, and the length of time spent viewing each page. Several studies have investigated the correlations between user's interest and user's web page activity. Claypool et al., [22] find that the time spent on a page, the amount of scrolling on a page, and the combination of the two have a strong positive relationship with explicit interest. In a similarly work, Goecks and Shavlik [30] measure user mouse and scrolling activity in addition to user browsing activity. They report that their system is able to predict the surrogate measurements of user interest based on their browsing behavior with a high accuracy.

User's web browsing activity is also used to identify web spam by Liu et al [46]. The authors extract three features from user behavior pattern analyses and exploit a large-scale web access logs. Machine learning techniques and descriptive analysis on user behavior features of web spam pages are applied to exploit the difference between web spam pages and ordinary pages in user behavior patterns.

3.5.3 Representing User's Online Behavior

One of the most common strategies for representing user's online behavior is to leverage historical search queries [19]. The raw search queries are specific in representing user's information need, but they are non-stationary. Hundreds of millions of new queries are submitted to search engines every day. The predication performance of a model built on users historical search queries could decrease dramatically when used to segment users in the future. For example, a model could be learned based on search queries of a group of people who bought tablet PCs online to segment users for tablet PCs ads delivery. If the model is built before the "iPad" is invented, the model would not be likely to identify the users who submit queries about iPad as potential tablet PC buyers after iPad is released. This is because "iPad" is not in feature space of the model before it is invented. However, users looking for information about an iPad probably are also interested in other tablet PCs and would have responded to other tablet PCs ads. A study carried by Kumar et al. [40] also indicate that more than half of search queries contain direct references to some type of structured object.

A taxonomy of topics is another widely used strategy for representing user's web behavior [14, 70]. The topics in a manually-built taxonomy are often static and they do not change fast. For example, one of the topics in the taxonomy could be "cameras". Nevertheless, the topics can be too broad and imprecise to represent a user's web activities. For example, it may not be enough to represent a user's interest as "cameras", if the user has searched information about Canon 60D or browsed pages about Canon 60D. In this case, the user might be particularly interested in Canon 60D, and representing the user's activities by a broad topic could result in information loss in the user data. These types of topics are presented in the bid request.

User's online behavior can also be considered as either "active" or "passive events" [6]. Active events include issuing search queries, browsing webpages, and clicking ads. Passive events include viewing ads and visiting pages in which an action is not specifically required upon seeing the page. In [5] several different events are used to model user's profile, each with a corresponding feature extraction method. The authors use a large scale real world benchmark to show the scalability of the proposed approach when the number of customized campaigns increases. The experimental results also indicate that short-term user history has a relatively higher importance over long-term user history when it comes to targeting. Archak et al. [7] compress individual user histories into a graph structure that represents local correlations between ad events. They also introduced several scoring rules to capture global role of ads and the ad paths in the graph, as well as the structural correlation between an ad impression and the user conversion.

In addition to search queries, the content of web pages visited by a user can also be used to learn a user interest. Kim et al. [39] propose to learn a user interest hierarchy (UIH) from a set of web pages visited by the user. The web page is assigned to nodes in the hierarchy for processing learning and predicting interests. They propose a divisive hierarchical clustering algorithm and evaluate their approach based on the data obtained from 13 users on their web server.

While most of previous work focuses on user's temporal interest, Ahmed et al. [5] propose a time-varying hierarchical user model which takes into consideration both the user’s long-term and short-term interests, with the purpose of generating user profile for behavioral targeting. They use a coherent approach based on Bayesian statistics and the experimental results indicate that their approach excels at the task of predicting user response for displaying advertising targeting. Similarly, Hassan et al. [32] build a sequence model that incorporates time distributions and their experiments result show that the sequence and time distribution models are more accurate than static models based on user behavior. They also show empirically that user behavior alone can give an accurate picture of the success of the user’s web search goals, even without considering the relevance of the document display.

Li et al. [43] also propose an adaptive scheme to learn the changes of users interest from click-history data. They introduce independent models for long-term and short-term user preferences to compose a user profile that contains a taxonomic hierarchy for long-term model and a recently visited page history buffer for the short-term model. The experimental results indicate that their scheme is sufficient to model the up-to-date user profile, and is able to achieve about 29.14% average improvement over the compared rank mechanisms.

Unlikely using search queries or web page visited to model user online behavior, Provost et al. [55] propose to take into consider user’s pages on social networking sites, photograph sites, non-professional blogs, etc. when modeling user profile. They introduce a method that extracts quasi-social networks from browser behavior on user-generated content sites, with the purpose of finding relevant users for brand advertising.

3.6 Summary

In this chapter, recent studies on behavioral targeting and query representation techniques are presented. Despite the fact that publicly available query logs are scarce and dated, they have shown to be useful for mining user behavior and tackling IR application problems. Query logs also help in understanding user online behavior which, in turn, helps in advertisement personalization. However, the existing studies rarely discuss the challenges with user queries in behavioral targeting advertising. Traditional user segmentation is based on Bag of Words model which fails to take into consideration the semantic relations among queries. This motivates the research questions presented in the previous chapter. In next chapter, a user intent representation strategy and a query enhancement mechanism are proposed to address the challenges with search query. It discusses the problem of binary user classification based on user's online intents and proposes an alternative to define user's search intents for evaluation purpose, in the case that the dataset is sanitized.

CHAPTER 4

INTENT-BASED USER CLASSIFICATION

4.1 Introduction

Online advertising spending has been increasing at an unprecedented pace over the past decade. In order to increase the effectiveness of targeting advertising, models are built based on user's web activities, such as search queries, to personalize advertisements. There are hundreds of companies and many different approaches (e.g., context, social, cookie-based, etc.) being developed to improve targeting advertising. The largest internet companies, such as Google, Facebook, and Yahoo, are all advertising companies. Data from search activities, web surfing and social connections are all mined to optimize online advertising effectiveness.

With the rapid advancement of the World Wide Web (WWW), accurate prediction of user's online intents underlying their search queries has been playing an important role in satisfying user's online experience. It has been helping advertisement campaigns to target more relevant users, publishers to recommend web content, search engines to return personalized results, and many other service providers to facilitate user's online experience. For instance, a user with a travel plan in mind would have a higher probability of clicking on a flight advertisement. Thus from a perspective of a flight advertiser, identifying users who are likely to travel could help targeted ad delivery and increase revenue. Similarly, if a content publisher knows a user's online intent, it can recommend relevant content to match the user's interest.

As a rich source of information on web searchers' behavior, query logs have been utilized by advertising companies to deliver personalized advertisements and leveraged by researchers to tackle other application problems, such as query suggestion. To carry out research on behavioral targeting, it is desirable to have golden standard datasets available, which contain both query logs and ad click information. These types of datasets are used by advertising companies to train and test a model that predicts user's ad click behavior. However, they are not available in the academic community, which makes conducting research in this area difficult. The publicly available query logs are small, dated, and sanitized, since search engine companies are reluctant to release complete query log data. It is understandable considering that query logs can reveal private information and cannot be thoroughly sanitized.

The first component of the system framework (highlighted in brown in Figure 4.1) is discussed in detail in this chapter below. It introduces a user intent representation strategy and proposes a query enhancement mechanism to address the challenges with search queries. The system first takes a query log and the external dataset Delicious to label the users and build a click graph which is then used to augment user's search query in the query enhancement mechanism. The enhanced query representation is then used to represent user's intents. It focuses on the problem of binary user classification based on user's online intent and provides the description of the datasets along with evaluation of the proposed method. This chapter also proposes an alternative to define user's search intent for evaluation purpose, in the case that the dataset is sanitized.

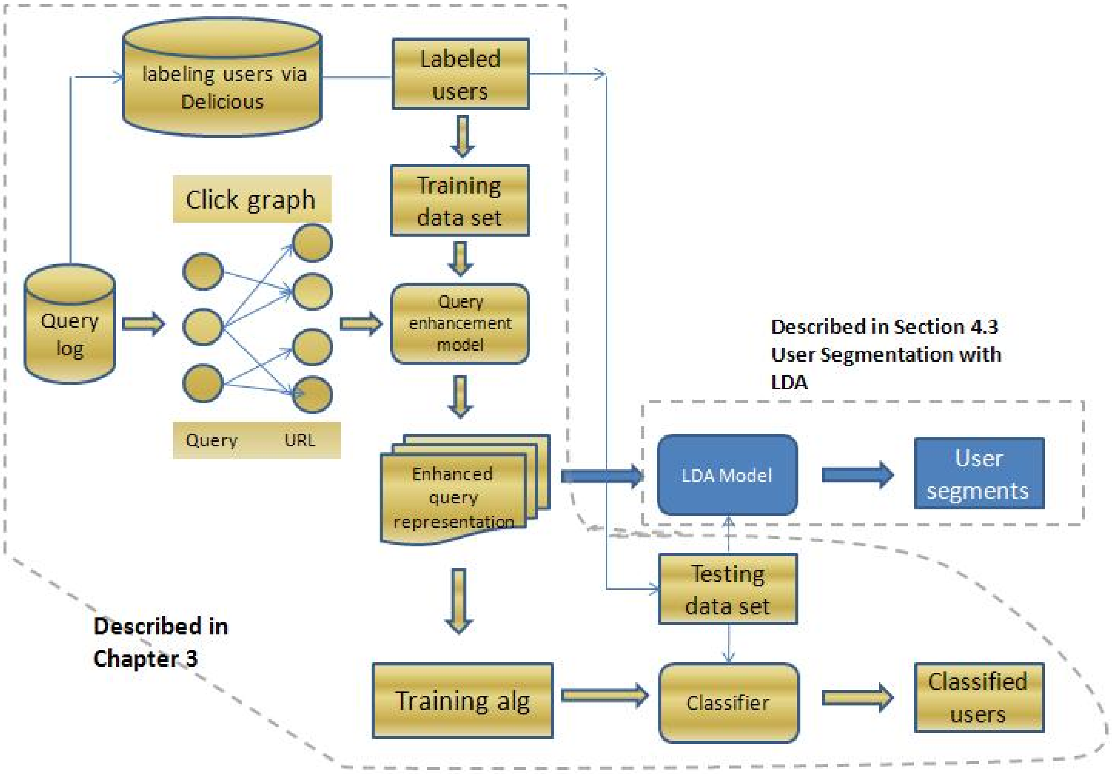


Figure 4.1 System framework. The first component of the system framework is highlighted in brown.

4.2 User Intent Representation

4.2.1 Baseline Model

In order to differentiate users by their online intents, the intent representation should consider user's online behavior which can be characterized by search queries. The queries issued by a user could contain hidden information about the user's intent. For example, queries like "map", "visa application" and "hotel reservation" have a strong indication that a user may also have an intention to purchase a flight, even if the user did not explicitly issue queries like "cheap flight" or "flight fares". Thus, a user's online intent can be built by considering all terms that appear in the user's queries.

