Analysis of Geographic Queries in a Search Engine Log

Qingqing Gan
Polytechnic University
Brooklyn, NY 11201
qq_gan@cis.poly.edu

Josh Attenberg
Polytechnic University
Brooklyn, NY 11201
josh@cis.poly.edu

Alexander Markowetz
University of Science & Technology
Hong Kong, S.A.R
alexmar@cs.ust.hk

Torsten Suel
Polytechnic University
Brooklyn, NY 11201
suel@poly.edu

ABSTRACT

Geography is becoming increasingly important in web search. Search engines can often return better results to users by analyzing features such as user location or geographic terms in web pages and user queries. This is also of great commercial value as it enables location specific advertising and improved search for local businesses. As a result, major search companies have invested significant resources into geographic search technologies, also often called local search.

This paper studies *geographic search queries*, i.e., text queries such as "hotel new york" that employ geographical terms in an attempt to restrict results to a particular region or location. Our main motivation is to identify opportunities for improving geographical search and related technologies, and we perform an analysis of 36 million queries of the recently released AOL query trace. First, we identify typical properties of geographic search (geo) queries based on a manual examination of several thousand queries. Based on these observations, we build a classifier that separates the trace into geo and non-geo queries. We then investigate the properties of geo queries in more detail, and relate them to web sites and users associated with such queries. We also propose a new taxonomy for geographic search queries.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Content Analysis and Indexing—*Indexing methods*; H.3.3 [**Information Systems**]: Information Search and Retrieval—*Search process*

General Terms

Measurement, Human Factors

Keywords

web search, geographic search, local search, query log mining

1. INTRODUCTION

Over the last decade, search engines have become the primary means of locating information for many people. For this reason, researchers have started investigating available search query logs, in order to better understand what people are searching for, how they are searching, and how this process can be improved. A number of recent studies [30, 11, 29, 4, 25], have looked at query logs from various perspectives, including Computer Science, Library and Information Science, and Social Sciences. Our perspective is primarily from Computer Science, where researchers mine query logs and click-through behavior to optimize system performance or provide more accurate results.

While the Web has removed many geographical limitations in media, communications, and e-commerce, many geographical aspects of the physical world are nonetheless reflected in the Web's content and structure. As a result, geography often provides a useful and intuitive constraint for Web search. This paper investigates *geographic search queries*, i.e., keyword queries that employ geographical terms in order to obtain search results related to a particular geographical location or area. Typical examples are "hotels new york", "building codes in seattle", "virgina historical sites", or "unemployment long island". Such queries frequently contain names of cities, states, or countries – often abbreviated, e.g., "CA", "NYC", or "SF". Alternately, they may contain streets names, informal synonyms (e.g., "big apple"), or refer to landmarks and neighborhoods (e.g., "SoHo" in New York). In some cases, users include zip codes or phone numbers.

Because of geography's important role in search requests, and the significant commercial potential of such queries (e.g., for hotels, real estate, or local businesses), search companies have recently invested significant resources into geographic (geo) search technologies (also called *local search*), i.e., methods aimed at giving improved answers to geographic search requests. Approaches range from integration of business directories (yellow pages) to answer fairly simple but lucrative queries (e.g., for hotels, shops, and restaurants), to a more detailed analysis of queries, page content, and site and link structure in order to facilitate more general queries. Geo search applications can use a standard keyword interface and extract geographic terms from queries, employ graphic interfaces such as interactive maps, or use the current location of a mobile user. In general, geo search engines combine knowledge regarding how people use geographic terms in queries, how such terms are used in pages, and how sites are organized and linked with respect to geography. They commonly also use external data sources, in particular gazetteers listing the names and locations of states, cities, or businesses. Geo search technology has recently been studied by a number of researchers, mainly focusing on the extraction of geographic information from page content and structure [22, 24, 2, 14, 20, 9], indexing and query processing [38, 7, 35, 21], and the automatic identification of geographic queries [10, 36].

Our main objective is to identify opportunities for improving geographic search engines. However, our observations should be of more general interest. We investigate real world queries of a large query log from a standard (non-geographic) search engine, namely 36 million queries from AOL. We study how people write geographic queries and how these should be processed by search engines. Our paper builds on work in [28] and [37] that analyzed geographic queries.

We are interested in what types of geographic queries (*informational*, *navigational*, *transactional*) users issue, what types of geographic terms they employ, and what they are looking for. We also study what sites users visited as a result of a geo query, how different geographic terms were used by the same user, and what non-geographic terms are associated with geographic terms.

The remainder of this paper is organized as follows. Section 2 provides a basic background and an overview of related work. Section 3 introduces the data set. Section 4 shows how geographic features can be used to classify queries into geo and non-geo queries. The next three sections investigate geographic properties of queries, users, and sites, respectively. The main focus lies on our taxonomy of geographic queries. Finally, we conclude in Section 8.

2. BACKGROUND AND RELATED WORK

There is significant literature on search engine logs, including studies of general search logs [30, 11, 29, 4, 25], and various papers focusing on special types of users and collections, e.g., multi-media search [12], intranet search [31], blog search [23], or search in other languages [19]. In particular, Kamvar and Baluja [15] studied the characteristics of mobile queries submitted to Google's search services for

LocWeb 2008, April 22, 2008, Beijing, China.
ACM 978-1-60558-160-6/08/04 . . . $5.00.

PDA and cellular phones. We note that while mobile and geographic (local) search are often thought of as being closely related technologies, they are certainly not the same. It can be argued that many mobile queries are in fact geographic in nature, and that for certain types of queries it may make sense to return results related to the current position of the user. Kamvar and Baluja [15] investigate various features of mobile queries, including query length and topics (but not geography), focusing on the user interface aspects of small screens and limited input capabilities. In contrast, we focus on queries issued by desktop and laptop users to a general search engine.

