Spatial Variation in Search Engine Queries

Lars Backstrom Jon Kleinberg
Dept. of Computer Science
Cornell University
Ithaca, NY 14853.
{lars,kleinber}@cs.cornell.edu

Ravi Kumar Jasmine Novak
Yahoo! Research
701 First Ave
Sunnyvale, CA 94089.
{ravikumar,jnovak}@yahoo-inc.com

ABSTRACT

Local aspects of Web search — associating Web content and queries with geography — is a topic of growing interest. However, the underlying question of how spatial variation is manifested in search queries is still not well understood. Here we develop a probabilistic framework for quantifying such spatial variation; on complete Yahoo! query logs, we find that our model is able to localize large classes of queries to within a few miles of their natural centers based only on the distribution of activity for the query. Our model provides not only an estimate of a query's geographic center, but also a measure of its spatial dispersion, indicating whether it has highly local interest or broader regional or national appeal. We also show how variations on our model can track geographically shifting topics over time, annotate a map with each location's "distinctive queries," and delineate the "spheres of influence" for competing queries in the same general domain.

Categories and Subject Descriptors: H.2.8 Database Management: Database Applications – Data Mining

General Terms: Measurement, Theory

Keywords: Web search, geolocation

1. INTRODUCTION

There has been growing interest in *local* aspects of Web search, associating geographic information with Web content [1, 2, 4, 11, 12, 13] and search engine queries [7, 19]. Such applications point to the outlines of a broad and largely open issue: understanding and quantifying the types of spatial variation that search queries can exhibit.

Many topics have a geographic focus of interest; sports teams, newspapers, schools, airlines, cell-phone companies, politicians, tourist attractions, cuisine, hobbies, weather events, and styles of music are just a few examples. This diversity of examples exposes a corresponding diversity in the way that spatial variation can be manifested: interest in a topic can be tightly concentrated at a particular location or spread diffusely over a broader region; it can have one geographic "center" or several; it can move over time. To characterize queries according to this continuum of possible geographic traits, we need a model and a source of data rich enough to be able to discover subtle distinctions in spatial properties.

Here we describe a framework for modeling the spatial variation in search queries, using data from search engine query logs, supplemented with geolocation techniques to assign accurate locations to a (large) subset of the IP addresses issuing the queries. In this way, we define the geographic focus of a topic by the locations of the people who search for it, rather than the locations of the servers hosting the Web content itself — in other words, according to the locus of user interest expressed by searches.

Our model is probabilistic, and discovers for each query a maximum-likelihood value for a *center* — the "hot spot" of interest — and a *dispersion* — the extent to which interest in the query is tightly concentrated around the query or more diffuse. Each of these two quantities has a natural meaning: the center provides a location, while the dispersion situates the query on a continuous spectrum ranging from local to national appeal. In this way, they function similarity to the concepts of *power* and *spread* considered by Ding et al. in the context of Web resources [4], but defined in a very different way here based on a probabilistic model over usage data.

Determining an accurate center and dispersion has potential applications in a number of contexts. It clearly has a role in focusing search-based marketing and advertising efforts by region, based on geographic distinctions among different queries. It can also be useful as a component of search engine rankings themselves, for refining or reordering query results based on geographic information. Finally, it can help in tools concerned with tracking news and current awareness, by distinguishing among different interest in news topics by locale.

Ultimately, then, the question is whether there is enough signal in raw query logs to produce values for the center and dispersion of a query with reasonable accuracy. To take a concrete example, will the query "Yankees" really localize to New York City — and will a much less searched-for term like "Royals" really localize to Kansas City — based purely on the latitudes and longitudes of queries, and despite the highly uneven distribution of locations from which these queries are being made, as well as the potentially obfuscating additional meanings of each? And will the range of dispersions for a topic such as baseball indeed distinguish teams (such as the Yankees) with a national following from those whose following is mainly local?

Supported in part by NSF grants CCF-0325453, CNS-0403340, BCS-0537606, and IIS-0705774, and by the John D. and Catherine T. MacArthur Foundation.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

Basic properties of the model. We find, in fact, that a natural generative model for query production based on geography can produce highly accurate centers and dispersions for a broad range of queries, even reflecting geographic distinctions that are subtle, short-range, and sometimes temporally varying. The model is based on a decomposition of the surface of the earth into small grid cells; we assume that for each grid cell x, there is a probability p_x that a random search from this cell will be equal to the query under consideration. In the basic form of the model, we then posit that each p_x should decrease with the distance of x from a "hot-spot" cell z; the cell z is then the center, and the rate of decrease of p_x as a function of its distance from z determines the dispersion. We develop an efficient algorithm to compute the maximum-likelihood values for the center and dispersion, capable of scaling to handle complete query logs.

