To What Degree Can Log Data Profile a Web Searcher?

Bernard J. Jansen[1], Mimi, Zhang[1], Danielle Booth[1], Daehee Park[1], Ying Zhang[3], Ashish Kathuria[2], and Pat Bonner[1]
College of Information Sciences and Technology [1]
Department of Electrical Engineering [2]
The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering [3]
The Pennsylvania State University
jjansen@acm.org, mzhang@ist.psu.edu, nephari@gmail.com, daehee@gmail.com, reneyingying@gmail.com, ashish14@gmail.com, thepatbon@gmail.com

Abstract
In this paper, we report ongoing efforts in a large scale research project to develop methods for profiling individual Web search engine users by leveraging data recorded in the transaction logs of search engines. Our research aim is to investigate how completely one can profile a Web searcher using log data. Taking a broad brush approach, we present an array of profiling attributes to illustrate the spectrum of user characteristics possible from log data. Specifically, we present ongoing research for determining a user's location, geographical interest, topic of interest, level of engagement, the degree of commercial intent, whether the user plans to make a purchase, and whether the user will click a link. We present the state of our ongoing research in user profiling along with that of other researchers. Our findings show that one can develop a fairly robust profile of a Web searcher using log data. We also discuss issues of determining the specific identity of the user. We conclude with a discussion of the implications for the areas of system development, online advertising, privacy, and policies concerning the use of such profiling.

Introduction
What can a query tell us about the person who creates it? Envision that the image in Figure 1 is a search box containing a query generated by a user and is about to be submitted to a major Web search engine (pick any one of the major engines).

Figure 1. Search Box with Search Button and Query

What can this query tell us about the background of this person? ASIS&T is generally used as the acronym for the American Society of Information Science and Technology, a professional organization of information scientists. Therefore, there is a good probability that this user is an academic, a researcher, a librarian, or a student in one of these disciplines. Leveraging demographic information for this organization (Vaughan & Hahn, 2005), we can discern that there is a 57 percent probability that this searcher is female (or conversely a 43 percent probability that the searcher is male); a 66.2 percent chance this user works in the library and information science field; a 55.6 percent probability this user has master's degree and 32.3 percent probability this user has a doctorate. There is a 53 percent likelihood that this person works in an educational organization.

With the background information of this person derived from the query, we can then use the Internet Protocol (IP) address to locate the geographical area, at least to the city, of the computer from which this query originates by using an IP placement application (e.g., http://www.hostip.info/). Search engine transaction logs usually record the IP address with each query submission. If this IP address belongs to a known public location, we could possibly derive more information regarding this user's identity. For example, if the IP address is from a city library that does not have computers available to the public, there is a reasonable chance that this user is a librarian at that institution.

What else can the query tell us? This user's search for information related to "Annual Meeting" indicates interest in the organization's conference. This organization's 2009 meeting is (or was depending on when you read this) scheduled in Vancouver, BC, Canada in November. So, we could infer that this person is searching for the conference's schedule (if the query is submitted prior to the meeting) or perhaps looking for presentations or papers from the meeting (if the query is submitted after the conference). This is simple to determine because transaction logging applications can easily record the date of the query submission. We could carry this example even further, trying to establish more aspects of the user and determining each characteristic of the user with a greater degree of accuracy. As this example illustrates, it is theoretically possible to leverage a single query to develop a fairly insightful profile of a particular user with a reasonable probability of accuracy (though we might have to use non-online means or sources for final verification).

The electronic records of queries, page views, and other aspects of Web searching are trace data. Each user of a Web search engine leaves trace data (Jansen, Taksa, & Spink, 2008b) of their interactions including queries, specific Websites visited, and duration of visits, among many others interactions. Leveraging this trace data by using search log analysis can provide considerable insights that one can utilize for various purposes, including improvements in algorithm design and useful content retrieval for the searcher. This trace data can also be used for personalization efforts of either system or content for an individual user or a set of users. However, these efforts have raised privacy concerns regarding how search engine companies and others store and use this data. In response to these worries, most major companies have implemented policies regulating the specific period of time (typically a few months) that log data is stored, after which the data is aggregated to a level higher than the individual user.

Millions of people use search engines every day and leave trace data of their interactions. At the individual level, the trace data can also be used for purposes other than system improvement. The use of a large amount of search data might shed insight into major cultural or societal trends (for example, support for a political candidate) or general trends of the particular subset of Web searchers (for example, traces of flu outbreaks). However, one cannot manually process trace data for millions of users and queries, like we did for the query in Figure 1. Automated methods are needed in these cases.