Search queries can be categorized according to several dimensions. Broder [5] first proposed three distinct categories of queries: (i) navigational, (ii) informational, and (iii) transactional. Of particular importance to our approach is the work by Rose and Levinson [27] who expanded Broder's work into a more detailed taxonomy, also consisting of three categories but differentiated further into ten search goals:
A. Navigational: The user has a distinct Web site or page in mind that he knows or assumes to exist. Navigational queries often contain fragments of URLs or names of organizations. The user commonly clicks on only one result, taking him directly to the desired page.
B. Informational: These queries are similar to those traditionally studied in IR, i.e., the user wants information about a certain topic, either broad (e.g., "history us") or narrow (e.g., "special nutrition for wound care"). Here, users often follow several of the resulting links.

- **Closed:** queries seek a single, closed answer.
- **Open:** queries seek open-ended answers or answers of unlimited depth.
- **Undirected:** queries target anything or everything about a particular topic.
- **Advice:** queries seek advice or instructions to complete a task.
- **Locate:** queries attempt to detect where a real world good or service can be obtained.
- **List:** queries search for lists of good pages on a topic, e.g., a Yahoo or ODP directory.

C. Resource: These queries target resources, not web documents.

- **Download:** queries target a resource which must be downloaded to be useful.
- **Entertainment:** queries search for pages which when viewed may provide entertainment.
- **Interact:** queries look for pages which require further interaction, for instance map or weather services.
- **Obtain:** queries seek documents which are useful on or off the computer, such as tax forms or government documents.

In [27, 17], researchers studied users' navigational behavior (in particular, click-through behavior), since a user's goal cannot always be inferred by just looking at a query. They find that over 60% of queries were informational, and a large fraction of the other nearly 40% seemed to seek a commercial transactions, rather than request product information. Distributions of search taxonomies are subject to changes in search technology and user behavior - somebody who a few years ago may have looked for the Web site of a company (navigational) for product information may now be willing and able to order the item directly from the site (locate). In this paper, we use the classification in [27], utilizing click-through data to identify the information need reflected by a query.

We are also interested in examining geo-queries categorized according to the topic-based taxonomy of Spink et al. [30]. Here queries are assigned to one of eleven categories according to what topic most closely matches their intent. These categories are, in decreasing order according to the fraction of all queries in a general query log in with each category:

1. Entertainment
2. Pornography
3. Business, travel, employment
4. Computers
5. Science and medicine
6. People, places, things, odds and ends
7. Society and religion
8. Education, humanitarian interests
9. The arts
10. Government
11. Unknown and other

Even before the web, researchers studied how to exploit geographic information embedded in documents for better text search and analysis; see [16] for a good overview of early work. Initial work on geographic search on the web appears in [6, 9, 22], and in recent years a significant amount of research has addressed this new challenge. Geographic queries were previously studied by Sanderson and Kohler [28] and by Zhang et al. [37]. The former provides a brief study of some of the properties of geographic queries, in particular frequency, topics, length, and spatial relationships. The latter study focuses on the issue of *geo modification* in consecutive queries, i.e., how users modify their choice of geographic terms when the previous query did not provide satisfactory results.

Assume a user looking for a nearby yoga class might look for "yoga park slope" (a neighborhood in Brooklyn). When this search returns poor results, she might try "yoga new york" and be swamped by many irrelevant results. Finally, "yoga brooklyn" satisfies her information need. For a single search task, she had to re-write the same query several times. One goal of geographic search technology is to avoid successive query modification through proper analysis of queries and collections. The automatic rewriting method in [37] provides one such approach (also related to the query expansion technique for geographic search in [8]). Our work here expands on [28] by providing a more in-depth analysis of the properties of geo queries. This paper also investigates the relationship between geography, page topic, and users, and is to our knowledge the first work in this direction.

Closely related to the analysis of geographic queries is the automatic detection of geo queries [10, 36, 37], and in general of geographic terms in text data [18, 2]. In particular, automatic detection is highly useful for measuring the statistical properties of geo queries in large logs. Such detection can be based either on individual queries, or can include past queries, past click-through behavior, or results returned by the engine. There have been many proposals on how to use knowledge mined from search query logs, such as click-through information, repeated identical or related queries by the same or different users, or co-occurrences of terms in queries, to deliver improved search results to users [3, 13, 33, 26, 34, 32, 1]. The study of geographic queries by the same or different users, or of click-through behavior on such queries, is also of interest in this context.

3. IDENTIFYING GEO QUERIES

This section lays the foundation for our study. We describe the underlying data, discuss basic geographic properties, and introduce a taxonomy of geographic queries. The relative frequency of geographical queries as well as their subtypes is evaluated on a manually geo-coded query set. Finally, we propose two classifiers to classify the entire query trace. These classifiers are highly accurate, as evaluated on the manually geo-coded samples. We then use these classifier to aid in our subsequent statistical evaluation of the entire trace.

3.1 Underlying Data

We study a trace of the AOL search engine, recording queries of roughly $650,000$ users over three months in early 2006. The trace consists of about 36 million lines of data, each containing five fields:

AnonID: an anonymous user-ID

Query: the actual query terms

QueryTime: when the query was issued

Item-Rank: the rank of the clicked result

ClickURL: the host-level result the user clicked on (if any)

In case the user clicked on multiple results to a single query, these events are recorded in the form of extra lines. For an in-depth description of the data, see [25].

Although real-life queries are often malformed and misspelled, the user's intent is usually quite clear. For example, "www.footballcamps-atlanta.google" is clearly malformed, but it is apparent what the user was looking for. Similarly, "noweign cruise lines" is misspelled, but has a clear intention.[1] When classifying queries by hand, we label according to the *intent of the user*, not according to any mistakes, when possible. This is done using the methodology of Rose and Levinson [27], utilizing click-through data for clarification when queries alone are insufficient for determining intent. The rationale is that query classification per se should be interested in a user's intent, not her way of expressing this intent. Also, most advanced search engines realize users' mistakes and propose corrected versions of the query. Due to limited resources, we do not perform spell-checking when performing automatic classification on the entire query trace.

