Geomodification in Query Rewriting

Vivian Wei Zhang, Benjamin Rey, Eugene Stipp and Rosie Jones
Yahoo! Research
3333 Empire Ave
Burbank, CA 91504
{zhangv, benjamir, estipp, jonesr}@yahoo-inc.com

ABSTRACT

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. We conducted experiments on geomodification in query rewriting. We examine both deliberate query rewriting, conducted in user search sessions, and automated query rewriting, with users evaluating the relevance 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 also find that automatically-modified queries are perceived as much more relevant when the geographic component is unchanged.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms: Algorithms

1. INTRODUCTION

In order to design information retrieval systems that take geography into account, we need to understand the 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 [10] examine user search queries in an Excite query log, finding that users frequently explicitly specify their geographic preference when querying a search engine. Gravano and co-authors [5] automatically classify queries into *local* and *global*, independent of whether they contain place-names, based on the prevalence and diversity of placed-names in search results for the 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.

Systems for performing spatial query-expansion [4] 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 [3] shows for example that user-understanding of "near" varies for different shopping contexts. Fu et al. [4] 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 [1]. We can introduce spatial ranking to an information retrieval system [8] but it would be good to include information about user preferences. 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 can obtain evidence from users' preferences from the way they modify their queries when interacting with a search engine. Query reformulation in search engines is extremely common [11, 6] but no previous work has studied the geographic component of query reformulation.

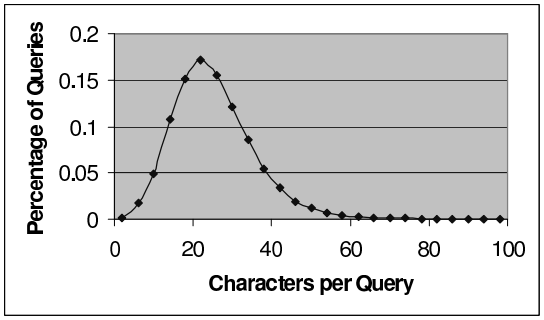
2. IDENTIFYING PLACE-NAMES IN QUERIES

In this section we give a brief overview of our proprietary system for identifying place-names in queries, 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 geospatial algorithms [2]. 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", etc. 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 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 $[0, 1]$. Context

GeoIR 2006 Seattle, WA USA



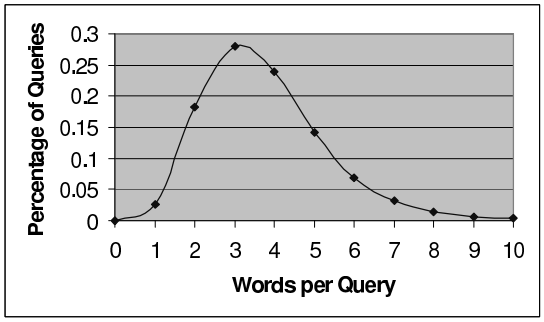


Figure 1: Distribution of number of characters and words per query for the queries with place names.

words and a database of non-places 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 [9]), and the location of the user.

2.2 Accuracy of Place-name Identification

An editorial team went through a sample of 10,000 queries and decided for every query whether the query is location-related or not. If a query is location-related, they then decided which location (e.g. Margate, UK vs Margate, Florida) the query is about. We ran our location-identifier through the same set of queries and determined that our software can reach near-human performences.

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 4 million queries from two weeks of Yahoo! query logs in the US, from February 14 – 18 2006. All queries were automatically spell-corrected. 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 [10].

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 [10]. Figure 3.1 shows the distribution of characters and words per query. 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 ([11]).

3.2 Distribution of Place-names in Queries

When we inspect the distribution of place-names in queries, we find that users use 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*.

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

Table 1: The distribution of queries into the categories *state*, *country* and *city*.

For the 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. Table 2 shows the top 20 popular countries in US queries.