As part of a stream of research, this paper examines how much we can learn about an individual Web search engine user from the data typically stored in transaction logs of these engines. Specifically, we are interested in building a profile of what a searcher is doing, why they are doing it, and what are they going to do next. We also discuss specific searcher identification. The overall aim is to see how complete of a searcher profile one can develop based solely on data that is stored in a typically search log.

Efforts in Using Log Data for User Profiling

Several researchers have utilized transaction logs for analyzing Web searching. Two of the first papers in this area examined searches on the Excite search engine (Jansen, Spink, Bateman, & Saracevic, 1998) and AltaVista (Silverstein, Henzinger, Marais, & Moricz, 1999). Since then, several studies have examined various aspects of Web searchers using transaction logs. (See (Markey, 2007a, 2007b) for a review of the Web searching research and (Penniman, 2008) for the historical roots of Web log analysis.) While researchers continue efforts to determine the actions and intentions of Web searchers using logs, many are increasingly apprehensive about privacy issues concerning the record that the Web search engines have of individuals (Zimmer, 2008).

However, there has been little systemic examination of how much of a user profile one can really develop from log data. Can one leverage search engine logs to gain important insights in individual Web searchers? If possible, how difficult is it to do so? How accurately can one profile Web search engine users based on log data? Can we leverage these profiles to improve system design? What are the risks of such profiling? These are some of the questions that motivate our research, but to address these questions, we must first discover how much we can learn about a user from log data.

Research Question

Our research question is: *How complete of a profile can one develop for a Web search engine user from transaction log data?*

To answer this question, we present methods to develop, understand, and gain insight into different characteristics of Web search engine users. For each component of the profile, we leverage the data that is typically stored in the transaction logs of Web search engines. We make the assumption that the user enters the search engine without logging in because once the user signs in via an account at a given search engine (i.e., iGoogle, myYahoo, etc.), much of the profiling becomes trivial, assuming that the registration data represents honest information. Therefore, our focus in this research is on the anonymous users of Web search engines.

Fields Typically in a Transaction Log

Each search engine records slightly different information during the episodes of interaction between the engine and a user. However, there is a de facto standard format that typically consists of some of the following fields:

- *User Identification*: the computer's IP address or a code to identify a particular computer based on the computer's IP address
- *Cookie*: an anonymous text file automatically assigned by the search engine server to identify unique users on a particular computer based on a browser instance
- *Date*: the date of the interaction as recorded by the search engine server
- *Time of Day*: usually measured in hours, minutes, and seconds as recorded by the search engine server on the date of the interaction
- *Query Terms*: the terms as entered by the user and recorded by the server
- *Vertical*: the source content collection that the user selects to search, such as *Web, Images, Audio, News*, and *Video*
- *Assistance*: usually a code denoting whether or not the query was generated via some contextual help feature of the search engine
- *Language*: the language in which the user prefers the content to be presented
- *Search Engine Results Page (SERP):* a code representing the set of both organic and sponsored links along with associate images and context displayed by the search engine in response to a user query
- *Click* – whether or not the user clicked on a sponsored or organic link
- *Rank* – the position in the results listing of the clicked link
- *Landing Page*: the Webpage pointed to by a uniform resource locator (URL) of a result on the SERP

Using these fields from a variety of search engine logs and standard transaction log techniques (Jansen, 2006), we explore the ability to develop profiles for Web search engine users, with results presented in the following section.

Results

In developing a user profile, we examine various aspects of a search engine user. We focus on deriving (a) what the user is doing, (b) what the user is interested in, and (c) what the user intends to do. Specific aspects that we aim to surmise are: (1) location, (2) geographical interest, (3) topical interest, (4) topical complexity, (5) content desires, (6) commercial intent, (7) purchase intent, (8) gender, (9) potential to click on a link, and (10) user identification. We acknowledge that there may be other aspects for investigation; however, we believe these are some core characteristics for demonstrating whether it is possible to adequately profile a user.

Since the research purpose of this manuscript is to explore profile development, we do not explain each method in great detail. Each algorithmic approach could be a research paper in itself. Where appropriate, we reference the existing literature in the field.

Where is the searcher? (user location): Determining the user's location (technically the location of the computer from which the query was submitted) can be done using a fairly straightforward IP address look-up script. Using IP addresses recorded in a search engine log, we developed a script that calls the *geo-location API* at www.hostip.info and updates the search log database to include city, state, and country. A sample of the enriched log with location data is shown in Table 1.

Table 1. Geographical Location of User Based on IP Address From a Search Log

IP Address	City	State	Country
x72.x78.68.x2			Germany
x2.x04.14.x09	Rockport	Massachusetts	USA
x61.x.2.x60	Helena	Montana	USA
x8.x11.x73.143	Leeds		England

(Note: For privacy reasons, the exact IP address has been masked.)