To detect geographic terms in queries, we use the US Census Bureau's gazetteer, which contains names and locations of counties, their subdivisions (district, borough, barrio), places (town, city, village, etc.), and ZIP Codes for all 50 states.

3.2 Hand-Tagging Geo Queries

We begin by extracting an initial sample of 6000 random queries from the data set. After discarding all queries consisting exclusively of URLs and some badly misspelled or malformed queries, 4495 queries remain. These are examined manually, and assigned one of four labels, according to their geographic intent and their use of common geographic terms. Thus, for each query we decide if it has a geographic intent, and if it contains the name of a city, county, or state according to the gazetteer. Note that other geographic terms also appear frequently, such as street names or names of landmarks or places of interest (e.g., "statue of liberty" or "empire state building"). The four categories are: (i) Geographic queries that contain a city, country or state name as a geographic term. (ii) Geographic queries that do not contain such terms. (iii) Non-geographic queries seemingly containing a geographic term, e.g., "whitney houston". This category includes many entity names, such as "Kentucky Fried Chicken", "New York Times" or "First Niagara Bank". (iv) Non-geographic queries without geographic terms. The numerical results of this classification are presented in Table 3.1.

Types of Queries	Num. of Queries
Geo with Geo terms	12.01%
Geo without Geo terms	0.93%
Non-Geo with Geo terms	24.44%
Non-Geo without Geo terms	62.62%

Table 3.1: Geo vs. non-geo queries.

Table 3.1 may give the impression that only 13% of the queries pursue a geographically focused task, but the real percentage should be somewhat higher. The AOL query trace is based on a standard search engine, with no explicit geo capabilities. Many users with a geographical search task in mind may only use such search engines to find a Web site that will allow them to restrain the geographic focus of their query in a second step. In our random sample, for example, we find about twenty five requests for mapping services (e.g., `mapquest.com`). These users are most likely pursuing a geographic search task. Similarly, users searching for "craigslist" will have to specify a metropolitan area of interest as soon as they access `www.craigslist.org`. Many queries for retail chains, e.g., Radio Shack, Nordstrom, or Target, are likely geographic in nature as users often seek to locate a store using the company's web site. We did not evaluate the number of such queries, as it would be difficult to guess if a user is interested in finding a local store or making an online purchase. In any case, 13% is probably an underestimate of the frequency of geographic search tasks.

In our experiments, we only consider geographic entities within the United States; thus, queries that refer to international locations or to the US as a whole are ignored. The rationale behind this decision is that any automatic query classifier needs to incorporate some understanding of the language issues, ambiguities and difficulties associated with the geographic query terms from a particular region. Such information is usually compiled for a single region or country at a time; for this reason, local search engines are commonly launched on a per-country basis. Since we are best able to manage these issues within the geographic and linguistic confines of the United States, we chose to focus our work on queries focused there.

After manual classification, we discovered 582 queries with geographic intent out of 4495 queries in the sample. We then looked at the query length (number of terms) of these queries; the results are shown in Table 3.2. Note that the columns titled "Non-Geo" and "Geo" indicate the distribution of geographic and non-geographic queries in terms of query length; thus, 14.48% of all geographic queries have 2 terms. The column titled "Geo of all" depicts the percentage of all queries with a given number of terms which have a geographic intent; thus, 18.78% of all queries with three terms are geographic queries.

Num. Query Terms	Non-Geo	Geo	Geo of all
1	25.54%	1.03%	0.52%
2	33.95%	14.48%	5.22%
3	19.54%	35.04%	18.78%
4	10.47%	26.21%	24.56%
5	5.19%	17.93%	30.86%
> 5	5.31%	5.31%	11.19%

Table 3.2: Number of terms in geo and non-geo queries.

This table confirms what was noticed in [28] and [37]: geo queries tend to have more terms than non-geo queries, and conversely the likelihood that a query is a geo query increases with the number of terms. However, one has to be very careful in interpreting these results. It should be expected that many classes of specialized queries, say geographic queries, people queries, or product queries, have more terms than average. If we imagine that each term in a query is chosen from some distribution, then the likelihood that a geo term (or people term, or product term) is present, and/or that a geographic or people or product intent is present, increases with the number of terms. Note also that classes such as geographic and health queries are not mutually exclusive, and that a longer query may be more likely to be in several classes. Thus, it is not impossible that most or even all such specialized classes of queries of interest have an above average number of terms. Finally, a very short query is less likely to be recognized as a geographic query even if the underling intent is geographic (e.g., as query "walmart" that tries to find the closest store on the company website). Related to this, [37] reports that 12.7% of query rewrites add a geo-specific term; thus, the original query probably had geographic intent. A good geographic search engine might use the user's location and previous geographic queries to return likely results of interest without a rewrite by the user.

3.3 Taxonomies for Geo-Search Queries

Following Rose and Levinson [27], we classified about 500 geo queries and about 500 non-geo queries from our sample into eleven distinct categories according to the apparent goal of the user, as inferred from the query itself and the associated click-through data. results, given in Figure 3.1, show significant differences between geo and non-geo queries. Geo queries are more frequently aimed at locating goods and services; non-geo queries are more likely aimed at entertainment, downloads, or lists of pages with further information.

[1] Norwegian Cruise Line is a large cruise operator.

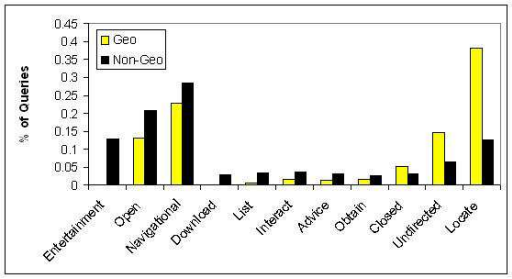

Figure 3.1: Distribution of geo and non-geo queries according to the taxonomy of Rose-Levinson. Note that the bars in each color sum up to a total of 1.0.