4. GEOGRAPHIC INFORMATION IN REFORMULATED QUERIES

Web searchers commonly reformulate their queries [11, 6], with actions such as inserting, deleting and substituting terms, as well as re-phrasing the query. By examining sequential queries issued in anonymous query sessions we can identify these 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.

4.1 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. Analyzing such data could be interesting, as it can be taken as a proxy for people's every day behavior. For instance, we could refine a user's definition of proximity or nearness depending on where they are (or at least the geographic region they are searching for).

To perform this experiment, we chose two different states in the US: California, and Vermont, and extracted pairs of consecutive queries from Yahoo! logs where one of the query contained a location in either California or Vermont.

country	percentage
United Kingdom	3.44%
Canada	2.63%
Malaysia	1.35%
Philippines	1.25%
Italy	0.95%
Mexico	0.93%
India	0.89%
France	0.87%
Australia	0.85%
Japan	0.61%
China	0.60%
Spain	0.53%
Germany	0.52%
Singapore	0.52%
Indonesia	0.46%
Ireland	0.35%
United Arab Emirates	0.33%
Brazil	0.30%
Pakistan	0.27%
Thailand	0.25%

Table 2: The most popular countries searched for in the United States and their distribution in queries with place-names.

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

Table 3: Distribution of types of geomodification for queries with places in US states California (CA) and in Vermont (VT).

4.2 Type of geomodification

From table 3, we can see that among query pairs with a location identifier, people specifying Vermont tend to modify the location into another one more often than people specifying California (20% vs 15%). 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.

4.3 Distance in Place-change Rewrites

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 into six ranges, ranging from local (0-10 miles) to very long distances (3000+) (see figure 2). 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 miles) much more often than people in Vermont, where in contrast queries tend to be reformulated to a county-level location (50-100 miles). The median distance for a California query rewrite is 615 miles and it is 1267 miles 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.

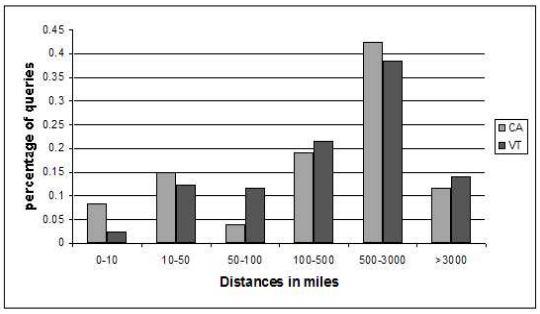


Figure 2: Reformulations with a place-change tend to involve shorter distances when one of the place is in California than when one of the places is in Vermont.

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 test. Rewrites of Edinburgh (in Scotland) to Dudley (in England) are presumably by users looking for information about castles in the British isles.

5. PERCEIVED RELEVANCE OF AUTOMATIC GEOGRAPHIC REFORMULATION

As we have seen, users often rewrite their queries by modifying the location part. In previous work [7], 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" → "dudley". 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 3) are location modifications. Thus we should understand how changing

the location of a query affects the quality of the rewrite.

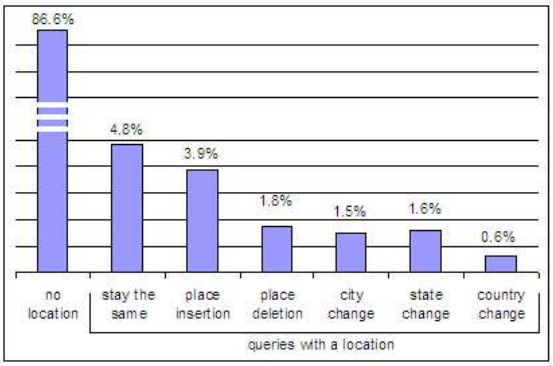


Figure 3: Type of substitution for auto-rewritten queries. We see that around 10% involve changing the location part of the query.

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
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 4), while state and country changes are good 25% and 16% of the time. A state-change tends to be labeled a fair (related but less relevant) rewrite 62% of the time, while a country change is fair 74% of the time. Poor (label 4) rewrites are more commonly identified for state and country changes than for city changes.