Given the use of ISPs and remote logon techniques, we acknowledge that this approach is not 100 percent accurate, and a good area of future research would be IP address lookup accuracy. Given the increased interest in geographical search restrictions in the keyword advertising market, the search engines have responded with geographical targeting techniques that appear to address this need. The IP look-up services state that geographical location using IP addresses is approximately 90 percent accurate. Our assumption, then, is that the use of the IP address of the originating computer is a fairly precise method. Additionally, with mobile search increasing, geo-positioning devices can be used to pinpoint user locations with even greater degrees of accuracy. As this technology becomes more ubiquitous in devices, location targeting will become even more accurate.

Table 2: Geo-Targeting of User Query (actual queries from a search log)

Query	Geo-Target
peru illinois movie theater showings	peru illinois
farms for sale in garrard county kentucky	garrard county kentucky
the pottery shop in commerce georgia	commerce georgia
manchester new hampshire outlet shopping	manchester new hampshire

Where is the user going? (local search or geo-targeting): Related to the user's location, there is attention to where the user is interested in going or, more specifically, what is the geographical target of interest of the user's query. This is a fairly straightforward characteristic to determine, namely leveraging a database of place and location names (along with common misspellings) to identify geographic terms in the query. We used a database of locations to determine the location references included in queries, which we label as geo-targets in Table 2.

What is the topic of interest? (topic identification of the query): The accurate topical classification of user queries is one of the most active research areas in Web search with the potential for increasing effectiveness and efficiency of Web search systems. Classifying Web queries by topic is a challenging problem, as Web queries are generally short (Wang, Berry, & Yang, 2003); because of this, they are ambiguous with few attributes for classification. One of the most successful approaches was provided by Beitzel, Jensen, Lewis, Chowdhury, and Frieder (2007) who used two label datasets from AOL to classify queries in a search stream, reporting that a combined methods approach led to an approximately 70 percent accuracy for topical classification.

Using one of the same AOL datasets (a manual topic classification log of approximately 22,000 queries), we replicated the approach proposed by Beitzel and fellow researchers (2007). We also found that a combined methods approach provided accurate classification when there was a match. However, this approach was unable to classify a large number of queries. Therefore, we automatically categorized

query search terms into semantic categories using the Open Calais tool (www.opencalais.com/calaisAPI). We present a snippet of results in Table 3.

Table 3. Topic Identification of User Queries (actual queries from a search log)

Query	Topic
angie everhart	person
1967 Ford	company auto
discount track lighting	shopping
side effects of arnica	health

Meta-tagging tools, like Open Calais, are improving and, when combined with approaches like those outlined in earlier studies (Beitzel et al., 2007; Özmutlu, Çavdur, & Özmutlu, 2008; Shen et al., 2006), one can use such tools for making measureable improvements in topical query classification. Accuracy rates for classification of Web queries are reportedly between 40 and 70 percent, which is high given the unbelievable range, structure, semantics, and syntax of Web queries. We have found that the major problems in topical classification of queries are misspellings, company names, people names, slang, and 'new' words – especially those involving entertainment, such as bands and games.

What are the patterns of user interaction? Along with general topic classification, a related area of investigation is the pattern of user interaction with the search engine. A user's interactions while searching can inform one of many things, such as the user's commitment to locate the desired information or the complexity of the searching topic. We have investigated query reformulation patterns (Jansen, Zhang M., & Spink, 2007b) and session identification (Jansen, Spink, Blakely, & Koshman, 2007a) using a six state classification, presented in Table 4.

Table 4. Query Reformulation Sessions

State	Description
New	first query from a user or the query is on a new topic from this searcher
Assistance	query generated by the searcher's use of some system assistance feature
Content Change	the user executed a query on another content collection
Generalization	the current query is on the same topic as the searcher's previous query, but the searcher is now seeking more general information
Reformulation	the current query is on the same topic as the searcher's previous query, and both queries contain common terms
Specialization	the current query is on the same topic as the searcher's previous query, but the searcher is now seeking more specific information

The results of this research showed that *Reformulation* and *Assistance* states account for approximately 45 percent of all query modifications. In session identification, we found that contextual and temporal approaches were the most effective in identifying session boundaries for Web searchers (Jansen et al., 2007a). We also noted reoccurring patterns for users' reformulating queries.

Expanding this research (Jansen, Booth, & Spink, 2009), we leveraged these patterns of interaction to predict a user's next action using an n-gram approach. N-grams are a probabilistic modeling technique used for predicting the next item in a sequence and are (n-1) order Markov models, where n is the gram (i.e., pattern) from the complete sequence. An n-gram model predicts state x_i using states x_{i-1}, x_{i-2}, x_{i-3}, …x_{i-n}. The probabilistic model is presented as: $P(x_i \mid x_{i-1}, x_{i-2}, x_{i-3}, \ldots x_{i-n})$, with the assumption that the next state only depends on the last $n - 1$ states. We employed n-grams to describe the probability of users transitioning from one query reformulation state to another in order to predict their next state. We developed first, second, third, and fourth order models and evaluated each model for accuracy of prediction, coverage of the dataset, and complexity of the pattern set. Results showed that a second or

third order model provided the highest predictability for the lowest cost. Table 5 shows the first order model probabilities.