Navigational queries of a geographic nature often point to regional sections of nation-wide corporation or service. We observe two typical cases: (1) *Site-Wide.* The geographic term is used to distinguish the desired Web site from other similar Web sites. For example, "DMV ny" targets `www.nydmv.state.ny.us`, while "DMV ca" targets `www.dmv.ca.gov`. Similarly, many different cities have bars or restaurants with identical names (e.g., Joe's Pizza) that are not affiliated in any way. (2) *Site-Internal.* Here the non-geographic terms already determines the desired Web site, and the geographic term targets a particular page or item inside this site (e.g., "craigslist boston").

The difference between "locate" queries in the context of geo vs. non-geo queries is pronounced. Most geo-query "locate" searches consist of the name of a particular store or a search for a service in an area, e.g., "florists phoenix" or "crobar nyc", while a typical non-geographic counterpart may contain the name of a good to buy online, such as "ellsworth kelly prints". Also, while there are many navigational queries among the geo queries, a majority of these are searches for local or state government agencies. Many "open" geographic queries are searches for local media, news, or people. Such topical differences are not conveyed by the taxonomy of Rose and Levinson.

Next, we turn to the topical classification scheme used by Spink et al. [30], which also consists of eleven categories, listed in Section 2. Labelling the same set if geo and non-geo queries, we get the results shown in Figure 3.2.

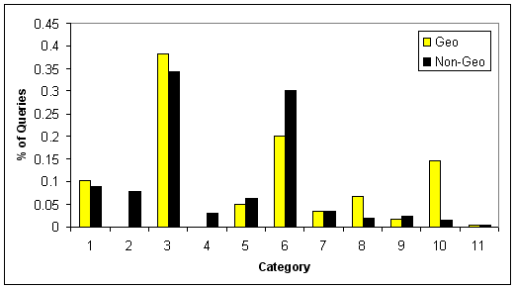

Figure 3.2: Distribution of geo and non-geo queries according to the topicality classification of Spink et al.

We again see some obvious difference in several categories. Category two and four are exclusively non-geographic: there were no queries asking for local pornography or local information about computers. Category 6 is dominated by geographic queries. There are frequent requests for local news and events, local government services, weather. On the other hand, many non-geo queries were about celebrities and national news. In category 5 (science and medicine), there were many queries for local medical services, but unsurprisingly very little local physics or other sciences. Category 8 shows that much information about schools and education is sought at the local level, for all levels of education. The same applies to category 10; there are frequent searches for branches of local government and official forms and information (e.g., about zoning laws and taxes). But as the taxonomy of Rose and Levinson, Spink's taxonomy also does not capture some important difference between geo and non-geo queries users, which are often within a category.

To address this, we propose a new query taxonomy for geographic queries that combines aspects of topicality and desired type of interaction. We came up with 23 categories as follows:

1. **Tourism/Travel:** hotels, maps, flights, transport, local attractions
2. **Government:** searches for government entities, info, and laws
3. **Real Estate:** houses, apartments, and commercial real estate
4. **Education:** requests for educational or school related information
5. **Business:** non-online business related searches, except when in another category
6. **Night Life:** including restaurants, entertainment, and casinos
7. **Undirected:** broad informational requests for a topic
8. **Medical:** hospitals, doctors, and general health and medical information
9. **Media:** news, radio, papers, magazines, and other media
10. **Employment:** searches seeking employment opportunities
11. **Automotive:** requests for automotive information and searches for automotive businesses
12. **Civic:** searches seeking civic, religious, and non-profit organizations
13. **Closed:** seeking an answer to a specific question
14. **Obtain:** seeking a specific document or resource that is useful on or off the computer
15. **List:** searches for a site which can provide further information. Seeking a hub rather than an authority
16. **Advice:** requests for advice or directions to complete a task
17. **Downloads:** requesting software or files to be downloaded to a user's computer
18. **Interactive:** requesting pages which require further interaction in order to be useful
19. **People:** seeking individual people
20. **Open:** open ended questions or requests for information
21. **e-Business:** attempts to find a online retailer of a product or service
22. **Entertainment:** queries seeking to be entertained by the contents of a page. Including pornography and pictures
23. **Navigational:** requests clearly looking for a specific web site

We note here that this taxonomy is specifically designed to allow better understanding of geo queries, and in particular the first twelve classes captures common types of queries that we found in our trace. The distribution of geo and non-geo queries in this finer-grained, hybrid taxonomy is shown in Figure 3.3. As we see, geo queries focus on the first 13 categories, and are less frequent in the others (with the exception of category 20). While there are significant number of commercial geo queries for hotels, restaurants, cafes, real estate, and local businesses, one interesting observation was the large number of local queries about government, civil organizations, education, and media that may not be well served by the current generation of geo search technology that is heavily focused on the former cases.

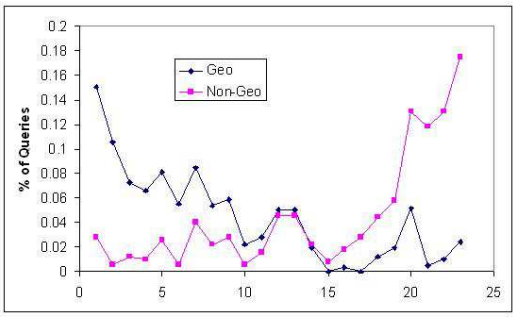

Figure 3.3: Distribution of geo and non-geo queries according to our hybrid classification

4. QUERY CLASSIFICATION

The sample data set used in the previous section is of insufficient size for many tasks. For example, making statements about frequently appearing terms in geographic queries requires more information than our sample set allows. Categorizing the entire AOL trace by hand is, however, not feasible. Instead, we use the manually labeled sample to bootstrap two classifiers. The first differentiates geographical queries from those without geographic intent, while the second classifies geographic queries roughly according to informational versus navigational queries. As our experiments show, both classifiers are sufficiently accurate, and thus they are subsequently used to classify all 36 million queries.

