International Journal of Geographical Information Science
Vol. 22, No. 3, July 2008, 1–20

Geographic Intention and Modification in Web Search

Rosie Jones, Wei Vivian Zhang, Benjamin Rey, Pradhuman Jhala and Eugene Stipp,
Yahoo! Research 3333 Empire Ave, Burbank, CA, 91504, USA

(November 30th 2006)

Web searchers signal their geographic intent by using place-names in search queries. They also indicate their flexibility about geographic specificity by reformulating their queries. By examining this data we can learn to understand web searcher flexibility with respect to geographic intent. We examine aggregated data of queries with locations, and locations identified from IP addresses, to identify overall distance preferences, as well as distance preferences by search topic. We also examine query rewriting: both deliberate query rewriting, conducted in web search sessions, and automated query rewriting, with manual relevance judgements of geo-modified queries. We find geo-specification in 12.7% of user query rewrites in search sessions, and show the breakdown into sub-classes such as same-city, same-state, same-country and different-country. We also measure the dependence between US-state-name and distance-of-modified-location-from-original-location, finding that Vermont web searchers modify their locations greater distances than California web searchers. We find that automatically-modified queries are perceived as much more relevant when the geographic component is unchanged. We look at the relationship between the *non-location* part of a query and the distance from the user. We see that people search for *child day-care* near their locations and *maps* far from where they are located. We also give distance profiles for the top topics which cooccur with place-names in queries, which could be used to set document priors based on document proximity and query topic.

1 Introduction

In order to design information retrieval systems that take geography into account, we need to understand users' geographic information needs. One way to survey the geographic distribution of web searchers' information needs is to analyze user queries which explicitly incorporate geographical information. Sanderson and Kohler (2004) examine user search queries in an Excite query log, finding that users frequently explicitly specify their geographic preference when querying a search engine. Gravano *et al.* (2003) automatically classify queries into *local* and *global*, independent of whether they contain place-names, based on the prevalence and diversity of place-names in search results for the

International Journal of Geographical Information Science ISSN 1365-8816 print/ ISSN 1362-3087 online
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810xxxxxxxxxxxxx

queries. For global queries they propose reranking documents to return the most global documents, while for local queries they propose appending the user's location, if the query does not already contain a location. Such a system could be even more effective if it could take into account both the user's location and locations at appropriate distances from the user, given the topic of the query. We will show distance profiles which could be used to set parameters for this kind of a system.

Systems for performing spatial query-expansion (Fu *et al.* 2005) could benefit from an empirical understanding of users' preferences: in some regions we may be able to justify greater distances in spatial query-expansions than in others. Cai (2007) shows for example that user-understanding of "near" varies for different shopping contexts. Fu *et al.* (2005) show systems for generalizing both locations and nearness. It is also possible to identify and disambiguate locations in web pages, as well as identify the correct place in a taxonomy of locations (Amitay *et al.* 2004).

Spatial ranking ranks geographic regions according to their relevance to a query-placename, by considering the degree of overlap between the query geographic-region and the candidate result regions. Different ranking functions weight overlap differently, by normalizing by query-region size, result region size, and so on (Larson and Frontiera 2004). We can introduce spatial ranking to an information retrieval system by further optimizing the weighting of the geographic and topical parts of the query. In doing this it would be good to include information about user preferences with regard to topical and geographic matching: for example, how do users view the trade-offs between less-topical matches, versus more-distant matches, when attempting to increase recall. In particular, if the distance-relevance varies with the topic, this is important to capture in the relevance function. We will discuss in Section 4 how the importance of proximity differs for users searching for "restaurants" than for "hotels".

We may be able to quantify notions of nearness for large populations of web-searchers by looking at their geographical preferences as exemplified in web searches. We know of no previous study looking at the relationship between user location and the locations they specify in their search queries. Query reformulation in search engines is extremely common (Spink *et al.* 2000, Jones and Fain 2003) but no previous work has studied the geographic component of query reformulation.

In Section 2 we give a brief overview of the technology we employ to identify placenames in web search queries. In Section 3 we summarize basic statistics about place-name occurrence in search queries. In Section 4 we look at distances between the place-name in a query and the user's location, as indicated by the IP address. We also show how different query topics have different distributions in distance from the user's location. In Section 5 we examine how

users manually modify the location in their query when they reformulate in search sessions. In Section 6 we examine how users might respond to a system for automatically modifying the placename in a search query through a manual evaluation of rewrite quality. In Section 7 we sum up our results and give recommendations for supporting user geographic intent in web search.

2 Identifying Places in Web Searches

In this section we give a brief overview of our proprietary system for identifying place-names in queries (Riise *et al.* 2003), which we will use as a black-box for automatic analysis in later sections. Our global locations database contains zip-codes, towns, suburbs, and states as well as colloquial names and places of interest (e.g. Eiffel Tower). Identifying places-of-interest has been addressed using web-page context and geo-spatial algorithms (Arampatzis *et al.* 2006). Knowing whether a query is related to a location is not as simple as looking up the potential place-name in our locations database, since there are towns called "Spears", "Cars", "Music", "Hotel", and so on. Once we have identified whether a query is location-related there is also the problem of identifying which of a potentially long list of locations the user has in mind. There are, for example, more than 900 places world-wide called "San Jose", including one in California, USA, and one in Costa Rica.