Table 5. Patterns of User Interaction with a Web Search Engine

	New	Content Change	Reformulation	Generalization	Generalization w/ reformulation	Specialization	Specialization w/ reformulation	Assistance	Total
New	0%	13%	21%	7%	7%	**22%**	9%	**21%**	100%
Content Change	0%	0%	17%	11%	8%	16%	7%	**41%**	100%
Reformulation	0%	11%	0%	14%	**18%**	19%	**23%**	15%	100%
Generalization	0%	10%	18%	0%	5%	**37%**	12%	18%	100%
Generalization w/ reformulation	0%	6%	**32%**	6%	0%	18%	27%	11%	100%
Specialization	0%	9%	**32%**	16%	**22%**	0%	12%	9%	100%
Specialization w/ reformulation	0%	6%	**28%**	14%	**36%**	8%	0%	7%	100%
Assistance	0%	**58%**	12%	7%	11%	5%	8%	0%	100%

(Note: High probabilities for each state are bolded.)

What type of content does the user desire? (user intent): Broder (2002) proposed a query taxonomy, later enhanced by (Rose & Levinson, 2004), of three board classifications (*informational*, *navigational*, and *transactional*) of user intent, defined as the type of content desired. Jansen, Booth, and Spink (2008a) presented a comprehensive classification of user intent consisting of three hierarchical levels. Additionally, the researchers developed an application using a decision tree approach to classify the user intent of Web queries automatically, with an accuracy of 74 percent. Others have also explored automatically determining the user intent of Web queries, notably Yates, Benavides, and Gonźalez (2006).

Leveraging this prior work, we utilized a k-mean clustering approach to group similar queries into more granular sets of user intent. A k-means clustering algorithm attempts to identify relatively homogeneous groups of cases based on selected characteristics. We clustered approximately four million Web queries and associated fields and analyzed the resulting clusters, comparing them against a manually labeled set of several thousand queries. Research results show that the accuracy of the k-means clustering approach is 94 percent in classifying like *informational*, *navigational*, and *transactional* queries into unique clusters.

However, we noted that there appeared to be topical influences. Therefore, taking a 20,000 plus AOL Web query data set sectioned by topic (Beitzel et al., 2007), we manually classified each query using a three-level hierarchy of user intent similar to that presented in (Jansen et al., 2008a). We noted differences in user intent across topics. Results show that user intent (*informational*, *navigational*, and *transactional*) varied by topic, as shown in Table 6.

As we can see from Table 6, there are certainly topical categories that trend toward *informational*, *navigational*, or *transactional* (15 to 24 percent depending on the category). We have bolded the topical categories that have *navigational* or *transactional* percentages above approximately 25 percent, along with notably high *informational* percentages. Most Web queries are *informational* in nature (Broder, 2002; Jansen et al., 2008a; Rose & Levinson, 2004).

Table 6. User Intent Classification Percentages by Topic

	Info.	Nav.	Trans.			Info.	Nav.	Trans.
Auto	**81.2%**	15.8%	3.0%		Organization	25.0%	**72.1%**	2.9%
Business	47.4%	**51.9%**	0.7%		Other	**55.6%**	26.1%	18.3%
Computing	**60.5%**	11.8%	**27.7%**		Places	**62.9%**	31.1%	6.0%
Entertainment	**79.7%**	6.1%	14.2%		Porn	11.6%	26.1%	**62.3%**
Games	65.5%	9.7%	**24.8%**		Research	**51.3%**	32.9%	15.8%
Health	**89.6%**	8.9%	1.4%		Shopping	33.4%	31.7%	**35.0%**
Holiday	48.3%	**50.8%**	0.9%		Sports	**51.7%**	30.2%	18.1%
Home	**60.9%**	21.0%	18.1%		Travel	47.4%	41.9%	10.7%
News	**50.9%**	35.1%	14.0%		URL	0.1%	**99.2%**	0.7%

Does the query have commercial intent? (pertaining to commerce or business): The major search engines generate most of their income through advertising, commonly displaying ads on the same page as organic search results. It would be advantageous for both the search engine and users if ads were displayed only when the user might participate or expressed interest in a commercial transaction. It is therefore profitable to have a mechanism for categorizing queries as either having or not having commercial intent (CI). Prior work in the area has used a 'circular' algorithm to determine CI in which researchers use data generated by the search engine (i.e., retrieved pages) to determine if the users' queries have CI (c.f., Dai et al., 2006).