We describe a range of tests indicating that our method is effective at accurately localizing queries when there is a natural "ground truth" value for the center: to mention just a few examples, the names of almost all professional baseball teams are localized to their home cities, the names of almost all U.S. Senators are localized to their home states, and the names of national parks are localized to their physical locations. Indeed, we show through a comparative evaluation that for localization accuracy on these types of queries, our probabilistic model significantly outperforms simpler geometric techniques, as well as state-of-the-art commercial software for query localization. This evaluation is based on the computed center; we also show that the dispersion follows a natural pattern for these classes of queries, ranging from queries that have broad appeal to those that are tightly focused geographically.

With a scalable method for determining centers and dispersions for queries, it becomes possible to assess the spatial variation of large collections of queries. We find that among the most 1000 frequent queries in the full log, there is surprising geographic concentration in a number of them. And with a simpler heuristic version of our probabilistic model, we perform a much more massive study — analyzing the 100,000 most frequent queries, and tagging each location on the surface of the earth with the queries that have the most significant concentration in that location with respect to the model. The resulting map of significant queries reveals trends that range from the global — such as the kinds of on-line social-networking sites favored in different parts of the world — to the very local — with striking geographic specificity as names of community colleges, high schools, and local newspapers vary between locations as little as ten miles apart.

Further extensions to the model. We can extend the model to handle other forms of spatial variation as well. To begin with, a number of queries have the property that their geographic focus noticeably shifts over time — for example, seasonal effects over slow time scales, and news events over more rapid time scales. We show how to extend the model to allow the center and dispersion to vary with time — incorporating an additional probabilistic component that favors relatively "smooth" temporal motion in these quantities. Tracking these quantities over time can then be done efficiently with a shortest-path computation in a graph derived from the spatial arrangement of the grid cells from which the queries are issued. Here too the results can be quite striking: for the query "Hurricane Dean," for example, one can see the query center moving day-by-day in a way that tracks the storm center's westward movement into Mexico — and with a dispersion that starts out very concentrated but expands widely as the hurricane approaches land and begins appearing in the national news.

We describe other extensions as well, including a method for modeling spatial variation with multiple centers, and a method for comparing the geographic concentration of multiple queries to determine the approximate "sphere of influence" of each.

Organization of the paper. The remainder of the paper is organized as follows. In Section 2 we describe the model and our algorithms for estimating the maximum-likelihood values for the center and dispersion. In Section 3 we provide examples of the method's performance on queries with a natural geographic focus, including a description of its evaluation relative to other approaches. We describe extensions to the model, including temporal variation and the problem of identifying multiple centers, in Section 4. In Section 5, we discuss methods to simultaneously locate many queries on a shared map, thereby illustrating regional variation at the level of large numbers of topics. Finally, we discuss related work in Section 6 and conclude in Section 7.

2. MODELING SPATIAL VARIATION

2.1 Methodology

Our data consists of Yahoo! search query logs. For each query, these logs give us both the query string and an approximation of the latitude and longitude from which the query originated (based on IP address). In order to reduce the effect of factors like network address translation, which allows many users to appear to come from the same IP, we first remove all queries that originate from IP addresses with particularly high search frequencies. The removed IPs account for less than 1% of all the queries. Furthermore, we focused only on North American queries, discarding any further west than 135° W, or further east than 60° W, as well as those south of 15° N or north of 60° N.

To reduce the variability introduced by people's usage habits, we further process the data so that no IP address is considered to issue more than a single query during the time window we consider. Thus, with respect to a particular query, we consider how many distinct IP addresses from each geographic location issued at least one query during a time window, and how many of those IP addresses issued the particular query under consideration at least once. (For ease of discussion, we will often refer to "users" issuing queries, although for the above-mentioned reasons, IP addresses are in fact our basic unit of analysis.)

2.2 Model

For a given query, we posit a simple generative model to explain the differential frequency with which the query appears across geographic regions. In this model, each query has a geographic center represented by a single point. This center corresponds to the point at which the query should occur most frequently, with frequency then falling off in distance from the center.