Using Bag of Words (BOW) model [62], all users can be considered as a user-by-term matrix, where each row of the matrix is a user and each column of the matrix is a term. In this model, search queries are represented as a collection of terms that appear in the queries, without considering the order of terms. In this way, a user who issues query "new york weather" will have the same intent as the user who issues query "weather new york", because both of the users are represented as terms "new", "york", and "weather". Therefore, each distinct term can be treated as a feature while all distinct terms in user's queries consist of the feature space.

In the baseline model, each term is weighted by the classical Term Frequency Inverse Document Frequency (TFIDF), which is the product of two statistics: term frequency and inverse document frequency. Let t be a term and d be a collection of queries from a user. In this case, the term frequency tf(t,d) is the number of occurrences of the term t in a user's query collection d, while the inverse document frequency is defined as follows:

$$\text{idf(t,D)} = log\frac{|D|}{1+\ |\{d\in D{:}t\ \in d\}|} \tag{4.1}$$

where |D| is the total number of users, and $|\{d\ \in\ D{:}\, t\ \ \in\ d\}|$ is the number of users whose queries contain term t. Then the weight for each term can be calculated as:

$$\text{tf*idf(t, d, D)} = \text{tf(t,d)}\times\ \text{idf}(t,D) \tag{4.2}$$

Therefore, in the user-by-term matrix $R^{d \times t}$ where d is the total number of users and t is the total number of terms that appear in user queries, a user's intent can be represented as a real valued vector. Clearly, the weight for a term increases when the term has a high frequency in a user's queries but decreases when it appears in too many users' queries.

4.2.2 Query Enhancement by Leveraging Query Log

In the past decade, web search have grown at an unprecedented pace. Typically queries issued by users contain very few terms. In an empirical study [36], about 62% of all queries contained one or two terms, and fewer than 4% of the queries had more than six terms. On the average, a query only contained 2.21 terms, which can carry only a small amount of information about the user. The tendency of users to use short and ambiguous queries makes it difficult to fully describe and distinguish a user's intent. For instance, the user intent behind query "Steve Jobs" will be represented as two terms in the BOW model: "Steve" and "Jobs", along with their weights in the feature space, which could describe an intent of a user who is either interested in the person "Steve Jobs" or looking for a job.

Another important aspect of user's search query is that, the volume of queries is huge and follows the Zipf's law, where a small amount of queries appear very often while most queries appear only a few times. This makes the query feature space sparse and hence could undermine a classifier's performance in predicting future unseen data. For example, "laptops" and "cameras" are frequent queries and there are advertisers bidding ads on these queries. However, "T61" and "D60" are more specific queries with much fewer occurrences. Without knowing “T61” is a laptop model and “D60” is a camera model, these queries would not lead to more focused advertisements.

Therefore, the challenge with intent representation using user query is two-fold:

- **Short and ambiguous queries making it difficult to describe and distinguish a user’s intent.** In addition, the amount of queries issued by different users over a period of time greatly varies. Even less information can be captured from the users who issue only a couple of search queries in a given period of time, which makes the problem even more challenging.
- **Sparseness of query space.** While frequent queries usually can lead to targeted advertisement, those “tail” queries do not have enough statistical learning instances to “match” with advertisement.

To address this challenge, the *click graph* [23], a bipartite graph between queries and URLs, has been used to describe the connection between queries and URLs, where edges connect a query with a clicked URL. Figure 4.2 is an example of a click graph with three queries and four URLs. One of the most useful features in the click graph is that, the edges of the graph carry some semantic relations between queries and URLs. For instance, queries “Steve Jobs” and “Apple” are co-clicked with URL “www.apple.com”, and hence are related to each other. Clearly, this graph can be employed to augment query “Steve Jobs” with “Apple” to provide more information about the user’s intent. Therefore, it is important that the queries are represented in a way that the semantic relations between each query can be measured so that closely related queries can be captured.

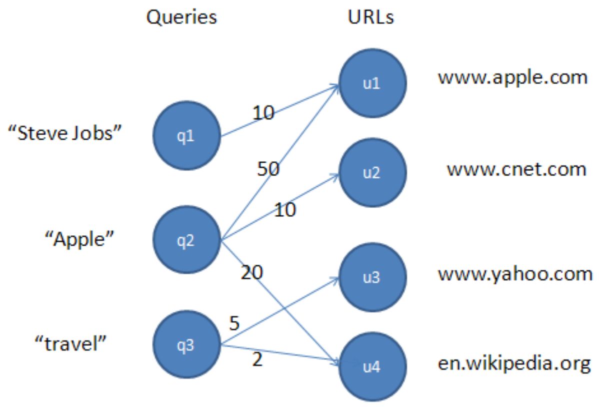


Figure 4.2 An example of click graph. The edges of the graph carry some semantic relations between queries and URLs.

Let *Q* = {q_1,q_2,…,q_i} be a set of i unique queries collected in a query log during a period of time. Let *U* = {u_1,u_2,…,u_j} be a set of j URLs clicked for these queries. For each edge (q_i, u_j), the click frequency are assigned as its weight to measure how frequent u_j was clicked by the user who issued query q_i. Intuitively, this click frequency *cf* can be considered as the Term Frequency in the classical TF*IDF model, where each query is a "document" and each URL is a "term". The click frequency matrix of Figure 4.2 is shown in Table 4.1.

Table 4.1 Click Frequency Matrix

	u_1	u_2	u_3	u_4
q_1	10	0	0	0
q_2	50	10	0	20
q_3	0	0	5	2

Similarly, the concept of inverse document frequency can be borrowed to measure the inverse query frequency, where the discriminative capability of a URL should be inversely proportional to entropy. Let $|I|$ be the total number of queries in the query log, and the inverse query frequency for the URL u_j is defined as:

$$iqf(u_j) = log \frac{|I|}{1+ |\{q \in Q: uj \in q\}|} \quad (4.3)$$

where $|\{q \in Q: uj \in q\}|$ is the number of queries that are associated with URL u_j. One of the important benefits of inverse query frequency, like inverse document frequency, is that it helps balance the bias of the clicks on those highly ranked URLs which usually tend to have more clicks (no matter whether those URLs are really relevant or not for that query).

To weight the edges in the click graph, a natural choice would be to incorporate the click frequency *cf* with inverse query frequency *iqf* in a similar TF*IDF model, which is defined as:

$$cf{*}iqf(q_i, u_j) = cf_{ij} \bullet iqf(u_j) \quad (4.4)$$

Therefore, each query qi can be represented as a vector where the feature space consists of URLs, and the weight can be measured by $cf{*}iqf(q_i, u_j)$.

As mentioned previously, the goal of query enhancement is to augment the query with closely related or similar queries. This is especially important for the queries that could lead to ambiguous meanings and for the users who only issued a few queries from which the user's intent can hardly be predicted due to the lack of information about the user. To measure the similarity between queries, the cosine function between two query vectors is adopted. It is calculated as:

$$\mathrm{Cos}(q_i,q_j) = \frac{\overrightarrow{qi}\cdot\overrightarrow{qj}}{||\overrightarrow{qi}||||\overrightarrow{qj}||} \tag{4.5}$$

where $\overrightarrow{q_\imath}$ indicates the vector of a query q_i.

After calculating the similarities between queries, for each query, the rest of the queries are ranked in the descending order of the similarities with the original query. The top *k* queries will be picked to augment the original query. Since the process can be executed offline with a large query log, the user's intent is represented by his/her issued queries along with the associated top *k* queries for each of the original query, and represent the terms in a BOW model. Table 4.2 illustrates an example of query enhancement results.

Table 4.2 Example of Query Enhancement Results

Query = microphone equipment
Stereo microphone Recording karaoke Audio gear Used microphone Digital recorder Microphone ebay Equalizer

4.2.3 Labeling Users

Before evaluating the impact of query enhancement on the user classification, it is important to label the positive users who have a specific online intent. The most straightforward way to identify the positive users is to see if the user has clicked a relevant ad. For instance, if a user clicks a flight ad, the user should be considered to have travel intent. However, as discussed previously, such datasets are not publicly available in academia, which makes it difficult to evaluate this approach. Therefore, one of the goals of this research is to come up with a reasonable alternative that defines a user's intent by utilizing external data.

An important aspect of user's online behavior is that, users tend to only make clicks on URLs which are of interest to them. Hence, it is reasonable to associate a user's online intent with the URLs clicked by that user. It is worth mentioning that a user may click multiple URLs during a period of time, and have multiple intents. This chapter aims to label the users by only considering one specific intent each time. However, it can be easily extended to other intents as explained later in this chapter.

Since the content of each URL can be described by different words or phrases, ideally each URL can be associated with a set of labels that cover the topics of the URL as comprehensive as possible. For example, the URL "www.united.com" is tagged with phrases such as "airline", "airfare", "travel", "flight", among many others. Therefore, Delicious, a social bookmarking web service is adopted as an external data source to identify the positive users and label them with a specific intent to build an evaluation dataset. It is one of the best researched folksonomy and each URL can be bookmarked and tagged by the entire community. When given a URL, it returns all the popular tags associated with that URL, which then can be used to match a selected intent.

From an advertiser's perspective, the title of an advertisement displayed to the users contains the information about the product or service that the advertiser wants to promote, while the keywords in the title reflect the user's intent if the user clicks the ad. Therefore, instead of arbitrarily defining an intent, the keywords in the title of an advertisement are used to indicate an online intent. For instance, keywords in the ad title "Cheap Flight Travel" can be used to label the positive users who have a travel intent and interested in purchasing cheap flight as follows.

Step 1: Remove stop words from ad title and extract the keywords.

Step 2: Get tags for each clicked URL from Delicious dataset.

Step 3: Tags and keywords stemming

Step 4: Get the URLs whose tags cover all the keywords extracted from the ad title. If none of the URLs has the tags that cover all the keywords, get the URLs whose tags cover the most of the keywords.

Step 5: Label the users as positive who have clicked any URLs from Step 4.