The biggest challenge in geographic query classification comes from ambiguous geographic terms. It is obvious to readers of the yellow press that queries such as "Paris Hilton" do not commonly refer to hotels in the capital of France. Similarly, "Cadillac" commonly targets automobiles, not a city in Michigan. In order to disambiguate queries containing these terms, we have to inspect their other terms. Abbreviations of state names such as "CA" often indicate a geographic meaning. This rule of thumb however does not apply to certain states like "MD", "LA", or "OR". Many such cases are hard to classify, even for humans.

4.1 Geo Non-Geo Classification

This first classifier detects geographical queries in two stages. First, a simple filter removes all queries without any geographic terms. In other words, queries with no locality terms are classified as non-geo queries; as shown earlier this affects about 1% of all queries that are geographic but have no city, country, or state name. After applying this filter, we are left with queries falling into categories "geo with geo terms" and "non-geo with geo terms". These are then classified according to the following features:

Property & Tourism Does the query contain terms about properties or hotels?[2]

State Does the query contain a state name, or its abbreviation?

State-Pos The position of the state name from the end of the query; e.g., 0 if it is the rightmost term in the query. We notice that when a state name is included in a query, the state name often appears at the end of the query.

Ambiguous State-Abbreviation Does one of the following state abbreviations appear as the only locality information in the query: "OH", "OR", "MD", "AS"(American Samoa) ? These abbreviations are often used in a non-geographic sense.

City Does the query contain a city name?

County Does the query contain a county name?

County-follow If the answer is true for the previous questions, is the county name followed by word "county", "village", "co", "borough" etc? People searching for a county or city often append such indicative terms.

State-follow If a city or county term appears in a query, does the term occur next or prior to a state name? The city or county must be inside that particular state.

Place-Size If a city or county term is found, how large is its population? If it is a very popular city or county in US, it is most likely that the query searches for that city/county. On the other hand, a small city is the target of few search queries.

Geo-Web-Freq If a city, county or state name is present, what is the frequency of this term in general Web documents?

Geo-Query-Freq If a city, county or state name is present, what is the frequency of this term in general search queries?

[2]In particular: *apartment*, *balcony*, *bath*, *bathroom*, *bed and breakfast*, *bedroom*, *building*, *condo*, *condominium*, *duplex*, *estate*, *flats*, *garage*, *home*, *hotel*, *house*, *inn*, *kitchen*, *lawn*, *lease*, *lodge*, *lodging*, *map*, *motel*, *property*, *real estate*, *realestate*(sic.), *rental*, *renting*, *sublet*, *view*, *villa*, *waterfront*, and their plural forms, e.g., *apartments*.

Class	Precision	Recall	F-Measure
Non-Geo	0.911	0.899	0.905
Geo	0.903	0.915	0.909

Table 4.1: Accuracy of the Geo-NonGeo Classifier

Place-Person If a city, county or state name is present, could this term also be a person's first or last name? First and last names were obtained from the US Census Bureau.

Name-Place If a city, county or state name is present, does this term appear prior to a last name or after a first name?

As shown in Table 3.1, there are actually more non-geographic queries containing geo terms than there are geographic queries. In order to produce a good classifier, we used training data consisting of 50% geographic queries *with* geographic terms and 50% geographic queries *without* geographic terms. In total, the training set consisted of around $1,200$ queries.

Utilizing the popular machine learning software, Weka[3], we evaluate our decision-tree based classifier using ten-fold cross validation. About 90.69% of all queries were correctly classified; see Table 4.1 for the results. Note that this accuracy is measured on the already filtered data, i.e., the classifier differentiates between geo and non-geo queries that both contain geographic terms. If used on all queries, its accuracy would be higher. Our classifier compares favorably to that of [10] in terms of accuracy. After applying the classifier to the entire AOL log, around 13.39% of all queries are identified as having geographic intent.

4.2 Informational vs. Navigational Queries

It is not feasible to automatically classify geographic queries according to any of the fine-grained taxonomies illustrated in Section 3.3. From a user's point of view there is a clear distinction between navigational or resource queries. A user wants to either find a website, or find a resource, e.g., buy something. However, the resulting queries often look similar, and can even be identical. Assume a user investigating the latest sportswear. She might search for "adidas", a navigational query to learn about available models. But a user intending to buy shoes online might also enter "adidas" and then proceed to the online store. This query now targets a resource; the query is the same, but the user's intention is very different. Thus, it is clearly not possible to infer user intent from queries alone, even for a human classifier. However, we can resort to a cruder taxonomy which is still meaningful and that allows for automatic classification. We hence limit ourselves to two simple categories, navigational and informational. The first contains all queries that are navigational according to the definition of Rose and Levinson, or that request a download. The second category contains all other queries.

This classifier differs from the previous in that it does not look at the query terms, but instead looks at users' click-through data. The underlying assumption is that for a navigational query, a user only clicks on a single result, as suggested in [17]. For an informational query, she may instead follow several links. This hypothesis is captured by the following two features used by our classifier:[4]

Avg. number of clicks per query This feature represents how many results a user clicks on after issuing a query. This number is averaged over all users who issued a particular query.

Click distribution This feature is based on the intuition that most clicks resulting from a navigational query focus on a few popular URLs. The click distribution of a query is defined according to the number of clicked times for each different URL associated with the same query. We look at 5 measures of distribution: average, mean, standard deviation, skew, and kurtosis.

Additionally, we investigate:

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

[4]For a detailed explanation of both features, see [17].

Class	Precision	Recall	F-Measure
Informational	0.85	0.951	0.898
Navigational	0.928	0.789	0.853

Table 4.2: Accuracy of Info-Navi Classifier

Geo-URL Does the clicked URL contain the name of a city, county or state?

The resulting classifier is reasonably accurate. Given a training set of around 400 hand labeled queries distributed evenly between informational and navigational, the classifier achieves an accuracy of 87.94%. Note that we only select queries with more than 10 clicks to evaluate our classifier. If a user issued an identical query several times and every time followed the same result, then we counted only a single click. Table 4.2 shows the accuracy numbers for this classifier.