2.1 *Identifying Place-names in Queries*

In order to identify place-names in queries, we use a function of pre-computed scores for each term in the query. Each term in the locations database has a pre-calculated "location-related probability" in the range $[0, 1]$ which is a context-independent prior probability of the term being a location, or a non-location homograph. Context words (e.g. "in", "at", "mr") and a database of non-places (e.g. "Paris Hilton" or "George Washington") affect the final location-related probability of a query. Locations with the same name are disambiguated based on their frequency in a corpus, population (similar to the approach used by Leidner (2006)), and the location of the user. In addition, for the analysis discussed in Section 4, we use the query IP address to identify the location it was issued from. In order to identify the location from the IP address, we can use information supplied by the Regional Internet Registry (RIR), which is a governing body that is responsible for the administration of Internet addresses in a specific geographic region. The RIR database contains IP addresses, Internet Service Providers (ISPs), and general geographic location. Using this source, it is possible to determine the Internet Service Provider (ISP) and the state and city for an IP address.

2.2 *Accuracy of Place-name Identification*

An editorial team manually labeled 10,000 queries and identified those which are location-related. Location-related queries were then manually disambiguated if needed (for example, deciding whether a query was more likely to be about Margate, UK than Margate, Florida, given the other terms in the query). We ran our location-identifier through the same set of queries and determined that our software can reach near-human performences, *i.e.* about the same accuracy as inter-annotator agreement, on the combined task of identifying and disambiguating place-names.

2.3 *Measuring Distance*

In Sections 4 and 5 we perform analysis in terms of the distance between two places relevant to a query. To obtain these distances we first map each place to a longitude and latitude. To obtain distances between two place-names, we use the standard spherical distance (Wikipedia 2007). When the place-name is a general area, such as a state-name, we map the position to the bounding box. For disjoint places we can then define distance as distance from the center of one bounding box to another. The distance from a place to itself is always zero. We have interesting choices to make when defining distance between one place-name and a second which encloses it (eg distance from "Los Angeles, California" to "California") or conversely from a place-name to a location inside it. We respect topological containment here: when a distance is from a location to the location inside it (eg from "California" to "Los Angeles") we define the distance as zero – since a web searcher looking for things in California may have their needs satisfied by things in Los Angeles. From the inner location to the outer we use the distance between the centers of the bounding boxes, since a web searcher looking for something in the city of Los Angeles may have to go a long distance if results are generically for the state of California. This the distance is not symmetric, but matches intuitions about web search results satisfying user needs.

An interesting additional topological aspect we did not consider in this work is adjacency: for example it might be interesting to consider that two cities or states are adjacent, rather than using the distance of the centroids of their bounding boxes. We leave this kind of consideration for future work.

3 Geographic Information in Queries

We use the location identification algorithm described in Section 2 to identify place-names in queries. In the remaining sections, all place-names are those

identified using this algorithm.

We randomly sampled queries from Yahoo! query logs in the US. For each query we resolve the IP address it came from to give the geographic location the query was issued from. No identifying information was attached to the queries and we performed all analysis on aggregated data, in accordance with Yahoo!'s privacy policy. All queries were automatically spell-corrected, using a state-of-the-art high precision search engine spell-corrector, with no special treatment of place-name spell-correction. We examined this sample using our software and found that 12.7% of queries contained a placename. This is comparable to the 14.8% found by Sanderson and Kohler (2004).

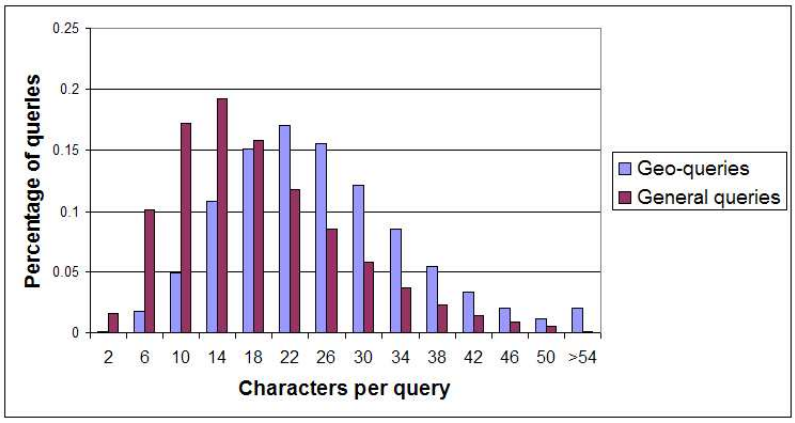
3.1 *Characteristics of queries with a place name*

To start our analysis, we looked at the number of characters and words in the queries which have place-names. The average number of characters per query is 25.1. The average number of words per query is 3.8 which is comparable to the 3.3 words found by Sanderson and Kohler (2004). Figure 1 shows the distribution of characters and words per query, contrasted with the distribution for all queries. As Sanderson and Kohler also found, both are greater than the statistics published for general search queries. For general search queries, the average number of characters per query has been reported as 15.5, while the average number of words per query has been reported as 2.7 (Spink *et al.* 2000). An interpretation of this data is that geospecified queries are about a word longer than typical queries, since they are like typical queries but with the addition of a place-name. We could use this information to aid in classifying whether a query contains a location: query length could be an additional predictor or modifier of the prior, in addition to other information.

3.2 *Distribution of Place-names in Queries*

When we inspect the distribution of place-names in queries, we find that queries contain city names much more commonly than country names, and country names more commonly than state names. This may indicate that most users are looking for specific information at the city level. The country-level queries may be due to users' interests in culture or travel planning. Table 1 shows the distribution of queries into the categories *state*, *country* and *city*.