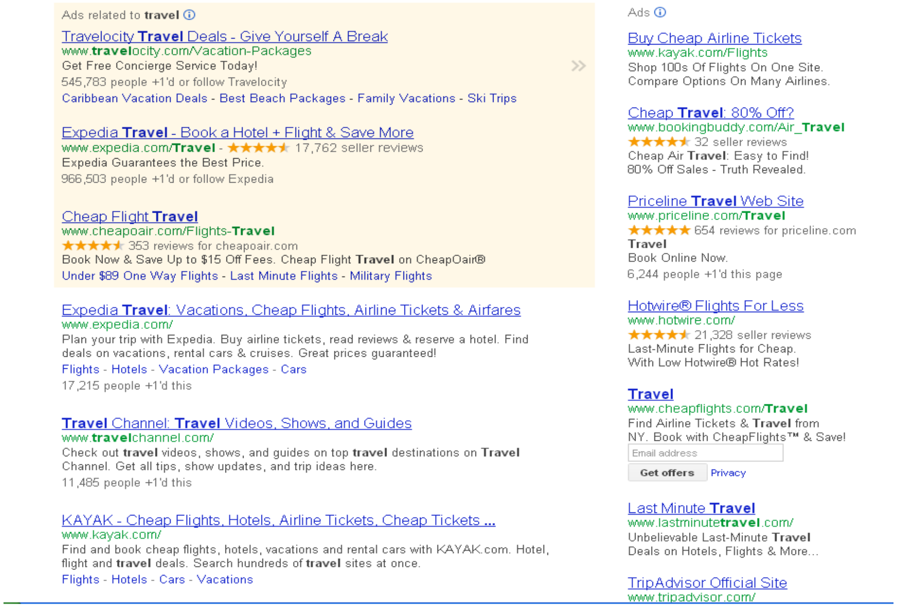
Lack of enough training datasets (labeled instances) could cause overfitting or high-bias when learning a classifier. There are three major benefits of using Delicious as an alternative to label users. Firstly, the tags associated with each URL are comprehensive, and can be added by any Delicious user. This is very important because it is unwise to miss out any positive users. Secondly, the dataset in Delicious is large and updated every day. Almost all of clicked URLs in the query log can be found in Delicious dataset. Finally, this approach does not need any manual effort while still creates reasonable training datasets for behavioral targeting research in academia. Table 4.3 demonstrates some of the URLs whose tags cover the keywords in the ad title.

Table 4.3 Examples of the URLs with Tags Cover the Keywords in the Ad Title "Cheap Flight Travel"

URLs	**Tags**
Kayak.co.uk	travel, flights, search, cheapflights, cheap, flight, comparison, airline, holiday, Tickets
travelzoo.com	travel, deals, airfare, flights, vacation, search, airline, shopping, cheap, shop
skyscanner.com	travel, flights, airfare, airlines, search, cheap, flight, airline, tickets, discount
jetblue.com	travel, airlines, flights, airline, airfare, usa, jetblue, cheap, inspiration, webdesign
airasia.com	travel, flights, asia, airlines, airline, thailand, malaysia, cheap, flight, lowcost
flycheapo.com	travel, airlines, lowcost, cheap, search, europe, airfare, airline, flight

4.3 User Classification

The performance of user classification has a great impact on the effectiveness of behavioral targeting advertising as it only makes sense to deliver ads to those who have an intent which is of interest to the advertiser. Ideally, an advertiser should be able to define an intent domain related to its product or service, and the user classifier automatically classifies a group of users based on this intent. Therefore, the user classifier discussed in this section makes binary decisions regarding whether a user has a particular intent which is indicated by the title of an advertisement. The proposed approach is evaluated in six domains: Travel, Jobs, Real estate, Automobiles, Diet, and Cameras, while this approach is general enough to be applied to other domains as well. Under each domain, the titles of the ads displayed on Google search are used as the specific intents to evaluate our approach. Figure 4.3 shows the returned search results for the query "Travel", where the sponsored ads are displayed on the top of the results and on the right-hand side of the page. The titles of these ads are then processed and used to label user's intent as described above in step 1 to step 5 in Section 4.2.3. The same method is used to evaluate the other five domains.



Ads related to **travel**

Travelocity **Travel** Deals - Give Yourself A Break
www.**travel**ocity.com/Vacation-Packages
Get Free Concierge Service Today!
545,783 people +1'd or follow Travelocity
Caribbean Vacation Deals - Best Beach Packages - Family Vacations - Ski Trips

Expedia **Travel** - Book a Hotel + Flight & Save More
www.expedia.com/**Travel** - ★★★★★ 17,762 seller reviews
Expedia Guarantees the Best Price.
966,503 people +1'd or follow Expedia

Cheap Flight **Travel**
www.cheapoair.com/Flights-**Travel**
★★★★★ 353 reviews for cheapoair.com
Book Now & Save Up to $15 Off Fees. Cheap Flight **Travel** on CheapOair®
Under $89 One Way Flights - Last Minute Flights - Military Flights

Expedia **Travel**: Vacations, Cheap Flights, Airline Tickets & Airfares
www.expedia.com/
Plan your trip with Expedia. Buy airline tickets, read reviews & reserve a hotel. Find deals on vacations, rental cars & cruises. Great prices guaranteed!
Flights - Hotels - Vacation Packages - Cars
17,215 people +1'd this

Travel Channel: **Travel** Videos, Shows, and Guides
www.**travel**channel.com/
Check out **travel** videos, shows, and guides on top **travel** destinations on **Travel** Channel. Get all tips, show updates, and trip ideas here.
11,485 people +1'd this

KAYAK - Cheap Flights, Hotels, Airline Tickets, Cheap Tickets ...
www.kayak.com/
Find and book cheap flights, hotels, vacations and rental cars with KAYAK.com. Hotel, flight and **travel** deals. Search hundreds of **travel** sites at once.
Flights - Hotels - Cars - Vacations

Ads

Buy Cheap Airline Tickets
www.kayak.com/Flights
Shop 100s Of Flights On One Site.
Compare Options On Many Airlines.

Cheap **Travel**: 80% Off?
www.bookingbuddy.com/Air_**Travel**
★★★★★ 32 seller reviews
Cheap Air **Travel**: Easy to Find!
80% Off Sales - Truth Revealed.

Priceline **Travel** Web Site
www.priceline.com/**Travel**
★★★★★ 654 reviews for priceline.com
Travel
Book Online Now.
6,244 people +1'd this page

Hotwire® Flights For Less
www.hotwire.com/
★★★★★ 21,328 seller reviews
Last-Minute Flights for Cheap.
With Low Hotwire® Hot Rates!

Travel
www.cheapflights.com/**Travel**
Find Airline Tickets & **Travel** from
NY. Book with CheapFlights™ & Save!

Last Minute **Travel**
www.lastminute**travel**.com/
Unbelievable Last-Minute **Travel**
Deals on Hotels, Flights & More...

TripAdvisor Official Site
www.tripadvisor.com/

Figure 4.3 Travel related ads. The sponsored ads are displayed on the top of the results and on the right-hand side of the page

4.3.1 Datasets

In this study, AOL query log is used to perform user classification. It is the most recent publicly available query log for the academic community that was released by AOL in 2006, which contains more than 30 million queries sampled in three months from over

650,000 users [51]. The dataset includes AnonID, Query, QueryTime, ItemRank, ClickURL and Time. The detailed data format is summarized in Table 4.4.

Table 4.4 Detailed Dataset Format

AnonID	An anonymous user ID number
Query	The query issued by the user
QueryTime	The time at which the query was submitted for search
ItemRank	The rank of the URL if clicked
ClickURL	Clicked URL
QueryTime	The time at which the query was submitted

In order to avoid noise, the users who have more than 1000 clicks within one day are filter out (they are most likely robots). In addition, stop words, punctuation marks and queries that appear less than 2 times are also removed. A quarter of the AOL dataset is taken to performance query enhancement, which contains 220,138 unique queries and 233,291 unique URLs. For the rest of the AOL dataset, 5000 users who fulfill both of the following two conditions are randomly picked for each intent classification experiment.

a) The users have issued queries in the first 7 days (01 March – 07 March)

b) The users have clicked URLs after the first 7 days (08 March – 31 May)

The queries issued in the first 7 days are used to build the bag of words representation and the URLs clicked after the first 7 days along with the Delicious dataset are used to label the users for each intent. After the preprocessing, 5000 labeled users are collected and each of them is represented by the bag of words model as a baseline. To compare with the baseline, query enhancement is applied before building the bag of words model, and *k* is set to be 10.

4.3.2 Evaluation Metrics

For each user classification experiment, the goal is to exam how our approach compared with the baseline model. After enhancing user's query, each user's intent is represented in a BOW model (as opposed to using user's raw queries in the baseline model which is introduced in Section 4.2.1). The users are classified based on the different online intents across six domains. The dependent variable in logistic regression is used to indicate the label of the user, while the independent variables are the TF*IDF values of the words in the BOW model. After fitting the logistic regression model on the training data, the coefficients of the independent variables are learned. The evaluation metrics used in this experiment is the positive precision.

The reason why positive precision is used in this study is that advertisers always want to deliver ads to those who have a high probability of having an intent related to the product. With a given advertising budget and the cost of displaying their ad to a user, advertisers tend to focus on the precision of positive users. Precision has been widely used as an evaluation metric in prior works on online advertising [42, 59, 73], while other studies tend to use click-through-rate (CTR) as their evaluation metric [18, 42]. However, CTR cannot be directly measured by using the datasets in this study, because the AOL datasets do not contain user's ad click data.

The performance of the classification is evaluated on each of testing datasets through filling the table as below.

	Labeled user class	
Predicted positive	tp	fp
Predicted negative	fn	tn

The positive precision is defined as:

$$Positive\ precision = \frac{tp}{tp+fp} \tag{4.6}$$

4.3.3 Experimental Results

As discussed in the previous section, the title of the advertisement is used to indicate a specific online intent for the evaluation purpose. In order to make the experiments fair, all the ads are used as different intents across the six domains to evaluate the proposed approach. More specifically, the titles of the ads used in the experiments are listed in Table 4.5, Table 4.6, Table 4.7, Table 4.8, Table 4.9 and Table 4.10.

Table 4.5 Travel Related Ads Title

1	Travelocity Travel Deals - Give Yourself A Break
2	Expedia Travel - Book a Hotel + Flight & Save More
3	Travelocity Travel Deals - Travelocity.com
4	Cheap Flight Travel
5	Buy Cheap Airline Tickets
6	Cheap Travel: 80% Off?
7	Priceline Travel Web Site
8	Hotwire® Flights For Less
9	Travel
10	Last Minute Travel
11	TripAdvisor Official Site

Table 4.6 Job Related Ads Title

1	New Jersey Jobs - Your New Job is right Around the Corner
2	Find Jobs - Find Job Openings In Your Area
3	Find Jobs in Your Area - indeed.com
4	New Jersey Jobs (Hiring)
5	Local Jobs Hiring Now
6	CareerBuilder Job Search
7	10 Best Job Search Sites
8	2013 Jobs Hiring $25+/Hr

Table 4.7 Real Estate Related Ads Title

1	New Jersey Real Estate - remax.com
2	Real Estate - Weichert.com
3	Real Estate For Sale - Zillow.com
4	Coldwell Banker
5	Century 21 Official Site
6	RealEstate.com
7	HUD Homes low as $10,000
8	MLS.com -Search for homes
9	Real Estate in NJ

Table 4.8 Automobiles Related Ads Title

1	Elmwood Park Auto Mall
2	Auto For Sale List
3	NJ Used Cars for Sale
4	Auto Loans USA
5	2014 New Chrysler Models

Table 4.9 Diet Related Ads Title

1	15-Day Weight Loss Trial
2	“Garcinia Cambogia” on Oz
3	Jenny Craig official Site
4	“Green Coffee Diet” on Oz
5	#1 The Fresh Diet
6	Weight Loss - Warning
7	Free Custom Diet Plans

Table 4.10 Cameras Related Aads Title

1	Panasonic Digital Cameras – New Advanced Lumix Digital Cameras
2	2014 Best Cameras
3	Cameras Store
4	Coldwell Banker
5	Digital SLR Camera
6	Digital Camera Mobile Lab

Tables 4.11 to 4.16 demonstrate the user classification results based on a 5-fold cross validation in six domains.