5. GEOGRAPHIC QUERY PROPERTIES

There are important differences between geo and non-geo queries; users look for different "things" when searching locally than globally. The classifiers presented in the previous section facilitate the study of properties of geo queries on a large scale. First, we classify the entire AOL trace into geo and non-geoqueries. Then, we analyze term frequencies for both types of queries. Finally, we explore the distribution of geographic and non-geographic queries in different topical categories as well as geographic distribution.

5.1 Frequent Terms

Table 5.1 outlines the five most frequent terms for geographic and non-geographic queries, taken from the results of our automatic classifier. Note that no geo terms (city, county, or state names) or stop words are counted; this applies to all remaining sections. Unsurprisingly, the most frequent terms in non-geographic queries are unrelated to geography, while other terms are more likely to appear in geo queries than in non-geo ones.

Query Type	Top-5 terms
non-geographic	"free" "google" "new" "yahoo" "pictures"
general geographic	"hotel(s)" "sale" "real estate" "beach" "home(s)"

Table 5.1: Top-5 query terms

5.2 Frequent Terms at Varying Granularity

Do geographic queries at different granularity (e.g. county vs. city) address different information needs? This is indeed the case, as shown in Table 5.2, which outlines the most frequent terms in different granularity. (We note here that county vs. city is not just a different granularity, but also often an indication of more rural or suburban versus urban environments, complicating the picture a bit. City residents are often more likely to refer to their location by city name rather than the county the city is located in, which may have little relevance to them.)

5.3 Indicative Terms

Some terms are more likely to appear in geo queries than in non-geo queries, of a non-geographic nature, and vice versa. Table 5.3 displays the five terms that are most likely to be in a geo queries. This is computed as the number of times a term appears in geographic queries divided by the number of instances in which the term appears in the general query log. This could be used to further improve the performance of our classifier. For example, the term "estate" is much more likely to appear in a geo query. Here, we only take into account query terms which appear more than 1000 times in the whole query log, reducing noise induced by infrequent terms.

Query Granularity	Top-5 terms
city level	"hotel" "beach" "city" "news" "auto"
county level	"county" "real estate" "house" "property" "home"
state level	"jobs" "lottery" "sale" "park" "department"

Table 5.2: Top-5 query terms

Term	Likelihood to appear in a geographic query
estate	81.61%
shores	81.59%
cemeteries	81.05%
appraiser	80.98%
lodging	80.79%

Table 5.3: Terms most likely to appear in geographic queries

5.4 Geo Queries and Topical Categories

In Section 3, we showed that geo and non-geo queries focus on different search topics. To explore this notion in the larger dataset, we relate our queries to web sites covered by the *Open Directory Project* (ODP). Thus, we assume that a query falls into some category iff the clicked URL (i.e., website, since click-though data is provided on a site level only) associated with this query is covered under that category. We limit ourselves to the ODP top-level categories. For each category, Figure 5.1 shows the number of geo and non-geo queries.

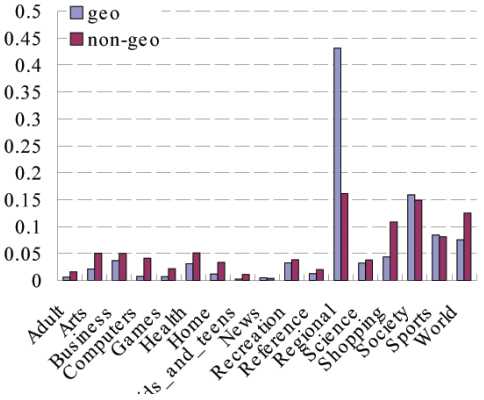

Figure 5.1: Query distribution over different topics

Note that we filter out duplicate query/click pairs from the same user. A small portion of sites are covered by more than one category. Of course, categories are not entirely exclusive. In particular, many sites (e.g., a local football club) are commonly classified by location ("regional") as well as topic ("sports"). Obviously, the "regional" category applies to a larger number of geographic queries. In order to compare geo and non-geo queries in terms of their distribution over topics, we removed the regional category and plotted the results again, shown in Figure 5.2. We can see that geographical queries clearly tend towards a few categories in ODP, such as Society and Sports. This also includes a large number of clicks on pages of religious, civic,and governmental sites.

5.5 Geo Query Distribution over US States

This section investigates how geo queries are distributed among different states in the US. A geographic query includes at least one location term, i.e., a city, county, or state name. We assign a state to

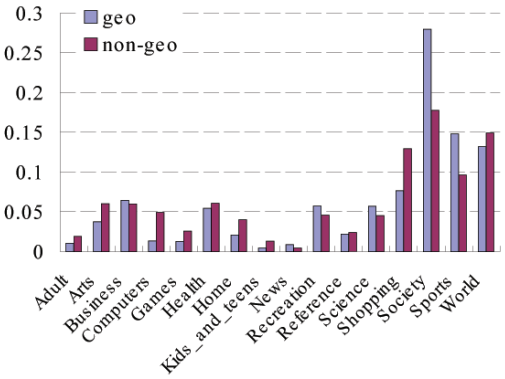

Figure 5.2: Query distribution, without "regional"

each query according to this term. In the case that only a city name is found and is associated with more than one state, we associate this query with the city having the largest population. For example, there are more than five "Brooklyn" in the US, but we assign "New York"' as the state for any such query.

In our experiments, we look at the popularity of different states in geographic queries. The five most popular states are: Florida, California, Texas, New York, and Ohio. Combined, queries about those five states count for 36.72% of all geographic queries in our data set. This is not surprising as these are also very populous states. Also, people show different interests for different states. For example, "Kids and teens" is the most popular topic in both Florida and New York, while the same topic is the least popular one in other states (possibly due to the importance of tourism for these states). Detailed results on this experiment are omitted for space reasons.