In table 5 we see examples of why city changes are more acceptable, since they frequently involve changing a city to

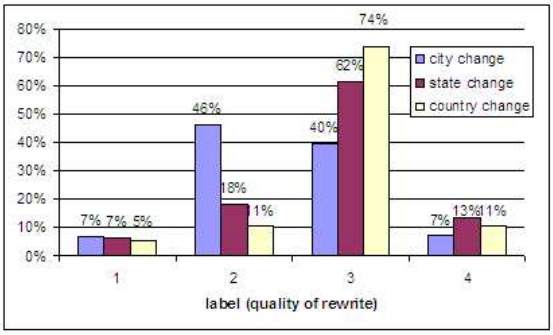


Figure 4: The perceived quality of an auto-rewrite depends on the type of location modification. City changes were more likely to be labeled 1 or 2 (good rewrites), while country changes were more likely to be labeled 3 or 4 (fair or bad rewrites).

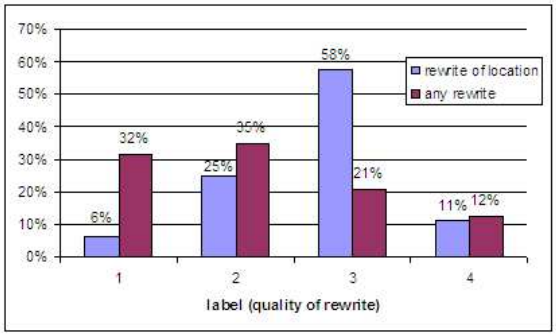


Figure 5: Perceived quality of query auto-rewrites. Overall rewrites with location-changes were more likely to be perceived as fair or poor (label 3 and 4) than rewrites in general.

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 5: Examples of place name rewrites

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 5), location change seem to be much riskier. Indeed, even the city change has a precision much lower than the average rewrite (50% compared to 67%).

6. CONCLUSIONS

We have described some of the phenomena that can be observed in user query rewrite sessions. Users modify the geographic component in their queries, the types of modifications they make may vary depending on their location, and users perceive changing the location part of a query as changing the meaning more than changing other parts of the query.

7. REFERENCES

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

[2] A. Arampatzis, M. van Kreveld., I. Reinbacher, C. Jones, S. Vaid, P. Clough, H. Joho, and M. Sanderson. Web-based delineation of imprecise regions. *The Journal of Computers, Environment and Urban Systems (CEUS)*, 2006. Hearst-like "X such as Y" to ascertain what Y is, then geo-spatial algs.

[3] G. Cai. Contextualization of geospatial database semantics for mediating human-gis dialogues. *Geoinformatica*, 2006.

[4] G. Fu and A. I. A. Christopher B. Jones. Ontology-based spatial query expansion in information retrieval. In *Lecture Notes in Computer Science, Volume 3761, On the Move to Meaningful Internet Systems: ODBASE 2005*, volume 3761, pages 1466 – 1482, 2005.

[5] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing web queries according to geographical locality. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 325–333, New York, NY, USA, 2003. ACM Press.

[6] R. Jones and D. C. Fain. Query word deletion prediction. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 435–436, New York, NY, USA, 2003. ACM Press.

[7] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *WWW '06: Proceedings of the 15th international World Wide Web conference*, 2006.

[8] R. R. Larson and P. Frontiera. Spatial ranking methods for geographic information retrieval (gir) in digital libraries. In *Proceedings of the Research and Advanced Technology for Digital Libraries: 8th European Conference, ECDL 2004, Lecture Notes in Computer Science Series, LNCS 3232*, pages 45–57, 2004.

[9] J. Leidner. 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)*. Springer, 2006.

[10] M. Sanderson and J. Kohler. Analyzing geographic queries. In *Proceedings of Workshop on Geographic Information Retrieval SIGIR*, New York, NY, USA, 2004. ACM Press.

[11] A. Spink, B. J. Jansen, and H. C. Ozmultu. Use of query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policy*, 10(4):317–328, 2000.