Table 4.11 User Classification Results in Travel Domain

Travel											
ads	1	2	3	4	5	6	7	8	9	10	11	Avg.
Baseline	0.593	0.580	0.657	0.745	0.714	0.770	0.742	0.814	0.829	0.710	0.693	0.713
QueryEnhancement	0.637	0.631	0.710	0.793	0.778	0.825	0.790	0.878	0.860	0.762	0.746	0.764
Difference	0.044	0.051	0.053	0.048	0.064	0.055	0.048	0.064	0.031	0.052	0.053	0.051

Table 4.12 User Classification Results in Job Domain

Jobs									
ads	1	2	3	4	5	6	7	8	Avg.
Baseline	0.614	0.626	0.718	0.707	0.748	0.708	0.723	0.714	0.694
QueryEnhancement	0.662	0.680	0.749	0.772	0.809	0.741	0.779	0.786	0.747
Difference	0.048	0.054	0.031	0.065	0.061	0.033	0.056	0.072	0.053

Table 4.13 User Classification Results in Real Estate Domain

Real Estate										
ads	1	2	3	4	5	6	7	8	9	Avg.
Baseline	0.731	0.695	0.634	0.710	0.689	0.758	0.713	0.680	0.736	0.705
QueryEnhancement	0.811	0.743	0.696	0.758	0.724	0.802	0.766	0.722	0.814	0.759
Difference	0.08	0.048	0.062	0.048	0.035	0.044	0.053	0.042	0.078	0.054

Table 4.14 User Classification Results in Automobiles Domain

Automobile						
ads	1	2	3	4	5	Avg.
Baseline	0.725	0.714	0.638	0.749	0.802	0.725
QueryEnhancement	0.790	0.745	0.756	0.820	0.865	0.795
Difference	0.065	0.031	0.118	0.071	0.063	0.070

Table 4.15 User Classification Results in Diet Domain

Diet								
ads	1	2	3	4	5	6	7	Avg.
Baseline	0.643	0.722	0.641	0.748	0.751	0.735	0.731	0.710
QueryEnhancement	0.687	0.771	0.701	0.768	0.815	0.794	0.824	0.765
Difference	0.044	0.049	0.060	0.020	0.064	0.059	0.094	0.055

Table 4.16 User Classification Results in Camera Domain

Camera							
ads	1	2	3	4	5	6	Avg.
Baseline	0.690	0.646	0.687	0.729	0.648	0.703	0.684
QueryEnhancement	0.775	0.732	0.742	0.766	0.705	0.774	0.749
Difference	0.085	0.086	0.055	0.037	0.057	0.071	0.065

Table 4.17 Summary of User Classification Results across Six Domains

	Travel	Jobs	Real estate	Auto	Diet	Camera	Avg.
Baseline	0.713	0.694	0.705	0.725	0.710	0.684	**0.705**
QueryEnhancement	0.764	0.747	0.759	0.795	0.765	0.749	**0.763**
Difference	0.051	0.053	0.054	0.070	0.055	0.065	**0.058**

In all of the six domains, the performance of the proposed user classification compared to the baseline model is statistically significant at two-tailed p value < 0.05, using a paired t test. This suggests that, by incorporating the proposed query enhancement in user classification, the performance of intent-based user classification can be significantly improved. Table 4.17 demonstrates the summary of user classification results across the six domains. The average difference in classification performance across six domains is 0.058, which yields 8.2% improvement compared with the baseline. The proposed query enhancement approach not only improves user classification performance, it also has a great impact on user segmentation performance, which will be discussed in detail in the next chapter.

The amount of labeled instances (training data) is vital to any classification problems. In this experiment, Delicious dataset is used as an alternative to label users. The tags associated with each URL are comprehensive, and can be added by any Delicious user. This is very important because it is unwise to miss out any positive users. In addition, the proposed user labeling approach does not need any manual effort while still creates reasonable training datasets for behavioral targeting research in academia. However, there are also several limitations involved in the Delicious dataset, which will be discussed in detail in Section 6.2.

In the process of query enhancement, top *k* similar queries are added to the original query. Based on empirical results, *k* is set to be 10 in this experiment. In order to achieve optimal classification results, two factors need to be considered when determining *k*: the size of the datasets and the computing resources. In practice, additional empirical effort needs to be devoted in order to achieve optimal results. Further discussion on this issue can be found in Section 6.1.

The logistic regression is adopted as the classifier in the experiments because it is a probabilistic classifier and uses a logistic function ranging from 0 to 1. The output can be simply considered as probability distributions. This also helps advertisers decide how much they should bid to show the ad based on the probability in the real time bidding system. It is worth mentioning that the advertisement titles used in this experiment are all from real ads displayed in the search results on Google, and the experiment can be easily extended to other domains. After the advertiser decided the title of the ad he or she wants to display, the classifier can be trained offline and a new user can be classified as interested in the ad or not interested in the ad automatically. This improvement of user classification can greatly help advertisers deliver their ads to users who are likely to be interested in their ads, and hence, click the ads.

This chapter focuses on binary user classification while next chapter will investigate the impact of the proposed query enhancement on user clustering under a topic model. In behavioral targeting advertising, users are grouped into different segments and advertisers always want to deliver ads to the users in the segment where the users are more likely to be interested in their products or services. Therefore, the next chapter formulates the user clustering problem from a behavioral targeting perspective, and describes a user segmentation approach based on Latent Dirichlet Allocation (LDA), where the semantics of user behaviors are taken into consideration.

CHAPTER 5

USER CLUSTERING FOR BEHAVIORAL TARGETING ADVERTISING

5.1 Introduction

The previous chapter proposed a user intent representation strategy and a query enhancement mechanism to address the challenges with search query. The experimental results demonstrated that users can be better classified based on their online intents by applying query enhancement. As discussed in Chapter 2, a user needs to be classified in the real time bidding while the advertiser bids to show the ad based on the likelihood the user has a specific intent. On the other hand, users can also be grouped offline in advance for behavioral targeting. If the publisher knows a user belongs to a segment where the users in that segment tend to be interested in a particular product, the publisher can target related ads to that user. This chapter, therefore, aims to answer the third research question:

Does the intent-based user segmentation improve the performance of behavioral targeting significantly?

Publishers and other service providers always want to have their ads displayed to the most relevant users in sponsored search. From an online service provider's perspective, it could be extremely useful to identify users who have a high probability of clicking its ads and display them in the sponsored search results. Therefore, the goal of this chapter is to improve grouping of the similar users into segments according to their online intent and to determine whether the new segmentation approach yields better segmentation results (PUR – Positive User Rate).

Behavioral Targeting aims to deliver relevant ads to potential consumers by analyzing user's online behavior. One of the curial problems in Behavioral Targeting is user segmentation with the purpose of grouping users into user segments with similar behaviors. If users with similar purchase intentions are successfully clustered into the same segment, an advertiser can potentially better profit from their online campaigns, as the ads are all delivered to the users who are more likely to click on the ad and convert than other users. At the same time the users may have better online experience as well, because the ads displayed to them are relevant to their interests. Thus, user segmentation has a great impact on the performance of behavioral targeted advertising and it is worth investigating how much intent-based user segmentation can help behavioral targeting.

This chapter refers to the preliminaries and datasets explained in Chapter 3. The problem of user segmentation is formulated as follows. For a given set of online users, each user's historical online behavior such as search queries are used to depict his/her interests. Each user is labeled either as a positive user or negative user for a given advertisement using the approach introduced in Section 4.2.3. The objective is to group all users into appropriate segments based on their search queries with the purpose of improving positive user rate (PUR) in any segment, which could in turn improve the ad click probability within those user segments as opposed to the massive and irrelevant ads. The main challenge with using user queries for segmentation is that, users who have the same online intent but have no common queries between them can be very hard to be grouped into the same segment. To overcome the disadvantages of traditional Vector

Space Model [61] which fails to exploit the semantic relation between user queries, the proposed approach is motivated by the use of topic models in the field of information retrieval and adopts Latent Dirichlet Allocation (LDA) [12] to represent user's online intent. LDA has been widely used in the field of information retrieval which effectively mines the relationship between words and documents with a hidden variable known as topic. Under the LDA model, the relationship between users and queries can be considered parallel to documents and words. Note that the query enchantment mechanism proposed in Chapter 3 is also applied to process user's query prior to building the LDA model for user intent representation.

The rest of this chapter is organized as follows. In the next section, a brief background on user segmentation is reviewed. Section 5.3 formulates the problem and describes the proposed user segmentation approach based on LDA. Finally Section 5.4 presents the experimental configuration and results along with analysis.

5.2 User Segmentation Background

Behavioral targeting is an online advertising methodology that aims to deliver personalized advertisement based on user's online behavior. It has been receiving more and more attention in advertising industry where a fair amount of commercial systems using behavioral targeting have been developed, such as Yahoo! Smart ads [64], which allows advertisers to target relevant users based on demographic and geographic, Doubleclick [1], which integrates special features such as user's browser type and operation systems to improve user segmentation, and Burst [16], which uses online survey for behavioral targeting.

Instead of relying on the contextual information of web pages for ad delivery, behavioral targeting enables advertisers to target advertisement to the audience who are more likely to be interested in the content of the ads by leveraging user's historical online behavior such as their queries submitted to search engines. Due to "one size fits all" problem that exists in most of the traditional online advertising methods, behavioral targeting has been playing an important role in deliver the right ads to the right audience. In recent years, web service providers, such as search engines and websites, have all started analyzing user's online behavior in order to provide a more satisfactory online experience for users and improve the effectiveness of advertising campaign for advertisers.

Traditional user segmentation approach for behavioral targeting includes the following three types: manual user segmentation, user classification, and user clustering. Manual rule-based user segmentation requires human effort to segment users manually which is time consuming. This method is rarely used by the current commercial systems because of the large scale of the data used for behavioral targeting in real life. The previous chapter discussed user classification for online advertising, and this chapter focuses on user clustering.