Instead of using external data, we employed an algorithm that functions independently from the search engine results, using only queries to determine if there is CI. There are many keywords that a user may include in a query that can reliably indicate CI, including words such as "buy," "price," "purchase," and so on. It is easy to label queries containing such keywords as having CI, but it becomes more complicated when queries do not contain such keywords. However, by utilizing these "obvious" CI keywords and a large set of manually labeled queries, it is possible to develop a method for extrapolating the CI of associated keywords.

The approach that we took was as follows: (a) a search log is parsed for queries that include keywords that obviously denote CI (c.f., "buy," "price," etc.); (b) for every query selected, the remaining key words are extracted; (c) using one of these associated keywords, the search log is parsed again for all queries containing it; (d) ratio of obvious CI queries to all queries containing that keyword is calculated; (e) if that keyword has a ratio of at least 1/10, it is considered to be a keyword that, in the future, will denote commercial intent for a query.

In order to evaluate the effectiveness of this approach, MSN Adlab has an online tool that measures the CI of queries (http://adlab.msn.com/OCI/OCI.aspx). It uses the algorithm reported in (Dai et al., 2006). The researchers performed an analysis of their algorithm with human evaluators, and they found it to be more than 90 percent accurate when compared to the human evaluations. Therefore, we used this tool to evaluate the effectiveness of our algorithm. The results of our approach concurred with MSN Adlab approximately 88 percent of the time, with an average probability of CI of approximately 74 percent. Naturally, as the manually labeled search log size increases, the accuracy of our CI algorithm will increase.

Is the user getting ready to make a purchase? (query indicates buying intent): We are pursuing a methodology to estimate a user's situation in the buying funnel accurately based on a search query. Hotchkiss (2004) defines the buying funnel as a staged process. Each phase represents a searcher's psychological step to approach a transaction, with a prime focus on the stages of *awareness*, *research*, *decision*, and *purchase* purposes.

Applying this concept of the buying funnel to sponsored search marketing would allow advertisers to purchase potentially highly targeted clicks and views. Currently, pay-per-click technology allows advertisers to buy traffic based on broad, phrase, and exact keyword matching. The return on investment

of an advertising campaign would markedly improve if it could be targeted specifically to phases of the buying funnel. While benefiting the advertisers with lower costs of conversion, search engines would deliver more accurate results to the user.

The number of words in a query plays an active role in determining the user's situation in the buying funnel. The use of long tail keywords (Wolfram, 1999) suggests that the user is further along in the buying funnel because he/she has accumulated enough knowledge to refine the search query. A long tail keyword is highly-targeted and niche. It is crafted by including many specific details about the target product/service. The use of a long tail keyword suggests that the user has already passed through the *awareness* and *research* phases of the buying funnel.

Thus, when a user's search is highly specific, he/she is more likely to make a purchasing decision compared to conducting broad, preliminary research on a product. Hotchkiss' study (2004) indicates that 70 percent of the participants start their searches with broad keywords because it is easier and allows them to whittle down the results.

Consider the following simulation of building up a search query in a given user session to a long-tail keyword:

- **Awareness:** "*digital cameras*" (recognition of the need for a product)
- **Research:** "*digital camera reviews*" (gathering information about products)
- **Decision:** "*nikon coolpix 5400 51 mp vs sony 5 mp cybershot dsc t9*" (comparison shopping for evaluation of alternatives)
- **Purchase:** "*amazon.com nikon coolpix 5400 51 mp digital camera*" (decision for what and where to purchase)
- **Search query:** "*amazon.com nikon coolpix 5400 51 mp digital camera*"
- **Analysis of Purchase Query**
  - **Grouping:** [store] [brand] [model name] [specification] [product type]
  - **Word count:** 8 and **Group count:** 5

We are currently conducting empirical analysis on the data log of a sponsored search campaign. Initial results show that branded and product queries (i.e., queries containing a store or product name) have markedly higher click through rates relative to unbranded queries.

What is the user's gender?: Microsoft has a tool (http://adlab.microsoft.com/Demographics-Prediction/) that can predict a user's gender (along with age and other demographic information) based on a user's online behavior, including what queries they search online and which Websites they visit. The distribution results in response to a seed query show the age breakdown of MSN Search users, based on a one-month MSN Search log and regardless of search query used. Predicted distribution depicts the breakdown by age of MSN Search users for a single search query, based on the MSN Adlab's predictive model.