For queries with place names, we found that of 73.8% places searched for are within the United States, while 26.2% of places searched for are outside the United States. This shows that there are significant international interests for users who submit queries on the United States site.



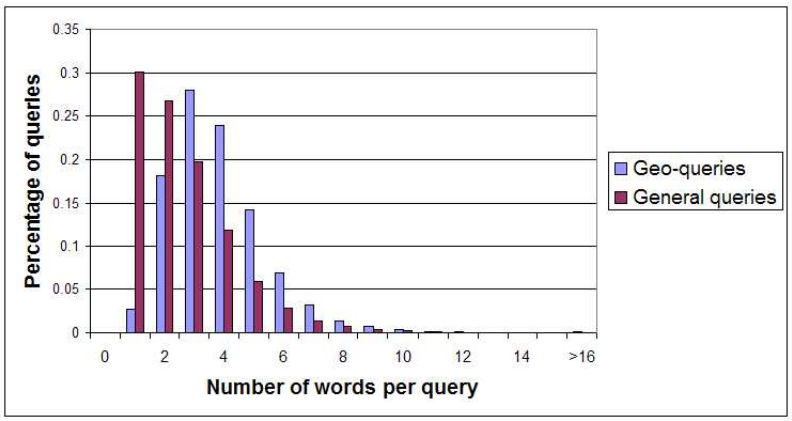


Figure 1. Distribution of number of characters and words per query for queries with place names. These queries are longer than the average search query.

location level	percentage
city	83.77%
state	2.54%
country	13.69%

Table 1. The distribution of web search queries containing place-names into the categories *state*, *country* and *city*.

4 Distance from Home of Locations Specified in Queries

When we look at aggregated distances between the locations specified in queries, and the locations of the IP address they are issued from, we can begin to understand how user topics are geographically distributed relative to the search location. In this section we examine the distribution of the distances

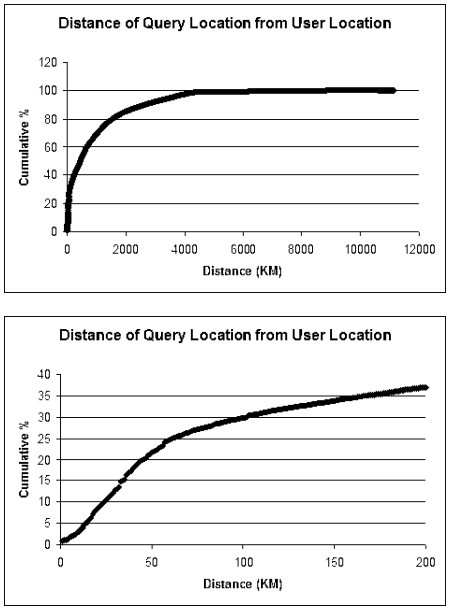


Figure 2. Cumulative distribution of the distance between the source IP address and the place-name identified in a query. The bottom figure is a zoom of the top figure. Geo-specified queries tend to be near the searcher's location.

overall, and see that most queries containing place-names are for locations close to the user. We also look at how different topics affect this distribution, since users may be interested in nearby schools, for example, but hotels in more distant locations. We take a dataset filtered to include only IP addresses in the United States and queries containing locations in the United States, to eliminate web searches for overseas travel or cultural interests.

4.1 *Distance Overview*

Internet access and search engines allow us to look for items without geographic constraints. However, when we look at the cumulative distribution of user searches by the distance from home specified in the query we see in Figure 2 that the vast majority of geo-specified queries are for locations near the searcher. When we zoom in we see that around 20% of the instances of place-names specified are for locations within 50km of the user's location, and 30% are within 100km.

quantile	max. distance
1	0
2	15
3	25
4	35
5	47
6	66
7	104
8	173
9	253
10	364
11	466
12	566
13	686
14	867
15	1053
16	1268
17	1569
18	2002
19	2690
20	3571

Table 2. Distance quantiles for distances of user-location from query-specified-location. Each quantile is shown with the maximum distance for that quantile.

4.2 *Distance By Query Topic*

Cai (2007) showed that user-understanding of "near" varies for different shopping contexts. By examining topics in user search queries, we can obtain a much more nuanced picture of user understanding of acceptable distances. In order to identify the topic of a query, we remove the part of the query which is the place-name, then consider the remaining string to be the "topic". For example, for the query "california maps" the topic is "maps".

In order to generate topical distance profiles, we generated bins based on quantiles from on the overall sample of IP-query-locations distances. The first bin is for queries at distance zero from the IP-location (same place or sub-region), while the remaining 19 bins each contain one-twentieth of the non-zero distance IP-query instances. The highest quantile was excluded to remove outliers. The distances for the quantiles are shown in Table 2.

In Figure 3 we show distance-profiles for 18 common query topics. Each bar is a distance quantile, and the height of the bar is the proportion of that topic's queries in the quantile. We see that users prefer to search for restaurants close to home, while hotels can be further away, and real estate is relatively uniformly distributed. Queries for "lottery" show a peak around distances at the order of magnitude of states, since lotteries are typically run at the state level and we see queries for, for example, `colorado lottery` and `pennsylvania lottery`.

These profiles could be used in several ways to aid web search retrieval. One is to inform topic-specific geographic relaxation or modification on the query side. The second is to assign a topic-specific prior distribution over geographic regions for documents retrieved in response to the query.