As mentioned earlier, user segmentation is a key process in behavioral targeting. The goal is to guarantee that users with similar online intents are grouped in the same segment. However, that information cannot be derived directly. The most common way is using the user behavior to represent user interests and purchase intents. Therefore, the assumption here is users with similar web behaviors have similar intents. In this way, user segmentation for behavioral targeting can be achieved by assigning each user in one segment where the users with similar behaviors are in the same segment. Since advertisers always tend to choose the most relevant segments to target their advertisements, the positive user rate in the segment is extraordinarily crucial for the effectiveness of the online campaigns.

5.3 User Segmentation with LDA

The problem with traditional Bag of Words model for user segmentation is that, the segmentation is only based on the ‘content’ of user’s queries, without considering the semantic relation between queries. This leads to the fact that users who have the similar online intent but have no common queries between each other can be very difficult to be grouped into the same segment. To address this challenge, this chapter proposes to project user’s queries to a topic level which allows mining of the semantics underlying user’s query. The second component of the system framework (highlighted in brown in Figure 5.1) is discussed in detail in this chapter below.

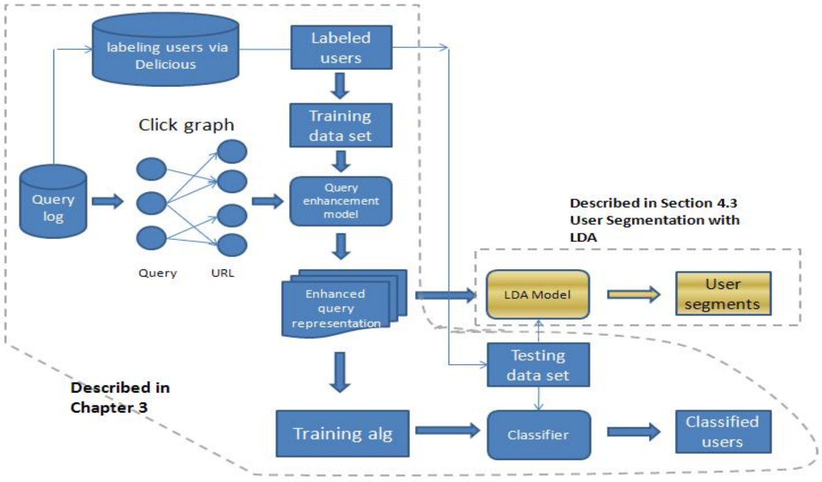


Figure 5.1 System framework. The second component of the system is highlighted in brown.

The proposed approach is motivated by the use of topic models in the field of information retrieval and adopts Latent Dirichlet Allocation (LDA) to present user's online intent on a topic level. Note that the query enchantment mechanism proposed in Chapter 4 is also applied to process user's query prior to building the LDA model for user intent representation.

5.3.1 LDA Model

Latent Dirichlet Allocation (LDA) [12] is a generative probabilistic model for collections of documents, where it considers every document as a distribution over the topics in a corpus and every topic as a distribution over the words of the vocabulary. Figure 4.2 is the graphical model representation of LDA, where M denotes the number of documents; N is the number of words in a document; θ_d is the topic distribution for document D; and z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document, while α and β are the corpus-level parameters, which can be assumed to be sampled once in the process of generating a corpus. The key inferential problem in LDA is to find the posterior distribution of the hidden variables given a document:

$$P(\theta,z|w, \alpha, \beta) = \frac{p(\theta,z,w|\ \alpha,\beta)}{p(w|\ \alpha,\beta)} \tag{5.1}$$

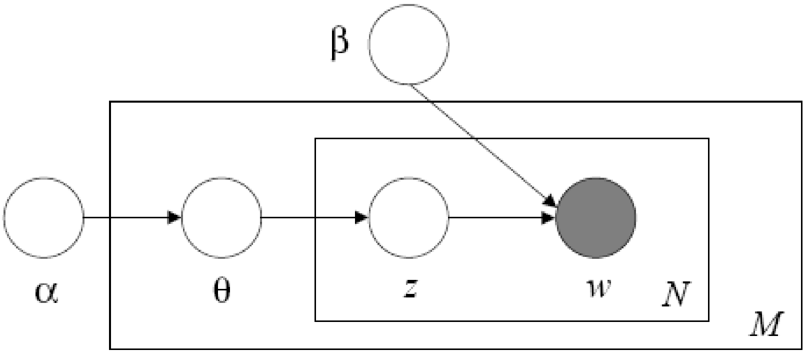


Figure 5.2 Graphical model representation of LDA. M denotes the number of documents; N is the number of words in a document.

The idea behind LDA is that documents can be represented as random mixtures over latent topics, and the topic, on the other hand, is characterized by a distribution over words. LDA assumes the following generative process for each document w in a corpus D:

1. Choose N ~ Poisson(ξ)
2. Choose θ ~ Dir(α)
3. For each of the N words $w_{n:}$
   (a) Choose a topic z_n ~ Multinomial(θ)
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, which is a multinomial probability conditioned on the topic z_n

In LDA, words are assumed to be generated by topics while those topics are infinitely exchangeable within a document. Thus, the probability of a sequence of words and topics follows the following form:

$$\mathrm{p(w, z)} = \int p(\theta)(\prod_{n=1}^{N} p(z_n \mid \theta)\mathrm{p}(w_n|z_n))d\theta \tag{5.2}$$

where θ is the random parameter of a multinomial over topics.

By marginalizing over the hidden topic variable z, LDA can also be understood as a two-level model. The word distribution:

$$p(w|\theta, \beta) = \sum_z p(w|z, \beta) p(z|\theta) \tag{5.3}$$

The following steps define generative process for a document w:

1. Choose θ ~ Dir(α)
2. For each of N words w_n:

   Choose a word w_n from $p(w_n \mid \theta , \beta)$.

In this way,

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N} p(w_n| \theta, \beta)\right)d\theta \tag{5.4}$$

It is worth mentioning that, a simple clustering model tends to only involve a two-level model where Dirichlet is sampled once for a corpus, a multinomial clustering variable is selected once for each document in the corpus, and a set of words are selected for the document conditional on the cluster variable [12]. As a result, in a simple clustering model a document only can be associated with a single topic.

5.3.2 LDA based User Segmentation

As discussed in Chapter 4, all terms that appear in the user's queries are used to build user's online behavior, and all users can be considered as a user-by-term matrix, where each row of the matrix is a user and each column of the matrix is a term. For a specific online advertisement, each user can be labeled as a positive user or a negative user using the method proposed in Subsection 4.2.3.

Under LDA model, each document can be represented as a probability distribution over topics. Note the fact that a query consists of terms, thus in the context of user segmentation, a user u_i is treated as a document d_i while each query q_{ij} issued by the user u_i is treated as a word in the corresponding document. Therefore, each topic z_i can be considered as a segment, and each user can be assigned into the topic segment that gives the highest probability.

5.4 Experiments and Evaluation

In this section, the same datasets from Chapter 4 are used to carry out user segmentation experiments for each of the intents across the same six domains in Chapter 4, which describes the datasets and data processing in detail in Subsection 4.4. Under each experiment, after computing all $p(z_k|u_i)$ (the probability a user u_i belongs to topic z_k) where k is the number of topics and i is the number of users, each user is assigned into the topic group that gives the highest probability.

5.4.1 Evaluation Metrics

The positive user rate in segment k is defined as:

$$PUR(S_k) = \frac{\#\text{ of positive users in } S_k}{\#\text{ of all users in } S_k} \tag{5.5}$$

while PUR over all users before segmentation as:

$$PUR = \frac{\text{\# of all positive users}}{\text{\# of all users}} \tag{5.6}$$

Because online service providers always aim to target the user segment with highest PUR, the segment PUR(S_k) is chosen when calculating the PUR improvement as:

$$\triangle\ (PUR) = \frac{PUR(S_k) - \ PUR}{PUR} \tag{5.7}$$

PUR(S_k) is determined by the following two constraints:

a) Maximum: choosing the segment that has the maximum PUR. This is reasonable since service providers always tend to recommend the user segment that has the highest ad click probability to advertiser for ads delivery.

b) Majority: the number of users in this segment cannot be less than average. This condition is also necessary, because it reduces some special situation. For example, some user segments may only have 1 user and he/she is a positive user. Obviously, this segment cannot be recommended to the advertiser even though it has the highest PUR.

While one of the objectives of this study is to compare the baseline performance with the proposed user segmentation approach, it would also be interesting to examine if the proposed query enhancement mechanism can improve the performance of the baseline in the context of user segmentation for behavioral targeting. As discussed in Section 3.2.2, K-means, a popular clustering algorithm, has been widely used in recent studies on user segmentation. It is also a fast clustering algorithm with good scalability and high efficiency. The assumption behind K-means is that clusters in the data are more or less spherical, ideally normally distributed. Since the user intents are presented as

TF*IDF vectors, a Mardia's test [47], a method of assessing the degree to which multivariate data deviate from multinormality, has been performed to make sure that the TF*IDF vectors follow a multivariate normal distribution. Therefore, K-means is adopted as a baseline to carry out experiments on user segmentation. More specifically, two experiments are carried out: the proposed LDA based user segmentation vs. K-means based user segmentation, and K-means based user segmentation vs. K-means based user segmentation with the proposed query enhancement mechanism.

5.4.2 User Segmentation Results

For each experiment, PUR improvements under different numbers of segments are investigated. The experimental results are shown in Tables 5.1 to Tables 5.6.