This particular tool runs into the same issue faced by many recommender systems and personalization efforts, namely that one cannot tell if the users' actions are for them or for someone else. As an example, Amazon makes books recommendations based on what a person has ordered in the past. However, ordering a book as gift messes up the recommendations, unless one notes the purchase as a gift. Therefore, we believe the more logical approach for guessing the gender of the user is to assign each query a gender bias – the probability that this query has a masculine or feminine disposition. This is a research avenue that would require a large dataset to be effective but would be a worthwhile endeavor.

Will the user click on a result? (predicting user actions): From a search engine perspective, one of the most important user actions is the click on a link presented in the SERP. We are investigating the factors that affect the click through of Web searchers. The goal is to determine more efficient methods to optimize click through rate. In one study, we used a neural network to detect the significant influence of searching characteristics on future user click through (Zhang Y., Jansen, & Spink, 2009a). Neural

networks are powerful data modeling tools and are able to capture and represent complex relationships between input and output, so they have broad application. Our results showed that high occurrences of query reformulation, lengthy searching duration, longer query length, and the higher ranking of prior clicked links correlated positively with future click through.

In a follow-on study investigating click through, we used time series analysis to evaluate predictive scenarios of click through (Zhang Y., Jansen, & Spink, 2009b). Time series analysis is a method often used to understand the underlying characteristics of temporal data in order to make forecasts. In this study, we used time series analysis to investigate users' actions using log data. We used a one-step prediction time series analysis method along with a transfer function. The period rarely affected *navigational* and *transactional* queries, while rates for *transactional* queries varied during different periods. For this data set, the average length of a searcher session was approximately 2.9 interactions, and this average was consistent across time periods. Our results showed that searchers who submitted the shortest queries (as measured by number of terms) clicked on the highest ranked results in the next period. These studies focused on an aggregate set of users (all the user searching episodes contained in the transaction log). We are continuing this line of research investigating individual user sessions.

Who is this particular user? (user identification): The aim of this research is to determine what the user is doing, why they are doing it, and what the user will do next, not to indentify who exactly the user is. However, it is an interesting exercise to get an idea of the difficulty of this task. We explored this effort anecdotally from several angles.

First, we explored ways to leverage a single query to identify a user (i.e., being able to refer to a particular person by name). As our leading example in Figure 1 shows, there are certain queries where we can narrow a user down to a particular location, along with other attributes. However, narrowing down to a single individual with a high degree of probability from a single query is extremely difficult except in a few cases (and even then one cannot be sure).

Therefore, we then examined an individual session. Given that most sessions are extremely short, this avenue also proved difficult. So, we examined all sessions in a given day. Again, identifying a given user proved nearly impossible from solely a day's queries.

Next, we looked at temporal spans. Unfortunately, there are limited log datasets available to academia that span more than one day, and even fewer where user identification carries across days. The most well known is the AOL date set which encompasses approximately 20 million Web queries from 650,000 AOL users over a three month period. However, even in this dataset, determining the identity of a specific user is extremely difficult, although much easier than with a log of only a single day. In fact, to our knowledge, of the 650,000 users, only a handful of users have been specifically identified leveraging the search data. Even here, locating an individual from the search log data must be cross referenced with other public records. Furthermore, even then it is impossible to be 100 percent accurate in linking search log data (without a logon) to an actual person without using other public records or communication means.

Generally, though, one can say that the longer the period of data collection, the higher the probability of correctly identifying a user. How long? Given that most search engines have implemented policies that anonymize data after about a year to eighteen months, this is probably the period where it becomes notably easier to indentify individual users.

Discussion and Implications

Based on the research results so far, we developed a user profiling framework (Figure 2). We classify user aspects into two levels: *internal* and *external*. The *internal* aspects are shown in the pink circle (Figure 2) and refer to attributes of the users themselves. We can develop the demographic profile including location, gender, identification, and such from the log files. The *external* aspects are shown in the blue circle (Figure 2). They relate to the behavior or interest of the users, including the users' system usage pattern. The *internal* and *external* aspects interact with each other. We can infer the users' *external* aspects from the users' *internal* aspects. The users' *external* aspects reflect the users' *internal* aspects.