Now let us consider, for each quantile, the dominant topic for that quantile. For each topic, we calculated the probability distribution over distance

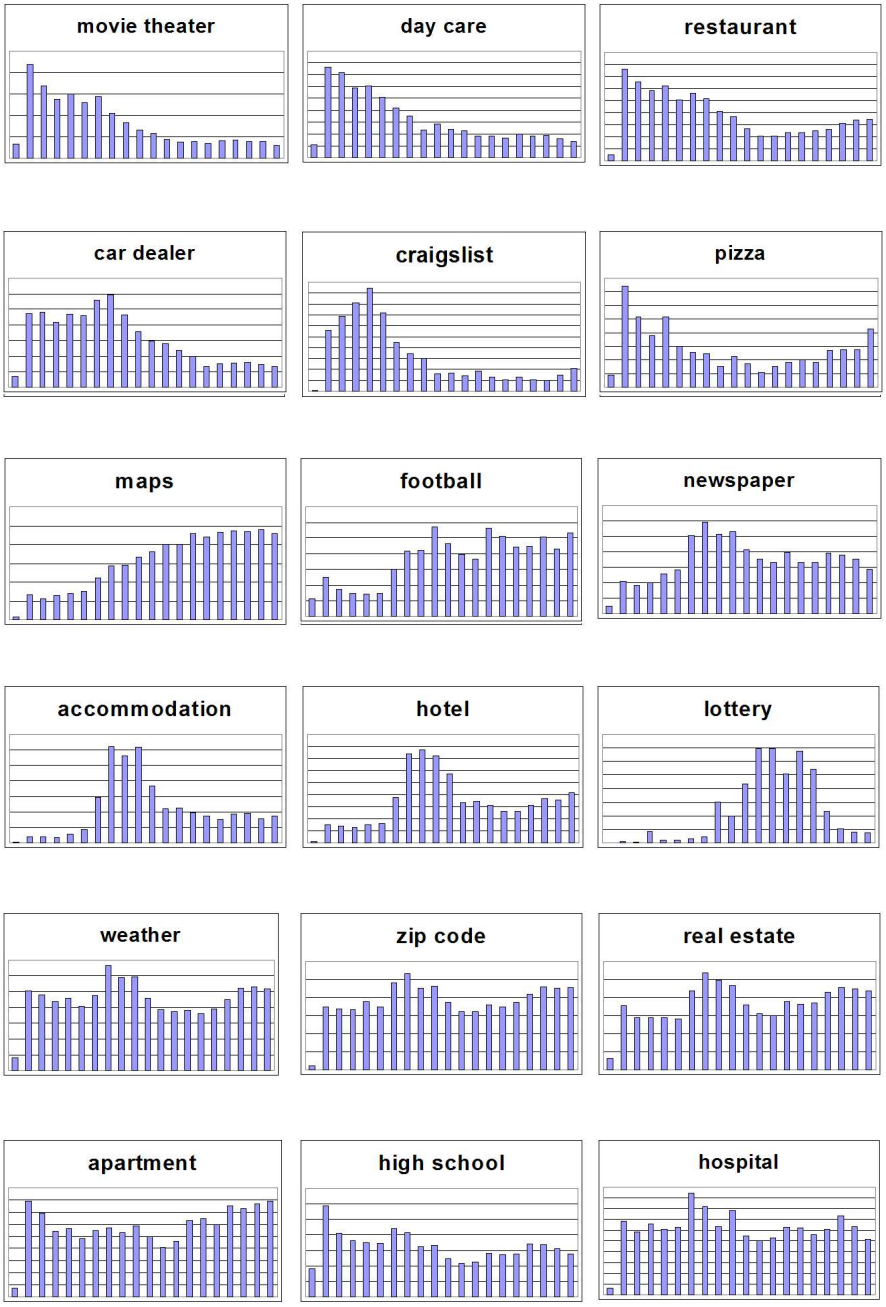


Figure 3. Aggregated profiles for common query topics, in terms of distance from the IP location to the location specified in the query. Each bar is a distance quantile for 20 quantiles, and the height of the bar is the proportion of that topic's queries in the quantile.

quantile	max. distance	topic
1	0	high school
2	15	restaurant
3	25	craigslist
4	35	craigslist
5	47	craigslist
6	66	craigslist
7	104	job
8	173	accommodation
9	253	accommodation
10	364	accommodation
11	466	lottery
12	566	lottery
13	686	lottery
14	867	lottery
15	1053	lottery
16	1268	lottery
17	1569	map
18	2002	map
19	2690	map
20	3571	map

Table 3. Top conditional topic per distance quantile. Each quantile is shown with the maximum distance for that quantile. We see that the topic "high school" has the highest conditional probability of the searcher being in the same region as the location, while the top nearby topic is "restaurant". "lottery" is the most probable intent for larger distances (typically a web searcher searches within his or her state) while "map" is characteristic of greater separation: people tend to search for maps outside their state.

quantiles, then chose the topic with the maximal $p(quantile|topic)$ as the characteristic topic for that quantile. In Table 3 we see that the dominant topic for queries with place-names in the same location as the query-IP is "high school" (more of *high school*'s probability mass is in quantile 1 than any other topic). "Restaurant" is the dominant topic for queries 0–15km from the query-IP-location. "Craigslist"[1] dominates quantiles 3 through 6, representing searches at the city and nearby city level. "Job" is the dominant topic for quantile 7. "Accommodation" is the dominant topic from 104 km to 364 km from the searcher's location. State-wide lotteries dominate searchers in regions 364 to 1268 km from the searcher. The top quantiles are dominated by "map" searches, which are often for places far from home.