Table 5.1 User Segmentation Results in Travel Domain

Travel												
ads	1	2	3	4	5	6	7	8	9	10	11	**Avg.**
5 segments												
Kmeans	45.4%	50.8%	44.2%	51.4%	46.1%	42.9%	41.0%	43.2%	52.3%	48.9%	43.5%	**46.3%**
Kmeans+QE	49.7%	56.3%	52.1%	57.6%	53.5%	48.8%	50.6%	54.8%	60.5%	55.2%	51.7%	**53.7%**
LDA+QE	54.5%	60.3%	58.7%	68.0%	66.4%	61.2%	57.3%	64.0%	67.5%	62.0%	60.7%	**61.9%**
10 segments												
Kmeans	57.7%	52.0%	53.7%	58.5%	54.9%	61.0%	50.3%	57.1%	56.6%	52.8%	51.2%	**55.1%**
Kmeans+QE	63.6%	58.3%	61.4%	65.6%	63.1%	66.2%	59.8%	64.9%	65.5%	62.5%	60.4%	**62.9%**
LDA+QE	76.1%	69.2%	74.0%	83.3%	81.5%	79.4%	74.4%	80.2%	73.9%	80.2%	77.4%	**77.2%**
20 segments												
Kmeans	64.2%	71.2%	70.0%	74.8%	80.7%	74.5%	73.4%	79.4%	82.2%	72.8%	70.7%	**74.0%**
Kmeans+QE	73.0%	80.8%	77.2%	90.2%	89.9%	83.6%	88.4%	83.0%	91.5%	84.0%	79.5%	**83.8%**
LDA+QE	89.7%	93.6%	88.4%	102.9%	103.2%	98.0%	95.8%	95.3%	106.4%	98.1%	92.9%	**96.8%**
40 segments												
Kmeans	92.6%	88.0%	87.8%	96.4%	95.5%	89.1%	87.7%	90.4%	94.1%	83.9%	98.8%	**91.3%**
Kmeans+QE	103.4%	94.7%	98.4%	112.4%	105.8%	97.6%	94.2%	101.6%	107.3%	95.2%	111.5%	**102.0%**
LDA+QE	114.9%	107.0%	120.5%	130.3%	127.8%	123.2%	108.6%	118.9%	122.4%	115.0%	126.4%	**119.6%**

Table 5.2 User Segmentation Results in Job Domain

Job									
ads	1	2	3	4	5	6	7	8	**Avg.**
5 segments									
Kmeans	42.6%	39.5%	45.2%	43.1%	37.7%	40.4%	42.3%	41.5%	**41.6%**
Kmeans+QE	50.5%	44.7%	51.6%	49.4%	47.6%	53.4%	55.1%	53.3%	**50.7%**
LDA+QE	54.2%	51.0%	60.2%	56.8%	59.0%	64.5%	61.2%	64.8%	**59.0%**
10 segments									
Kmeans	47.2%	50.2%	62.3%	55.4%	57.9%	62.9%	52.7%	51.5%	**55.0%**
Kmeans+QE	56.0%	61.5%	67.4%	62.4%	63.0%	73.5%	64.7%	61.7%	**63.8%**
LDA+QE	67.8%	70.2%	75.0%	69.3%	76.1%	81.4%	72.5%	72.8%	**73.1%**
20 segments									
Kmeans	63.3%	68.7%	74.2%	71.5%	67.7%	72.1%	70.9%	66.8%	**69.4%**
Kmeans+QE	78.5%	77.5%	80.5%	82.5%	78.9%	83.4%	84.8%	82.6%	**81.1%**
LDA+QE	92.0%	88.4%	91.6%	94.9%	99.2%	97.0%	94.4%	91.9%	**93.7%**
40 segments									
Kmeans	85.8%	71.9%	92.4%	86.2%	89.5%	86.4%	90.6%	78.6%	**85.2%**
Kmeans+QE	97.5%	82.7%	104.5%	99.6%	102.4%	93.2%	107.5%	96.1%	**97.9%**
LDA+QE	114.4%	106.9%	112.5%	110.4%	115.8%	106.3%	120.6%	109.0%	**112.0%**

Table 5.3 User Segmentation Results in Real Estate Domain

Real Estate										
ads	1	2	3	4	5	6	7	8	9	**Avg.**
5 segments										
Kmeans	43.7%	42.8%	48.8%	44.9%	47.1%	42.1%	45.0%	47.3%	42.4%	**44.9%**
Kmeans+QE	51.8%	48.4%	54.3%	52.6%	52.9%	50.8%	55.3%	55.8%	53.1%	**52.8%**
LDA+QE	63.5%	59.8%	69.7%	65.2%	68.8%	70.4%	67.0%	71.6%	67.8%	**67.1%**
10 segments										
Kmeans	50.3%	47.9%	57.4%	53.2%	56.9%	54.6%	52.5%	58.0%	55.6%	**54.0%**
Kmeans+QE	58.8%	54.4%	66.2%	60.6%	69.1%	65.9%	69.2%	70.5%	63.7%	**64.3%**
LDA+QE	71.3%	64.6%	81.0%	74.7%	84.5%	88.1%	78.1%	89.8%	85.2%	**79.7%**
20 segments										
Kmeans	66.0%	70.6%	77.9%	71.8%	73.3%	72.8%	69.5%	79.7%	75.5%	**73.0%**
Kmeans+QE	75.1%	78.0%	85.6%	79.9%	82.0%	88.4%	84.2%	90.8%	87.5%	**83.5%**
LDA+QE	88.2%	87.3%	95.4%	86.6%	93.6%	103.2%	98.2%	108.5%	104.0%	**96.1%**
40 segments										
Kmeans	80.9%	79.1%	86.1%	87.4%	83.5%	95.6%	84.8%	86.5%	91.2%	**86.1%**
Kmeans+QE	90.5%	87.0%	94.0%	103.5%	96.7%	112.2%	91.7%	108.3%	104.6%	**98.7%**
LDA+QE	108.0%	101.4%	116.6%	113.0%	104.0%	124.5%	105.5%	126.2%	117.7%	**113.0%**

Table 5.4 User Segmentation Results in Automobile Domain

Automobile						
ads	1	2	3	4	5	**Avg.**
5 segments						
Kmeans	49.6%	52.5%	47.1%	49.5%	58.4%	**51.4%**
Kmeans+QE	60.2%	59.7%	57.9%	58.0%	68.3%	**60.8%**
LDA+QE	74.5%	68.0%	66.2%	66.8%	73.0%	**69.7%**
10 segments						
Kmeans	58.1%	61.6%	58.0%	58.4%	64.6%	**60.1%**
Kmeans+QE	69.3%	67.2%	70.8%	67.4%	75.5%	**70.0%**
LDA+QE	78.0%	81.2%	78.2%	78.5%	83.3%	**79.8%**
20 segments						
Kmeans	75.5%	78.7%	81.1%	75.0%	77.5%	**77.6%**
Kmeans+QE	86.2%	89.5%	90.5%	87.9%	86.0%	**88.0%**
LDA+QE	98.8%	95.4%	104.4%	96.6%	98.7%	**98.8%**
40 segments						
Kmeans	100.3%	96.0%	106.2%	97.1%	104.9%	**100.9%**
Kmeans+QE	113.5%	109.5%	114.0%	105.2%	113.5%	**111.1%**
LDA+QE	128.8%	125.4%	130.5%	120.8%	123.6%	**125.8%**

Table 5.5 User Segmentation Results in Diet Domain

Diet								
ads	1	2	3	4	5	6	7	**Avg.**
5 segments								
Kmeans	55.3%	46.9%	51.0%	53.5%	48.2%	53.8%	49.0%	**51.1%**
Kmeans+QE	62.9%	57.4%	61.6%	59.9%	60.4%	64.5%	57.4%	**60.6%**
LDA+QE	71.7%	69.5%	66.3%	70.7%	73.1%	72.2%	68.8%	**70.3%**
10 segments								
Kmeans	61.1%	55.0%	59.8%	62.3%	56.6%	62.0%	58.8%	**59.3%**
Kmeans+QE	72.0%	68.8%	70.5%	74.5%	67.0%	71.8%	70.6%	**70.7%**
LDA+QE	83.3%	85.1%	76.0%	80.2%	82.5%	79.3%	82.7%	**81.3%**
20 segments								
Kmeans	82.5%	74.0%	79.0%	85.5%	80.7%	77.9%	75.5%	**79.3%**
Kmeans+QE	94.5%	92.1%	88.5%	98.4%	94.3%	90.2%	88.3%	**92.3%**
LDA+QE	107.7%	101.0%	99.3%	112.0%	105.5%	96.6%	104.4%	**103.8%**
40 segments								
Kmeans	108.3%	101.2%	93.6%	90.0%	105.5%	92.0%	93.9%	**97.8%**
Kmeans+QE	116.2%	110.8%	104.0%	111.8%	120.3%	106.5%	110.8%	**111.5%**
LDA+QE	132.2%	136.6%	126.3%	121.1%	131.9%	125.5%	130.4%	**129.1%**

Table 5.6 User Segmentation Results in Camera Domain

Camera							
ads	1	2	3	4	5	6	**Avg.**
5 segments							
Kmeans	47.0%	52.2%	46.5%	44.0%	48.3%	45.5%	**47.3%**
Kmeans+QE	58.8%	63.9%	61.0%	52.4%	58.0%	56.4%	**58.4%**
LDA+QE	67.1%	73.0%	66.3%	63.0%	71.6%	65.9%	**67.8%**
10 segments							
Kmeans	59.0%	54.4%	62.9%	57.0%	60.2%	55.5%	**58.2%**
Kmeans+QE	66.1%	68.3%	70.0%	69.4%	71.1%	67.3%	**68.7%**
LDA+QE	75.2%	73.0%	80.4%	74.5%	77.8%	79.9%	**76.8%**
20 segments							
Kmeans	80.4%	83.5%	81.1%	72.9%	75.5%	78.0%	**78.6%**
Kmeans+QE	93.3%	95.0%	91.4%	90.4%	87.5%	89.3%	**91.2%**
LDA+QE	107.0%	104.3%	110.6%	106.9%	98.5%	103.7%	**105.2%**
40 segments							
Kmeans	96.7%	102.3%	96.7%	93.0%	100.4%	97.7%	**97.8%**
Kmeans+QE	106.0%	110.1%	105.0%	102.7%	113.3%	107.1%	**107.4%**
LDA+QE	115.4%	124.0%	113.3%	117.7%	122.9%	118.3%	**118.6%**

In the above user segmentation experiments, the performance of the proposed LDA based user segmentation is compared with the performance of K-means based user segmentation to see whether the semantic approach improves performance of the traditional clustering algorithm. In order to examine the impact of the proposed query enhancement mechanism on user segmentation, the experiments also compare the performance of K-means based user segmentation with the performance of K-means based user segmentation with the proposed query enhancement mechanism. To investigate whether the proposed approach is domain-independent, experiments are carried out independently across six domains, and in each domain the averaged results of individual ads are taken as the final outcome.

Through user segmentation, it is clear that the behavioral targeted advertising can significantly improve the positive user rate, if the advertisements are delivered to the proper segments of users. The experimental results indicate that the proposed query enhancement mechanism can be used to improve the effectiveness of user segmentation, as the average PUR improvement rates under "K-means + QE" strategy are increased over simple K-means strategy in different number of segments across all six domains. The PUR improvement rate can be as high as 136.6% by using the proposed user's intent representation technique with query enhancement mechanism under LDA model. By further analysis, the proposed "LDA + QE" strategy significantly exceeds K-means and "K-means + QE". This fact proves that semantic approach is appropriate to be utilized in behavioral targeting and the results verify the correctness of the proposed strategy.

For user intent representation, LDA is adopted over other simple Dirichlet-multinomial clustering models. Unlike simple Dirichlet-multinomial clustering model, LDA involves three levels where the topic node is sampled repeatedly within the document, and documents can be associated with multiple topics. This is similar to the fact that a user can have multiple intents, if a user is considered as a document and his or her intent as a topic. Under the LDA model, the relationship between users and queries can be considered parallel to documents and words.