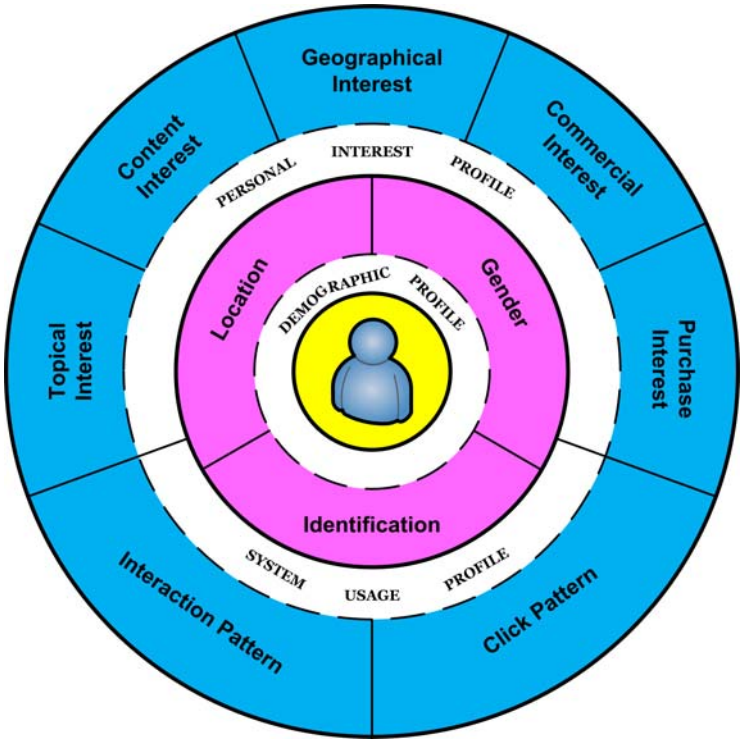

Figure 2. User Profiling Framework

There are several important implications for leveraging search engine transaction logs to profile users for a variety of areas, including personalization, online advertising, ecommerce, and privacy, among many others. More effective leveraging of log data for user profiling can yield more effectiveness and efficiency for Web searchers, search engines, and merchants.

Concerning the use of trace data, the fields within a typical search log are sparse, with rarely more than a dozen or so fields. With such a meager range of data, these logs must typically be enriched in some way to gather meaningful insights. Examples of enrichment include using the IP address to gather location information, combining the query with Webpages or an external database to identify the topic or location, and applying algorithms to session interactions for additional classifications. One can also take algorithmic approaches to enrich the data set by combining fields to generate new characteristics such as examining the changes of a sequence of queries within a session.

Our research findings indicate that one can develop a fairly in-depth profile of what a searcher is doing, gain insights into what is motivating the user, and build predictive models about the user's future actions. Specifically in terms of profiling, we can determine fairly easily where a user is and what the geographical area of interest is as expressed by the query. We are getting better at determining the user's topical interests and intent. We are beginning to be able to identify the commercial intent of the query and are working on predictive aspects, such as determining whether or not a user is a potential buyer or if a user will click on a result link. Although sparse, it appears that one can embellish and enrich logs enough to profile Web searchers; this has significant implications for system development, government regulation, corporate policies, social privacy concerns, and ecommerce.

Our future research will focus on the efficiency of these and other profiling efforts as well as improving the effectiveness of these approaches. These are challenging tasks, but ones for which it appears that steady progress is being made by a variety of researchers. Secondly, there has to be concerted efforts on the evaluation of these approaches (how accurate can we predict a user's location?; how sure can we be of the gender bias of a query?, etc.). These will be especially challenging given the scale of the data that is often needed (millions to billions of interactions and hundred of thousand to millions of users)

and hurdles in providing testing and evaluation datasets. For example, how does one test whether or not a query has a female or male orientation? How does one evaluate the accuracy of using IP to locate a given user? Responding to these questions will require large and specialized test beds of data. Finally, the privacy and anonymity issues must be resolved. How can one leverage the trace data that is left behind by the millions of online interactions to improve services, while balancing the protection of the identities of the individual users?

Conclusion

In this manuscript, we have presented several techniques for leveraging log data to build a profile of an individual Web search engine user. In time, these approaches will only improve in effectiveness, efficiency, and accuracy, as well as range of characteristics (i.e., richer profiles).

There are benefits that system designers can leverage from this profiling, including contextual help, targeted advertising, and more relevant results. These benefits have the potential to help all involved – the search engines, users, and especially the advertisers who provide the revenue to support the massive infrastructure needed for large scale Web search engines. However, there are potential risks with the ability to profile individuals that confront users and search engine companies. For the user, how much control or ownership does the individual have over the data generated, if any? For the search engine company, what are the risks of storing this profile data? How long should it be stored? Additionally, predictive models of user behavior can offer significant improvements in system performance. However, there is a short distance from prediction to influence and, perhaps, undue influence. What is an acceptable boundary? As the technology continues to improve, these policy questions have to be addressed. We believe that the research presented in this manuscript will assist in generating the much needed discussion.

References

Beitzel, S.M., Jensen, E.C., Lewis, D.D., Chowdhury, A., & Frieder, O. (2007). Automatic classification of Web queries using very large unlabeled query logs ACM Transactions on Information Systems, 25(2), Article No. 9.

Broder, A. (2002). A Taxonomy of Web Search. SIGIR Forum, 36(2), 3-10.