5 Geographic Information in Reformulated Queries

Web searchers commonly reformulate their queries (Spink *et al.* 2000, Jones and Fain 2003), with actions such as inserting, deleting and substituting terms, as well as re-phrasing the query. For example, a web searcher looking for a place to order pizza near their home might first query for "pizza", then "pizza altadena", then "pizza pasadena", then "pizza 91101" as they try different ways of finding pizza parlors that might deliver to their home. By examining sequential queries issued in anonymous query sessions we can identify these

[1] `craigslist.org` is a popular website with classified ads for many cities in the US, as well as internationally. Common queries are `craigslist san francisco` and `craigslist new york`

rewrites. One of the refinements a user might use is changing the location part of the query (around 10% of the query rewrites). They can specify a place name if the query did not initially contain one (E.g.: "dry cleaner" → "dry cleaner pasadena"), remove a place-name, or change it to another location (E.g.: "french restaurant in venice beach, california" → "french restaurant in santa monica, california"). In this section, we will study the type of geo-modifications performed by users in query reformulation. This modification of location intent from one query to the next may specify a searcher's willingness to travel: if "pasadena" and "altadena" are close enough to use as possible pizza sources, this may be considered "near" from the point of view of the searcher, when considered in the context of "pizza". This contrasts with the proximity measures we considered in Section 4, where we considered the proximity of the location from a single query to the searcher's home location, again in the context of a specific intent.

5.1 *Sampling Query Rewrites for Automated Analysis of Geo-Reformulation*

We take sequential pairs of queries and look for geo-modification from one query from a user to the next, aggregating over many users to find that, for example, "edinburgh" is frequently modified by web searchers to "glasgow". About 50% of sequential query pairs are reformulations (Spink *et al.* 2000, Jones and Fain 2003). In order to triage the coincidentally cooccurring pairs from deliberate reformulations, we used only query pairs which passed one of several filters. Firstly, we required all pairs to have occurred at least three times. The second filter is the log-likelihood ratio test (Manning and Schuetze 1999, Dunning 1993) which reduces the spurious cooccurrences from 50% of the data to about 10% of the data. The third is based on considering query pairs with small character or word edit distance which independently reduces the spurious coincidence rate to about 15% (Jones *et al.* 2006). The reason for including small edit distance rewrites is it allows us to include rare rewrites such as `akron canton airport` → `columbus airport`. After these filters, we applied a fourth, retaining only query pairs which have a location identified in both queries. This further reduces the misidentified pairs. These filters are summarized in the first four steps in Algorithm *1.*

5.2 *Geomodification and Geocorrection*

In Table 4 we see several place-names used in reformulations for "edinburgh" by web searchers in the United States. The reformulation from "edinburgh" to "glasgow" may come from US web searchers seeking various holiday destinations in Scotland. However, the reformulation from "edinburgh" to "edinburg

Query Rewrite Pair	LLR Score	
edinburgh	→ glasgow	7084
edinburgh	→ scotland	4267
edinburgh	→ york	1658
edinburgh	→ aberdeen	1273
edinburgh	→ london	1185
edinburgh	→ fraser	1089
edinburgh	→ uk	807
edinburgh	→ edinburg texas	731
edinburgh	→ edinburg	689
edinburgh	→ edinburg tx	686

Table 4. Place-names commonly appearing as rewrites for Edinburgh, along a significance score based on the log-likelihood ratio (LLR) test. Rewrites of Edinburgh (in Scotland) to London (in England) are presumably by users looking for information about castles or other sight-seeing in the British isles. Rewrites of Edinburgh to Edinburg TX may be *geocorrections* as we discuss below.

tx" is more likely to come from a user seeking the Texas city called Edinburg, and reformulating the query to spell-correct and disambiguate it. We would like to distinguish between deliberate geographic reformulation, which may reflect the proximity preferences of a user relaxing the geographic constraints of their query, and this kind of *geocorrection*, which we discuss in more detail below.

5.3 *Identifying Geospecified Reformulations*

We manually labeled 108 queries which had been through these three filters, and found that only 10 were coincidentally cooccurring query pairs. While this is a small sample, it gives us a feel for the general quality of the filters. Table 5 shows a summary of the accuracy of geo-modified reformulation. About 9% of the query pairs were coincidental cooccurrences which probably signaled different user intents. 80% were reformulations, with the user modifying either the location or the topic from one query to the next. (In general as calculated over much larger samples 73% are distance zero, ie the topic is changed rather than the location, as we will discuss further below.) The remaining 11% were what we call a *geocorrection*. This is where the user issues a query containing a placename eg "santiago", then their next query is a modification to disambiguate it, eg "santiago cuba". We also see users spell-correcting their queries, eg "ravenwood" → "ravenswood". These pairs lead to large apparent distances between query and reformulated query, but the distance does not reflect user flexibility about geographic distance.

We might expect that including international locations includes more possible candidates for place-name ambiguity. However, when we filter for query pairs with both locations identified as being in the US, the proportion of geocorrections increases (Table 6) which suggests that even within the United

Type	Description	Count	Example
Reformulation	Geomodification or intent modification	86	edinburgh → glasgow
Geocorrection	Correction of spelling or placename disambiguation	12	edinburgh → edinburg tx
Not a reformulation	Coincidentally cooccurring query pair	10	edinburgh → ryanair

Table 5. A manual labeling of 108 potential query reformulations with locations in both the original query and the reformulated query.

Reformulation	62
Geocorrection	9
Not a reformulation	6

Table 6. A manual labeling of 77 potential query reformulations with a location in the United States in both the original query and the reformulated query.