In the experiments of this study, each user is only allowed to belong to one user segment by assigning the user into the topic segment that gives the highest probability under the LDA model. Otherwise, it is unfair to compare the proposed approach with K-means because K-means, as the baseline model, permits one user to belong to only one user segment. However, the number of users in the segment can be adjusted in practice by allowing a user to fall into multiple segments. This can be done by setting up a threshold and if the probability of a user belonging to a topic segment is equal to or greater than the threshold, the user is assigned to that segment.

It is also worth pointing out that the PUR improvement increases as the number of segments increases. Yet, it is not wise to increase the number of segments to extreme; otherwise some segment may only have a few users, which is not useful for advertiser to deliver ads, even though those users are positive users and might have the purchase intent. When the segment number approaches to infinity and every user belongs to a distinct segment, the PUR of all the segmentation approach will be the same. In addition, increasing the number of topics in LDA to extreme also may cause the over-fitting problem.

From an advertiser’s perspective, even though PUR improvement increases as the number of segments increases, there should be a tradeoff between the PUR improvement and the number of segments, depending on various factors, such as the ways of pricing online advertising and the budget for the advertising campaigns. Further discussion can be found in Section 6.1.

CHAPTER 6

DISCUSSION, LIMITATIONS, AND CONTRIBUTIONS

This chapter discusses the limitations of this dissertation, outlines the contributions of this study, and summarizes the major findings.

6.1 Discussion

6.1.1 Number of Users in the Segment and Ads Pricing

As discussed earlier in Chapter 5, although PUR improvement increases as the number of segments increases, there should be a tradeoff between the PUR improvement and the number of segments, depending on various factors, such as the ways of pricing online advertising and the budget for the advertising campaigns. For instance, in the CPA (Cost-Per-Action) pricing model, where the advertiser compensates the publisher only for clicks that subsequently result in a sale or conversion against advertiser's campaign goal, the risk for the advertisers is low because they only need to pay when the ads generate their desired outcome. Therefore, in the CPA model, PUR is more important from a publisher's perspective. The publisher might want to increase the number of segments to achieve higher PUR in order to get better compensation.

On the other hand, CPM, which stands for Cost-Per-Mille, pays the publisher a certain amount of money for every 1,000 ad impressions served. In other words, publishers get paid for every impression and risk nothing on the ads performance, regardless of whether or not the ad leads to a click or other action. In the CPM model, the PUR has no influence on how the publishers get paid, so a publisher might just want to serve as many impressions as possible without considering the number of user segments. With ample budget, advertisers could choose to reduce the number of segments and increase the number of users in a segment by adjusting the threshold during the process of user segmentation to reach more potential customers. How to select the best pricing methodology and set up the appropriate budget for the advertising campaigns is very important in advertising industry, but it is out of the scope of this study.

6.1.2 User Labeling

There are accidental clicks on the URLs. A user may accidentally click on a URL that the user does not mean to. In this case, the clicked URL may have nothing to do with this user's online intent and the user can be incorrectly labeled.

In the experiment, a user was labeled as a positive user as long as the user clicks on a URL which has the Delicious tags covering all the keywords extracted from an ad title. In practice, the user labeling settings can be adjusted to increase the accuracy of user labeling. For example, instead of labeling a user as positive by a single click, a minimum threshold of clicks can be defined, because the user is more likely to have a specific intent if s/he clicks on more than one relevant URLs. However, if the threshold is too high, this approach could miss out some potential customers. Therefore, a trade-off should be taken into consideration when labeling the users.

6.1.3 Datasets

In this study, the proposed approach improved the baseline models significantly, and the experimental results are based on the AOL datasets and the Delicious dataset. Since ad clicks data is not available in the AOL datasets and the user labeling approach is based on the clicked URLs and the Delicious dataset, there is no guarantee that the same results or better can be achieved under other datasets where user's intent is indicated by ad clicks. However, the primary goal of this study is to address the major challenges with user queries in the context of behavioral targeting advertising by proposing a query enhancement mechanism, which has been proven to help increase the performance of both user classification and user segmentation. The process of query enhancement only needs user query log and does not rely on the Delicious dataset. Therefore, similar impact of the proposed query enhancement on user classification and user segmentation can be expected under other datasets as well. In other words, as long as user query log and ad clicks data are available, the proposed query enhancement is still able to help increase the performance of user classification and user segmentation over baseline models.

6.1.4 Top *k* Similar Queries

Depending on the size of the datasets and the computing resources, the number of top similar queries to be added in the query enhancement process can also vary. In this study, k is set to be 10 based on empirical results. If k is too big, not only irrelevant queries might be added to the user's original query, which undermines user intent representation, but it could also significantly increase the dimension of the feature space, which leads to higher computational cost. If k is too small, fewer queries are added to user's original query and less information can be captured about the user's intent, especially when the dataset is small and the search queries are collected over a short period of time. In practice, additional empirical effort needs to be devoted in order to achieve optimal results.

6.2 Limitations

This study has several limitations, including cold start problem, dataset limitation, scalability issues, and difficulties in predicting the ordering of user's intents by using external data source.

6.2.1 Cold Start Problem

The proposed query enchantment mechanism needs a relatively large query log to obtain better results, especially when there are not enough data about the users available at the beginning. When a small query log is used to calculate the similarities between queries, a desirable performance cannot be achieved and many queries are not even found in the query log. As a matter of fact, as a limitation, cold start is a widely known problem involved in data modeling. It is most prevalent in recommender systems.

6.2.2 AOL and Delicious Datasets Limitation

In the experimental design, if a user clicked on a URL that did not have an associated tag in the Delicious dataset, this user was excluded from the experiment. The practical implication of this is that, if majority of the URLs in a dataset have no associated Delicious tags, this dataset is not suitable for performing user classification or segmentation using the proposed approach, which labels user's intents by matching the clicked URLs with the Delicious tags. In this case, the excluded users cannot be classified or segmented. In practice, when applying the proposed approach, there needs to be attention paid to the amount of URLs that have Delicious tags. In the AOL dataset, about 67.6% of the URLs have at least one associated Delicious tag.

Tag quality and data availability are the two limitations involved in the Delicious dataset. While it is reasonable to associate user's intent with the clicked URLs, the tags associated with URLs in the Delicious dataset may not reflect the current content of the webpage, as a result of the latency between the page update and tag update. In this case, a user could be incorrectly labeled with an intent if the tags associated with the clicked URLs are out of date. Another limitation in the Delicious dataset concerns the availability of the dataset. Since essentially Delicious is a free social bookmarking web service for storing, sharing, and discovering web bookmarks, there is no guarantee that Delicious dataset will always be publicly available.

Alternatively, other than using Delicious, queries in the click graph might be used to tag web pages. With this new design, the need for external dataset no longer exists. However, the effectiveness of this approach is uncertain without further experimentation.

6.2.3 Scalability Issues

The proposed study also involves scalability issues. The query log used in this study contains 220,138 unique search queries and 233,291 unique URLs over three months. In real advertising industry, much bigger datasets are used to perform user segmentation under machine learning algorithms. As an experimental limitation, how well the proposed approach scales is not discussed in this study. Yet, the proposed LDA based user segmentation can be implemented under MapReduce framework for good scalability in industry. The users can be divided and processed among the processors in the map phase and all of the processors are given a copy of the counter, while the global parameters update happens during the reduce phase.

6.2.4 Difficulties in Predicting the Sequence of User's Intents

Another limitation of this work involves the difficulties in predicting the sequence of user's intents by using external data source. The sequence of user's intents may involve user's offline activities. For example, showing a flight advertisement after a user has already bought a flight ticket by phone may not be useful. Similarly, a user clicking on an airline website may not be interested in purchasing a flight because the user may be just trying to check in online with a ticket bought long time ago.

6.3 Contributions

The outcome of this study contributes to the field of online advertising in the following three aspects.

Firstly, this study introduces a user intent representation strategy and proposes a query enhancement mechanism by leveraging user query log. Unlike traditional user segmentation methods, which take little semantics of user behaviors into consideration, this study incorporates the query enhancement mechanism with a topic model to explore the relationships between users and their behaviors in order to segment users in a semantic manner. The proposed method can be used to improve the performance of both user classification and topic-based user segmentation in the context of online advertising, which could lead to more successful campaigns and better user satisfaction.

Secondly, the experimental results in this study confirm the effectiveness of behavioral targeting on user segmentation. One of outcomes of this study is to provide a validation of behavioral targeting for online advertising. The experimental results indicate that the PUR improvement rate can be as high as 136.6% by using the proposed user's intent representation technique with query enhancement mechanism under LDA model.

Finally, to the author's best knowledge, the proposed user labeling approach in this study is the first effort to address the problem of the lack of benchmark datasets available in the field of online advertising in academia. It provides an opportunity for scholars who do not have access to the entire user online datasets (especially ad click data) to carry out the research in similar areas. This approach does not need human effort and can be executed in a large scale.

6.4 Summary

This research aims to address the major challenges with user queries in the context of behavioral targeting advertising by proposing a user intent representation strategy and a query enhancement mechanism. This dissertation focuses on investigating the intent based user classification performance and the effectiveness of user segmentation under a topic model that helps explore semantic relation between user queries in behavioral targeting.

Three major research questions in this study are: How to represent a user's online intent? How well can users be classified based on their intents? and does the intent-based user segmentation improve the performance of behavioral targeting significantly?

The first research question, how to represent a user's online intent, is addressed in Chapter 4 where this research proposes a query enhancement mechanism by leveraging user query log. It provides more information about a user's interests and hence helps describe and distinguish a user’s intent. The second research question investigates the impact of the proposed technique on the intent-based user classification, where a user’s intent is presented by the issued queries as well as the augmented queries. In addition to classifying users, Chapter 5 addresses the third research question by examining the effectiveness of the proposed approach on user segmentation, which plays an extremely important role in nowadays behavioral targeting advertising. The experimental results demonstrated that the proposed approach could significantly improve the user classification performance. Six different domains were chosen to evaluate the proposed approach and all the six domains yielded good performance. This non-domain specific approach can be easily applied in all intent domains without any further efforts.

REFERENCES

[1] Advertisers/Agencies – DoubleClick: *http://www.google.com/doubleclick/advertisers/*. Accessed: 2013-06-20.

[2] Agichtein, E. et al. 2006. Improving web search ranking by incorporating user behavior information. *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (2006), 19–26.

[3] Agichtein, E. et al. 2006. Learning user interaction models for predicting web search result preferences. *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (2006), 3–10.

[4] Agichtein, E. and Zheng, Z. 2006. Identifying best bet web search results by mining past user behavior. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006), 902–908.

[5] Ahmed, A. et al. 2011. Scalable distributed inference of dynamic user interests for behavioral targeting. Conference on Knowledge Discovery and Data Mining (2011), 114–122.