6. GEO PROPERTIES OF WEB SITES

6.1 Geo vs Non-Geo Sites

In the previous section, we investigated geo queries. In this section, we extend our study to sites that are commonly associated with such queries. In particular, we look at what sites are are mostly visited by clicking through on geo queries, and how such sites are distibuted over topics and assosiated with geo terms. Figure 6.1 divides all sites receiving more than 10 clicks into ten bins. Bins are assigned according to the fraction of these queries that were geo queries. Thus, the first column on the left represents sites visited exclusively from geo queries, while the rightmost column represents sites visited only from non-geo queries. We can see that there is a strong bimodal behavior; many sites are either mostly geo or mostly non-geo in nature when characterized by the queries used to visit them. There is also a reasonable number of sites, shown in column 2 to 4, that have mostly non-geo queries but also some geo queries; such sites may have some limited amount fo geographic information on their site such as, such as a store location or company address.

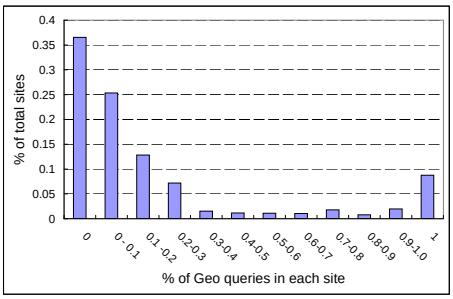

Figure 6.1: Distribution of sites according to the queries that are used to find them

Based on this, we define a geo site as a site where more than 80% of its associated queries are geo queries. Those sites where more 80% of the associated queries are non-geographic in nature, we call non-geo sites. Next, we look at the differences between geo and non-geo sites.

6.2 Geo Sites and Top-Level Domains

In Figure 6.2, we look at how geo and non-geo sites are distributed among different top-level domains. We see that .gov and .org sites are more often visited via geo queries, as such sites are more often associated with local government and civil organizations.

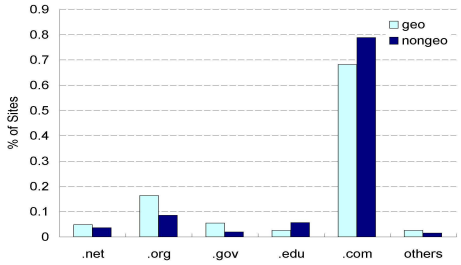

Figure 6.2: Distribution of geo/non-geo queries for different top level domains

6.3 Geo Sites and Topical Categories

Now we investigate the topical distribution of geo and non-geo sites, using again the ODP hierarchy. Confirming our previous findings, we see that geo-sites are more likely to be associated with the regional category. In fact, the vast majority of geo sites that were found in ODP were in the regional category. This indicates that our way of defining a geo site could in fact be used to identify good candidates for the regional category. More detailed results are again omitted due to space constraints.

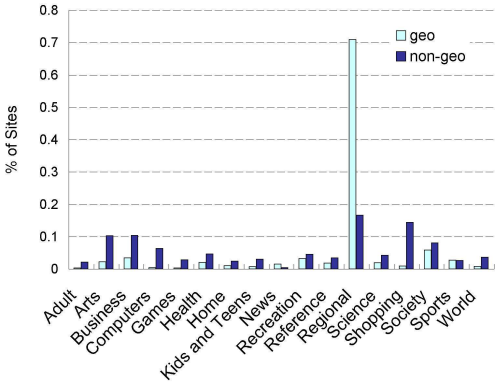

Figure 6.3: Distribution of sites in different categories

6.4 Local vs National sites

Some sites seem to appear only in results for queries regarding a particular area (say, "www.brooklynyoga.com" for Brooklyn), while other sites are associated with geographic query terms from around the country. Examples of such sites include "www.realtor.com" and "travel.yahoo.com". This tells us that some sites have a broad geographic relevence while others provide a service only to a particular area. In additional experiments, omitted for space reasons, we studied the properties of such local versus nationwide sites. In summary, as shown in this section, geo queries can be used to mine interesting facts about the sites that are visited via those queries.

7. GEOGRAPHIC USER PROPERTIES

This section studies user behavior in connection with geographic search tasks. Due to space constraints, we can only summarize some of our observations. We focused on users with at least 200 geographic queries, and then manually examined the users' searching behavior, looking at the following questions:

Do users repeatedly conduct searches on the same geographic area? The answer is yes. Indeed, one could probably easily infer the hometowns of many of these users from the geo terms in their queries, as

users exhibit a tendency to conduct searches for local services. The non-geo terms associated with a user's geo-terms also reveal much of a user's relationship with an area. Thus, if terms such as "school", "yoga" or "real estate" tend to appear with geo terms, we have reason to believe that the user lives nearby. On the other hand, terms like "hotel" or "vacation" might indicate the user lives somewhere else.

Do people in a single session of querying reformulate their queries, trying different names for the same area? That is, how frequent is *geo modification*, as discussed in Section 2? Indeed, not too often. There are different ways to define search sessions. Manually checking the search history, we can identify instances when a person changes the topic of a search, and thus define a user search session as a series of queries on a similar topic over a continuous block of time. This period can vary from several minutes to several days, as long as a user stays focused on a topic. When people search for local information or services, they are often fairly confident about the appropriate geo terms. Thus, when users modify their queries, they more often modify the non-geo terms. Users occasionally change the geographic constraint present in the query while maintaining the non-geographic portion of the information request. We found that in most of these cases, the user is querying about a location away from their likely home. The geographic terms are sometimes adjusted to point to different parts of a city, since in some cases a tourist or traveler may be flexible about where to go for a temporary stay. We note that the state names show very strong consistency across a user's search session.

How are user queries clustered locally? For a particular user, one can derive their main geographical focus as the state or area addressed by most of the geo queries of this user. This is likely the place of residence of the user. Similarly, one can define secondary and further clusters, potentially recent travel destinations of this user.