States, place-name ambiguity is a significant problem for web searchers. It could be helpful to design retrieval systems to support users in this (for example, identifying when a placename is ambiguous, or suggesting disambiguations based on the user's registered location).

In order to address these geocorrections we added an additional filter. Any query pair with spatial distance over 50 km and with word insertions as the only change was identified as a geocorrection (for example "farmington high school" → "farmington ct high school"). Any query pair with spatial distance over 50 km with a single character edit difference was also called a geocorrection (for example "ravenwood" → "ravenswood"). These filters were sufficient to identify all geocorrections in the sample of 108 query pairs, without causing any false positives. We used these filters on the data in subsequent sections, for analyzing the distances between a place a user searches for initial, and an alternative subsequent place they may query for as an alternative source for, say pizza, or location for a holiday.

5.4 *Distance Between Query Place-name and Reformulated Query Place-name*

In Figure 4 we see the cumulative distribution of distance between query location and reformulated query location. 73% of reformulations have distance zero. These zero-distance reformulations are due to three cases: (1) the topic is reformulated and the location is left unchanged, (2) the location is reformulated to a synonym, or (3) the location is reformulated to a location contained within the original location. When we zoom in, we see that over half of all reformulations of distance greater than zero are within 100 km of the original query location. This contrasts with IP-location to query-location distances, where we found only 30% of distances less than 100 km. Once a user has specified the location of interest in the query, they are not very flexible in reformulating. An interesting further extension to this study would be to examine triples of IP-location, query-location and reformulated-query-location to see the impact

Algorithm *1* Algorithm for identifying geo-reformulations from query sessions, and distinguishing geo-corrections from geo-reformulations.

Identify sequential query pairs: two queries in succession from the same user
Keep only pairs which

- occur more than three times
- occur more than chance using log-likelihood ratio test OR
- have levenshtein character edit distance $< 40\%$ (to capture spelling changes) OR
- have word overlap $> 40\%$ (to capture query refinement)

Keep only pairs with a location in both query and rewrite
Remove pairs with spatial distance > 50 km and with word insertions as the only change (these are geo-corrections)
Remove pairs with spatial distance > 50km and single character edit distance (these are geo-corrections)

type of location rewrite	CA	VT
same place	20%	16%
place change	15%	20%
place insertion	34%	33%
place deletion	31%	31%

Table 7. Distribution of types of geomodification for queries with places in US states California (CA) and in Vermont (VT).

of geographic distance from the user's location, on geographic reformulation flexibility.

5.5 *Sampling Query Rewrites in California and Vermont*

Users may have a different way of rewriting their query depending on the location they are searching for. We could refine a user's definition of proximity or nearness depending on the geographic region they are searching for. To perform this experiment, we chose two different states in the US: California, and Vermont, and used query reformulation pairs where one of the query contained a location in either California or Vermont.

People specifying Vermont in their queries tend to modify the location more often than people specifying California (20% vs 15%, shown in Table 7). This could mean that web results are better defined for locations in California, or simply that it is easier to find things online in California than in Vermont.

When both the initial query and the reformulated one had a place-name, we computed the distance between these two places. We binned the distances

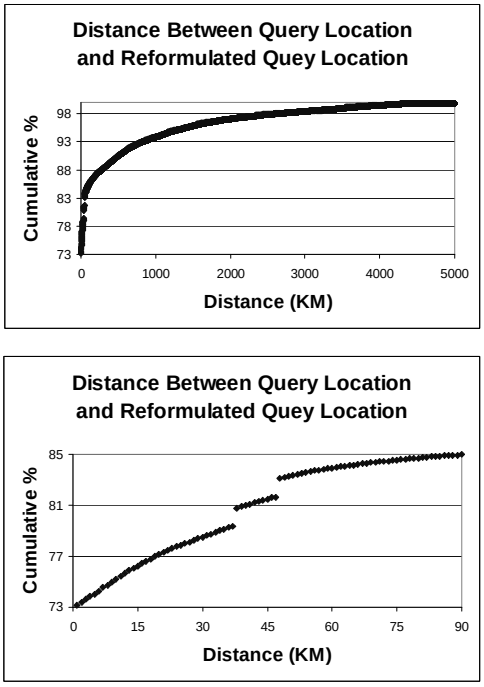


Figure 4. Distance between the place-names in a query and its reformulation, for query pairs constrained to be within the United States, with geocorrections automatically filtered. Query reformulations tend to lead to the same location, or a nearby location. The second graph is a zoom of the first.

into six ranges, ranging from local (0-10 km) to very long distances (3000+) (see Figure 5). We can have very long distances in reformulations when, for example, a query referring to California is reformulated into a a query about a different state or country.

The main difference between the two states is that people in California reformulate their queries to a neighborhood location (<10 km) much more often than people in Vermont, where in contrast queries tend to be reformulated to a county-level location (50-100 km). The median distance for a California query rewrite is 615 km and it is 1267 km for Vermont. Here again, we can see that Californians find what they're looking for much closer to home than people from Vermont.

These two experiments suggests that for queries with a location, including web results spanning not only the given location, but also surrounding locations would help people from Vermont more than people from California. And the type of surrounding locations should be at the neighborhood level for California and the at county level for Vermont.