[6] Aly, M. et al. 2012. Web-scale user modeling for targeting. *Proceedings of the 21st international conference companion on World Wide Web* (2012), 3–12.

[7] Archak, N. et al. 2010. Mining advertiser-specific user behavior using adfactors. *Proceedings of the 19th international conference on world wide web* (2010), 31–40.

[8] Baeza-Yates, R. et al. 2005. Query recommendation using query logs in search engines. *Current trends in database technology-EDBT 2004 Workshops* (2005), 395–397.

[9] Baeza-Yates, R. and Ribeiro-Neto, B. 1999. *Modern information retrieval*. New York, NY; ACM.

[10] Baeza-Yates, R. and Tiberi, A. 2007. Extracting semantic relations from query logs. *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (2007), 76–85.

[11] Beeferman, D. and Berger, A. 2000. Agglomerative clustering of a search engine query log. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (2000), 407–416.

[12] Blei, D.M. et al. 2003. Latent dirichlet allocation. *Journal of machine Learning research*. 3, (2003), 993–1022.

[13] Bouras, C. and Tsogkas, V. 2011. Clustering User Preferences Using W-kmeans. *2011 Seventh international conference on signal-image technology and internet-based systems (SITIS)* (2011), 75–82.

[14] Broder, A. et al. 2007. A semantic approach to contextual advertising. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), 559–566.

[15] Bucklin, R.E. and Sismeiro, C. 2003. A model of web site browsing behavior estimated on clickstream data. *Journal of Marketing Research*. (2003), 249–267.

[16] BURST MEDIA: *http://burstmedia.com/*. Accessed: 2013-06-20.

[17] Chen, Q. et al. 2007. Improving query spelling correction using web search results. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (2007), 181–189.

[18] Chen, T. et al. 2010. Transfer learning for behavioral targeting. *Proceedings of the 19th international conference on world wide web* (2010), 1077–1078.

[19] Chen, Y. et al. 2009. Large-scale behavioral targeting. *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (2009), 209–218.

[20] Chowdhury, G. 2010. *Introduction to modern information retrieval*. Facet publishing.

[21] Chuang, S.L. and Chien, L.F. 2003. Enriching web taxonomies through subject categorization of query terms from search engine logs. *Decision Support Systems*. 35, 1 (2003), 113–127.

[22] Claypool, M. et al. 2001. Inferring user interest. *Internet Computing, IEEE*. 5, 6 (2001), 32–39.

[23] Craswell, N. and Szummer, M. 2007. Random walks on the click graph. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), 239–246.

[24] Cucerzan, S. and Brill, E. Spelling correction as an iterative process that exploits the collective knowledge of web users. *Proceedings of EMNLP* (2004), 293–300.

[25] Cui, H. et al. Query expansion by mining user logs. *Knowledge and Data Engineering, IEEE Transactions on*. 15, 4 (2003), 829–839.

[26] Cui, J. et al. Real time google and live image search re-ranking. *Proceedings of the 16th ACM international conference on multimedia* (2008), 729–732.

[27] Dou, Z. et al. A large-scale evaluation and analysis of personalized search strategies. *Proceedings of the 16th international conference on world wide web* (2007), 581–590.

[28] Fox, S. et al. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*. 23, 2 (2005), 147–168.

[29] Gao, W. et al. Cross-lingual query suggestion using query logs of different languages. *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (2007), 463–470.

[30] Goecks, J. and Shavlik, J. 2000. Learning users' interests by unobtrusively observing their normal behavior. *Proceedings of the 5th international conference on Intelligent user interfaces* (2000), 129–132.

[31] Google – friend or froe? - New markets, new media and consumer insight - What we think - Home - WPP Annual Report & Accounts 2008: *http://www.wpp.com/annualreports/2008/what_we_think/insight/google.html*. Accessed: 2014-04-06.

[32] Hassan, A. et al. 2010. Beyond DCG: User behavior as a predictor of a successful search. *Proceedings of the third ACM international conference on Web search and data mining* (2010), 221–230.

[33] Huang, C.K. et al. 2003. Relevant term suggestion in interactive web search based on contextual information in query session logs. *Journal of the American Society for Information Science and Technology*. 54, 7 (2003), 638–649.

[34] Jansen, B.J. et al. 2005. A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology*. 56, 6 (2005), 559–570.

[35] Jansen, B.J. et al. 1998. Real life information retrieval: a study of user queries on the Web. *ACM SIGIR Forum* (1998), 5–17.

[36] Jansen, B.J. et al. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information processing & management*. 36, 2 (2000), 207–227.

[37] Joachims, T. et al. 2005. Accurately interpreting clickthrough data as implicit feedback. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (2005), 154–161.

[38] Kanungo, T. et al. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 24, 7 (2002), 881–892.

[39] Kim, H.R. and Chan, P.K. 2003. Learning implicit user interest hierarchy for context in personalization. *Proceedings of the 8th international conference on Intelligent user interfaces* (2003), 101–108.

[40] Kumar, R. and Tomkins, A. 2010. A characterization of online browsing behavior. *Proceedings of the 19th international conference on World wide web* (2010), 561–570.

[41] Kuo, R.J. et al. 2006. Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation. *Expert systems with applications*. 30, 2 (Feb. 2006), 313–324.

[42] Lacerda, A. et al. 2006. Learning to advertise. *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (2006), 549–556.

[43] Li, L. et al. 2007. Dynamic adaptation strategies for long-term and short-term user profile to personalize search. *Advances in Data and Web Management*. Springer. 228–240.

[44] Liu, B. et al. 2006. Measuring the meaning in time series clustering of text search queries. *Proceedings of the 15th ACM international conference on Information and knowledge management* (2006), 836–837.

[45] Liu, N. et al. 2010. Learning to rank audience for behavioral targeting. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010), 719–720.

[46] Liu, Y. et al. 2008. Identifying web spam with user behavior analysis. *Proceedings of the 4th international workshop on Adversarial information retrieval on the web* (2008), 9–16.

[47] Mardia, K.V. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika*. 57, 3 (1970), 519–530.

[48] Markatos, E.P. 2001. On caching search engine query results. *Computer Communications*. 24, 2 (2001), 137–143.

[49] Mei, Q. et al. 2008. Query suggestion using hitting time. *Proceedings of the 17th ACM conference on Information and knowledge management* (2008), 469–478.

[50] Papadimitriou, P. ALGORITHMS AND STRATEGIES FOR WEB ADVERTISING. Stanford University.

[51] Pass, G. et al. 2006. A picture of search. *Proceedings of the 1st international conference on Scalable information systems* (2006), 1.

[52] Perlich, C. et al. 2012. Bid Optimizing and Inventory Scoring in Targeted Online Advertising. (2012).

[53] Poblete, B. et al. 2008. Dr. searcher and mr. browser: a unified hyperlink-click graph. *Proceeding of the 17th ACM conference on Information and knowledge management* (2008), 1123–1132.

[54] Poblete, B. and Baeza-Yates, R. 2008. Query-sets: using implicit feedback and query patterns to organize web documents. *Proceedings of the 17th international conference on World Wide Web* (2008), 41–50.

[55] Provost, F. et al. 2009. Audience selection for on-line brand advertising: privacy-friendly social network targeting. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), 707–716.

[56] Qasim, U. et al. 2009. A partial-order based active cache for recommender systems. *Proceedings of the third ACM conference on Recommender systems* (2009), 209–212.

[57] Raeder, T. et al. Design Principles of Massive, Robust Prediction Systems.

[58] Ratnaparkhi, A. 1992. Finding predictive search queries for behavioral targeting. *Training*. 10, 27,920,032,253 (1992), 27–920.

[59] Ribeiro-Neto, B. et al. 2005. Impedance coupling in content-targeted advertising. *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (2005), 496–503.

[60] Sağlam, B. et al. 2006. A mixed-integer programming approach to the clustering problem with an application in customer segmentation. *European Journal of operational research*. 173, 3 (Sep. 2006), 866–879.

[61] Salton, G. et al. 1975. A vector space model for automatic indexing. *Communications of the ACM*. 18, 11 (1975), 613–620.

[62] Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*. 24, 5 (1988), 513–523.

[63] Silverstein, C. et al. 1999. Analysis of a very large web search engine query log. *ACm SIGIR Forum* (1999), 6–12.

[64] Smart Ads - Yahoo!: *http://advertising.yahoo.com/article/smart-ads.html*. Accessed: 2013-06-20.

[65] Spink, A. et al. 2000. Searching the web: The public and their queries. *Journal of the American society for information science and technology*. 52, 3 (2000), 226–234.

[66] Spink, A. et al. 2002. US versus European Web searching trends. *ACM SIGIR Forum* (2002), 32–38.

[67] Teevan, J. et al. 2005. Personalizing search via automated analysis of interests and activities. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (2005), 449–456.

[68] Teevan, J. et al. 2008. To personalize or not to personalize: modeling queries with variation in user intent. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (2008), 163–170.

[69] The Data Driven Web: Targeting, Optimization and the Evolution of Online Display Advertising: *http://winterberrygroup.com/ourinsights/wp*. Accessed: 2014-04-06.

[70] Trajkova, J. and Gauch, S. 2003. *Improving ontology-based user profiles*. University of Kansas, Electrical Engineering and Computer Science.

[71] Tyler, S.K. et al. 2011. Retrieval models for audience selection in display advertising. *Proceedings of the 20th ACM international conference on Information and knowledge management* (2011), 593–598.

[72] Wang, C. et al. 2002. Understanding consumers attitude toward advertising. *Eighth Americas conference on information systems* (2002), 1143–1148.

[73] Wang, X.J. et al. 2009. Argo: intelligent advertising by mining a user's interest from his photo collections. *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising* (2009), 18–26.

[74] Wen, J.R. et al. 2001. Clustering user queries of a search engine. *Proceedings of the 10th international conference on World Wide Web* (2001), 162–168.

[75] Wen, J.R. et al. 2002. Query clustering using user logs. *ACM Transactions on Information Systems*. 20, 1 (2002), 59–81.

[76] Xu, J. and Liu, H. 2010. Web user clustering analysis based on KMeans algorithm. *2010 International conference on information networking and automation (ICINA)* (2010), V2–6–V2–9.

[77] Yan, J. et al. 2009. How much can behavioral targeting help online advertising? *Proceedings of the 18th international conference on world wide web* (2009), 261–270.

[78] Zheng, K. et al. 2012. User Clustering-Based Web Service Discovery. *2012 Sixth International Conference on Internet Computing for Science and Engineering (ICICSE)* (2012), 276–279.

[79] Zhuang, Z. and Cucerzan, S. 2006. Re-ranking search results using query logs. *Proceedings of the 15th ACM international conference on Information and knowledge management* (2006), 860–861.