Dai, H.K., Nie, Z., Wang, L., Zhao, L., Wen, J.-R., & Li, Y. (2006, 23–26 May). In Detecting Online Commercial Intention (OCI) (pp. 829-837). Paper presented at the World Wide Web Conference (WWW2006), Edinburgh, Scotland.

Hotchkiss, G. (2004). Inside the Mind of the Searcher. Retrieved 15 March, 2005, from http://www.enquiro.com/research.asp

Jansen, B.J. (2006). Search log analysis: What is it; what's been done; how to do it. Library and Information Science Research, 28(3), 407-432.

Jansen, B.J., Booth, D., & Spink, A. (2008a). Determining the informational, navigational, and transactional intent of Web queries. Information Processing & Management, 44(3), 1251-1266.

Jansen, B.J., Booth, D.L., & Spink, A. (2009). Patterns of query reformulation during Web searching. Journal of the American Society for Information Science and Technology, 60(7), 1358-1371.

Jansen, B.J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real Life Information Retrieval: A Study of User Queries on the Web. SIGIR Forum, 32(1), 5-17.

Jansen, B.J., Spink, A., Blakely, C., & Koshman, S. (2007a). Defining a session on Web search engines. Journal of the American Society for Information Science and Technology, 58(6), 862-871.

Jansen, B.J., Taksa, I., & Spink, A. (2008b). Research and Methodological Foundations of Transaction Log Analysis. In B.J. Jansen, A. Spink & I. Taksa (Eds.), Handbook of Research on Web Log Analysis (pp. 1-17). Hershey, PA.: IGI.

Jansen, B.J., Zhang M., & Spink, A. (2007b). Patterns and transitions of query reformulation during Web searching. International Journal of Web Information Systems, 3(4), 328-340.

Markey, K. (2007a). Twenty-five years of end-user searching, part 1: Research findings Journal of the American Society for Information Science and Technology, 58(8), 1071-1081.

Markey, K. (2007b). Twenty-five years of end-user searching, part 2: Future research directions. Journal of the American Society for Information Science and Technology, 58(8), 1123-1130.

Özmutlu, H.C., Çavdur, F., & Özmutlu, S. (2008). Cross-Validation of Neural Network Applications for Automatic New Topic Identification. Journal of the American Society for Information Science and Technology, 59(3), 339–362.

Penniman, W.D. (2008). Historic perspective of log analysis. In B.J. Jansen, A. Spink & I. Taksa (Eds.), Handbook of Research on Web Log Analysis (pp. 18-38). Hershey, Pennsylvania, USA: IGI.

Rose, D.E., & Levinson, D. (2004, 17–22 May). In S. Feldman, M. Uretsky, M. Najork & C. Wills (Eds.), Understanding User Goals in Web Search (pp. 13-19). Paper presented at the World Wide Web Conference (WWW 2004), New York, NY, USA.

Shen, D., Pan, R., Sun, J.-T., Pan, J.J., Wu, K., Yin, J., et al. (2006). Query enrichment for web-query classification Transactions on Information Systems, 24(3), 320 - 352.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a Very Large Web Search Engine Query Log. SIGIR Forum, 33(1), 6-12.

Vaughan, L., & Hahn, T.B. (2005). Profile, needs, and expectations of information professionals: What we learned from the 2003 ASIST membership survey. Journal of the American Society for Information Science and Technology, 56(1), 95-105.

Wang, P., Berry, M., & Yang, Y. (2003). Mining Longitudinal Web Queries: Trends and Patterns. Journal of the American Society for Information Science and Technology, 54(8), 743-758.

Wolfram, D. (1999). Term Co-occurrence in Internet Search Engine Queries: An Analysis of the Excite Data Set. Canadian Journal of Information and Library Science, 24(2/3), 12-33.

Yates, R.B., Benavides, L.C., & Gonźalez, C. (2006). The Intention Behind Web Queries. In F. Crestani, P. Ferragina & M. Sanderson (Eds.), Lecture Notes in Computer Science: String Processing and Information Retrieval (SPIRE 2006) (Vol. 4209/2006, pp. 98-109). Glasgow, Scotland: Springer Berlin / Heidelberg.

Zhang Y., Jansen, B.J., & Spink, A. (2009a). Identification of Factors Predicting ClickThrough in Web Searching Using Neural Network Analysis. Journal of the American Society for Information Science and Technology, 60(3), 557-570.

Zhang Y., Jansen, B.J., & Spink, A. (2009b). Time Series Analysis of a Web Search Engine Transaction Log. Information Processing & Management, 45(2), 230-245.

Zimmer, M. (2008). The panoptic gaze of Web search engines. In A. Spink & M. Zimmer (Eds.), Web Searching: Interdisciplinary Perspectives (pp. 77-99). Dordrecht, The Netherlands: Springer.