More generally, these experiments suggest interesting analyses that can be

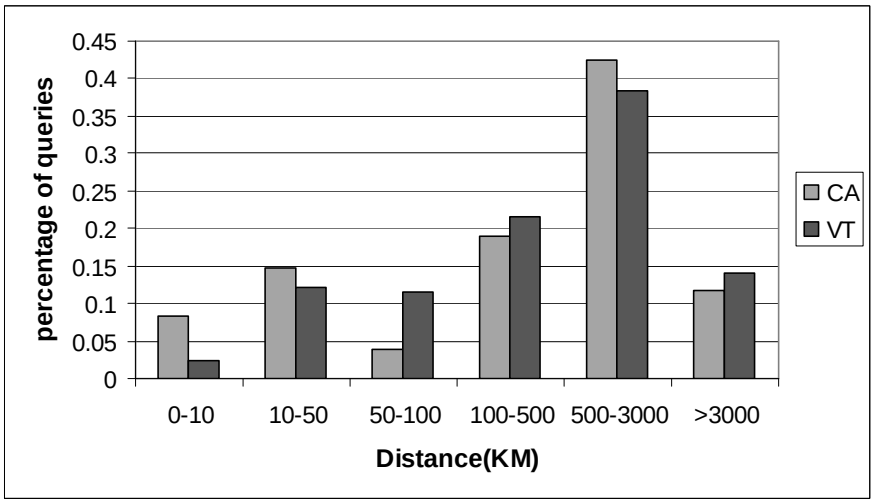


Figure 5. Reformulations with a place-change tend to involve shorter distances when one of the places is in California, than when one of the places is in Vermont.

performed when we drill down into distances in spatial reformulation according to user location. In future work it would be interesting to perform this kind of analysis across other types of geographic distinctions, such as rural and urban locations.

6 Perceived Relevance of Automatic Geographic Reformulation

As we have seen, users often rewrite their queries by modifying the location part. In previous work (Jones *et al.* 2006), we described an algorithm to mine sequential queries and use these to generate automatic rewrites. In generating automatic rewrites, we treat place-names the same as all other query terms. For example, the query "castles near edinburgh" has three phrases we could modify, and based on user query rewrite session distribution, candidate rewrites for each phrase include "castles" → "medieval castles", "near → "in" and "edinburgh" → "london". Table 4 shows place-names commonly used to replace Edinburgh, based on logs for users searching on the US Yahoo! web-site.

When we examine rewrites performed automatically by our location-agnostic rewrite system, we find that a substantial proportion of these rewrites (see Figure 6) are location modifications. Thus we should understand how changing the location of a query affects the quality of the rewrite.

We had human annotators evaluate the rewrites using the following labels:

1 user intent is respected
2 slight shift in user intent, but closely related

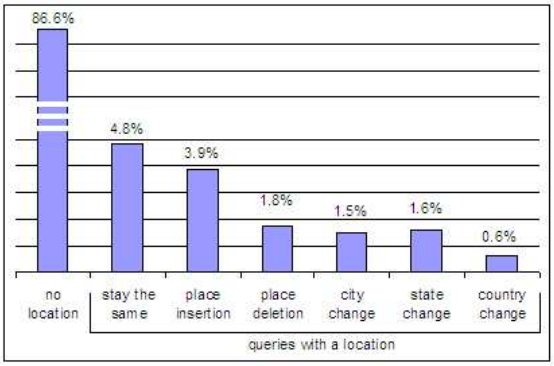


Figure 6. Type of substitution for auto-rewritten queries. We see that around 10% involve changing the location part of the query.

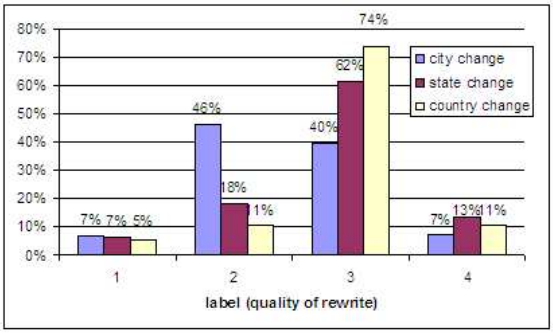


Figure 7. The perceived quality of an auto-rewrite depends on the type of location modification. City changes were more likely to be labeled 1 or 2 (good rewrites), while country changes were more likely to be labeled 3 or 4 (fair or bad rewrites).

3 related to initial query
4 unrelated

Labels 1 and 2 are considered to be good (excellent and good) rewrites, and labels 3 and 4 are considered to be bad (fair and poor). Of the query rewrites we had labeled, we isolated those in which a place name had been identified and modified (505 query pairs). For these queries, we identified their city name, state and country.

Human labelers find that a city name change is good 50% of the time (see Figure 7), while state and country changes are good 25% and 16% of the time. A state-change tends to be labeled as a fair rewrite (related but less relevant) 62% of the time, while a country change is fair 74% of the time. Poor rewrites

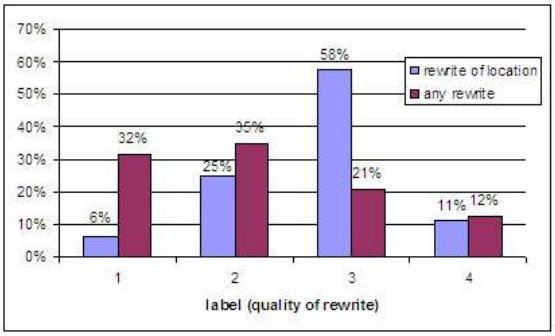


Figure 8. Perceived quality of query auto-rewrites. Overall rewrites with location-changes were more likely to be perceived as fair or poor (label 3 and 4) than rewrites in general.

initial query	suggestion	label	type of modification
elite vietnam	elite china	4	country change
indonesia calling card	australia calling card	4	country change
days inn toronto	days inn quebec	3	state change
land for sale in maryland	land for sale in california	4	state change
south korea	seoul	2	state change
days inn toronto	days inn mississauga	2	city change/ same state
syracuse newspapers	binghamton newspapers	3	city change / same state
disney orlando	disney florida	1	city change/ same state

Table 8. Examples of place name rewrites

(label 4) are more commonly identified for state and country changes than for city changes.