8. CONCLUSION

In this paper, we investigated geographic properties of search queries. Though, our main objective was to derive new techniques for geographic search engines, we believe our observations are of general interest. Our main contributions here are a more detailed study of geographic search queries, a new taxonomy for such queries, and experiments that relate such queries to the sites that are visited and the users that pose them. We believe that with improved understanding of users' query goals and websites' informational content, search engines can take measures to improve response relevance. Due to space constraints, we had to omit many details of our results.

There are many intriguing open questions left by our work. In particular, we would like to explore additional properties of the web sites associated with geographic queries, and of geographic search sessions, and study how user behavior on geo queries (particularly clickthrough data) can be harvested for better geographic search.

9. REFERENCES

[1] E. Agichtein and Z. Zheng. Identifying "best bet" Web search results by mining past user behavior. In *Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 902–908, 2006.

[2] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proc. of the 27th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 273–280, 2004.

[3] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proc. of the Sixth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 407–416, 2000.

[4] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized Web query log. In *Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 321–328, 2004.

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

[6] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting Geographical Location Information of Web Pages. In *2nd Int. Workshop on the Web and Databases (WebDB)*, pages 91–96, 1999.

[7] Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic web search engines. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 277–288, 2006.

[8] T. M. Delboni, K. A. V. Borges, and A. H. F. Laender. Geographic Web search based on positioning expressions. In *Proc. of the Workshop on Geographic Information Retrieval*, pages 61–64, 2005.

[9] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proc. of the 26th Int. Conf. on Very Large Data Bases (VLDB)*, pages 545–556, 2000.

[10] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing Web queries according to geographical locality. In *Proc. of the 12th Int. Conf. on Information and Knowledge Management*, pages 325–333, 2003.

[11] B. J. Jansen and U. Pooch. A review of Web searching studies and a framework for future research. *J. of the American Society for Information Science and Technology*, 52(3):235–246, 2001.

[12] B. J. Jansen, A. Spink, and J. Pedersen. An analysis of multimedia searching on AltaVista. In *Proc. of the 5th ACM SIGMM Int. Workshop on Multimedia Information Retrieval (MIR)*, pages 186–192, 2003.

[13] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. of the eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 133–142, 2002.

[14] C. B. Jones, A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In *Proc. of the 3rd Int. Conf. on Geographic Information Science*, pages 125–139, 2004.

[15] M. Kamvar and S. Baluja. A large scale study of wireless search behavior: Google mobile search. In *Proc. of the SIGCHI conference on Human Factors in Computing Systems*, pages 701–709, 2006.

[16] R. R. Larson. Geographic information retrieval and spatial browsing. In L. Smith and M. Gluck, editors, *GIS and Libraries: Patrons, Maps and Spatial Information*, pages 81–124, 1996.

[17] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in Web search. In *Proc. of the 14th Int. Conf. on the World Wide Web*, pages 391–400, 2005.

[18] J. L. Leidner. Toponym resolution in text: Which sheffield is it? In *Proc. of the 27th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 602–602, 2004.

[19] D. Lewandowski. Query types and search topics of German Web search engine users. *Information Services and Use*, 26:261–1269, 2006.

[20] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger. Design and implementation of a geographic search engine. In *8th Int. Workshop on the Web and Databases (WebDB)*, 2005.

[21] B. Martins, M. Silva, and L. Andrade. Indexing and ranking in GeoIR systems. In *Proc. of the 2. Int. Workshop on Geo-IR*, 2005.

[22] K. McCurley. Geospatial mapping and navigation of the web. In *Proc. of the 10th Int. Conf. on the World Wide Web*, pages 221–229, 2001.

[23] G. Mishne and M. de Rijke. A study of blog search. In *Proc. of the European Conf. on Information Retrieval*, pages 289–301, 2006.

[24] Y. Morimoto, M. Aono, M. Houle, and K. McCurley. Extracting spatial knowledge from the web. In *Proc. of the Symp. on Applications and the Internet*, pages 326–333, 2003.

[25] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proc. of the 1st Int. Conf. on Scalable Information Systems*, 2006.

[26] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. In *Proc. of the Eleventh ACM SIGKDD Int. Conf. on Knowledge Discovery in Data Mining*, pages 239–248, 2005.

[27] D. E. Rose and D. Levinson. Understanding user goals in Web search. In *Proc. of the 13th Int. Conf. on the World Wide Web*, pages 13–19, 2004.

[28] T. Sanderson and J. Kohler. Analyzing geographic queries. In *Proc. of the Workshop on Geographic Information Retrieval*, 2005.

[29] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large Web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[30] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the Web: the public and their queries. *J. of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

[31] D. Stenmark. One week with a corporate search engine: A time-based analysis of intranet information seeking. In *Proc. of the Americas' Conf. on Information Systems*, 2005.

[32] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 718–723, 2006.

[33] Q. Tan, X. Chai, W. Ng, and D. Lee. Applying co-training to clickthrough data for search engine adaptation. In *Proc. of the 9th Int. Conf. on Database Systems for Advanced Applications (DASFAA)*, 2004.

[34] J. Teevan, E. Adar, R. Jones, and M. Potts. History repeats itself: Repeat queries in yahoo's logs. In *Proc. of the 29th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 703–704, 2006.

[35] S. Vaid, C. B. Jones, H. Joho, and M. Sanderson. Spatio-textual indexing for geographical search on the web. In *Proc. of 9th Int. Symp. on Spatial and Temporal Databases (SSTD)*, 2005.

[36] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, and Y. Li. Detecting dominant locations from search queries. In *Proc. of the 28th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2005.

[37] V. Zhang, B. Rey, E. Stipp, and R. Jones. Geomodification in query rewriting. In *Proc. of the Workshop on Geographic Information Retrieval*, 2006.

[38] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W. Ma. Hybrid index structures for location-based web search. In *Proc. of the 14th ACM Int. Conf. on Information and Knowledge Management*, pages 155–162, 2005.