In Table 8 we see examples of why city changes are more acceptable, since they frequently involve changing a city to a nearby city (E.g.: "toronto" to "mississauga"). Another type of good rewrite is when the city name is unnecessary because the state name is enough to disambiguate the intent (E.g.: "disney florida").

Overall, compared to other rewrites, (see Figure 8), location change seem to be much riskier. Changes at the country and state level are generally considered poor, and even rewrites at the city level have a lower precision than the average rewrite (50% compared to 67%).

7 Conclusions and Future Work

By examining aggregated logs of queries containing place-names, and looking at the distances from IP-location and reformulated query-location, we have obtained some insights into user distance preferences in web search. We have observed that the locations in search queries are likely to be relatively close to the IP-location that issued the query. The shape of the cumulative distribution

for these distances could be used to inform priors on distances for retrieving documents when the query does not contain a location. We also examined IP-location to query-location distance profiles for specific query topics. We saw the distance profile varies greatly by query topic. This could allow us to build priors on distances for specific topics, for use either in query location relaxation, or document priors based on location.

We saw that arond 10% of query reformulations containing locations involve a *geocorrection*, in which a user either spell-corrects or explicitly disambiguates the location. We gave an algorithm for identifying these, to reduce noise in automated analysis. In general, we should provide location disambiguation assistance, much the way search engines today provide spell correction assistance.

Our study of automated query reformulation showed that annotators perceive changes to the location of the query to be much more harmful than modifications to the topic. In performing automated query reformulation, we should consider identifying synonyms, spell corrections or other small topic changes to queries, before attempting to modify the location.

An interesting further extension to this study would be to examine triples of IP-location, query-location and reformulated-query-location to see the impact of geographic distance from the user's location, on geographic reformulation flexibility.

We assumed in this work that the queries in a search query and search reformulation reflect a searcher's true geographic preferences. In practice, these may be constrained by search engine interfaces, paucity of results, etc. It would be helpful to follow-up this log-based study with a more interactive survey, in which web searchers are asked how far they are willing to look for things online.

REFERENCES

Amitay, E., Har'El, N., Sivan, R. and Soffer, A., 2004, Web-a-where: geotagging web content. In *Proceedings of the SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, United Kingdom (New York, NY, USA: ACM Press), pp. 273–280.

Arampatzis, A., van Kreveld, M.J., Reinbacher, I., Jones, C.B., Vaid, S., Clough, P., Joho, H. and Sanderson, M., 2006, Web-based delineation of imprecise regions.. *Computers, Environment and Urban Systems*, **30**, 436–459.

Cai, G., 2007, Contextualization of geospatial database semantics for Human-GIS interaction. *Geoinformatica*, **11**, 217–237.

DUNNING, T.E., 1993, Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, **19**, 61–74.

FU, G., JONES, C.B. and ABDELMOTY, A.I., 2005, Ontology-Based Spatial Query Expansion in Information Retrieval. In *Proceedings of the Lecture Notes in Computer Science, Volume 3761, On the Move to Meaningful Internet Systems: ODBASE 2005*, 3761, pp. 1466 – 1482.

GRAVANO, L., HATZIVASSILOGLOU, V. and LICHTENSTEIN, R., 2003, Categorizing web queries according to geographical locality. In *Proceedings of the CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, New Orleans, LA, USA (New York, NY, USA: ACM Press), pp. 325–333.

JONES, R. and FAIN, D.C., 2003, Query word deletion prediction. In *Proceedings of the SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Toronto, Canada (New York, NY, USA: ACM Press), pp. 435–436.

JONES, R., REY, B., MADANI, O. and GREINER, W., 2006, Generating query substitutions.. In *Proceedings of the WWW*, L. Carr, D.D. Roure, A. Iyengar, C.A. Goble and M. Dahlin (Eds) (ACM), pp. 387–396.

LARSON, R.R. and FRONTIERA, P., 2004, Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries. In *Proceedings of the Proceedings of the Research and Advanced Technology for Digital Libraries: 8th European Conference, ECDL 2004, Lecture Notes in Computer Science Series, LNCS 3232*, pp. 45–57.

LEIDNER, J., 2006, Experiments with Geo-Filtering Predicates for IR. In *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. (Lecture Notes in Computer Science, volume 4002)*, C. Peters, F.C. Gey, J. Gonzalo, G.J.F. Jones, M. Kluck, B. Magnini, H. Mller and M. de Rijke (Eds) (Springer), pp. 987–996.

MANNING, C.D. and SCHUETZE, H., 1999, *Foundations of Statistical Natural Language Processing* (MIT Press).

RIISE, S., PATEL, D. and STIPP, E., 2003, Geographical Location Extraction. *US Patent Application 20050108213*.

SANDERSON, M. and KOHLER, J., 2004, Analyzing geographic queries. In *Proceedings of the Proceedings of Workshop on Geographic Information Retrieval SIGIR* (New York, NY, USA: ACM Press).

SPINK, A., JANSEN, B.J. and OZMULTU, H.C., 2000, Use of query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policy*, **10**, 317–328.

WIKIPEDIA, "Great-circle distance — Wikipedia, The Free Encyclopedia", [Online; accessed 25-June-2007] 2007.