In addition to its central point, each query in this model has two other parameters associated with it: a constant, C,

giving the frequency at the query's center, and an exponent α determining how quickly the frequency falls off as one gets further away from the center. The model posits that when a random user at distance d from the query's center issues a query, it is equal to the query under consideration with probability $Cd^{-\alpha}$. For queries that are very local, such as the name of a small city, we expect a large value of α, indicating that people rarely search for that query unless they are quite close to the query's center. On the other hand, a query with essentially no geographic aspects or regional bias might have α very close to zero, indicating that the frequency is essentially uniform over geography. The polynomial functional form is employed here based on initial exploratory analysis of the data, and for tractability of the model; it is also sufficient for our purposes, as it is capable of ranging smoothly (as α varies) from a uniform distribution of user interest to distributions that are sharply peaked.

2.3 Algorithm

With this model in mind, we can focus on a particular query q and take a maximum-likelihood approach to discovering the parameters for q from the data. For a particular C, and α, we can compute the probability that the true query logs came from this model. For each log entry consisting of a user issuing a query, we compute $p = Cd^{-\alpha}$, where d is that person's distance from the query center. If that person issues query q, we multiply the overall probability of the data by p, and by $1-p$ otherwise. (To avoid underflow, we work by adding logarithms of probabilities, rather than actually multiplying.)

We now have a way to evaluate a particular set of parameters on a particular query; but it would be far too expensive to consider a wide range of parameters using a brute force method of simply trying many of them. Instead, we first observe that moving the center a little bit tends not to affect the overall log-odds very much. Thus, our search algorithm starts by trying centers on a coarse mesh. It then selects the best one, and uses a finer grain mesh on the region around that best one. This can be repeated until the desired accuracy is reached. In practice we find that starting with points at every two degrees of latitude and longitude, and ending with points at tenths of degrees works well.

Once we have selected a center, we now have to optimize the other two parameters. Our approach is based on Theorem 1, below, which establishes that the log-likelihood as a function of C and α is unimodal; we therefore develop techniques based on optimization of unimodal multivariate functions to find the optimal parameters. For scalability, we bucket all the queries by their distance from the center, enabling us to evaluate a particular choice of C and α very quickly.

To establish the necessary unimodality property, we proceed as follows. Let S be the set of log entries for query q (indexed by users who issued this q), and let d_i be the distance of a user i from the query's center. Then $f(C,\alpha) = \sum_{i\in S} \log Cd_i^{-\alpha} + \sum_{i\notin S} \log(1 - Cd_i^{-\alpha})$ is the log of the probability for parameters C and α.

LEMMA 1. *$f(C,\alpha)$ is concave in C.*

LEMMA 2. *$f(C,\alpha)$ is concave in α.*

THEOREM 1. *$f(C,\alpha)$ has exactly one local maximum over its parameter space.*

PROOF. For sake of a contradiction, imagine that there were two choices of parameters, (C_1,α_1) and (C_2,α_2), each of which was a local maximum. Unless $\alpha_1 = \alpha_2$ (in which case they can't both be maxima by Lemma 2), there is some d_0 such that $C_1 d_0^{-\alpha_1} = C_2 d_0^{-\alpha_2}$.

We now consider all functions $Cd^{-\alpha}$ that cross the point $(d_0, C_1 d_0^{-\alpha_1} = C_2 d_0^{-\alpha_2})$. Each is fully determined by the value of the exponent α, having $C(\alpha) = C_1 d_0^{\alpha-\alpha_1}$. We now consider the likelihood $f(C(\alpha),\alpha)$. By simply taking the second derivative of $f(C(\alpha),\alpha)$ with respect to α, we find that it is always negative, contradicting our original assumption of two maxima.

$$\begin{aligned} f(C(\alpha),\alpha) &= \sum_{i\in S} \log C_1 d_0^{\alpha-\alpha_1} d_i^{-\alpha} \\ &\quad + \sum_{i\notin S} \log 1 - C_1 d_0^{\alpha-\alpha_1} d_i^{-\alpha} \\ f''(C(\alpha),\alpha) &= \sum_{i\notin S} \frac{-(C_1 d_0^{-\alpha}(\frac{d_0}{d_i})^{\alpha})^2}{(1 - C_1 d_0^{-\alpha_1}(\frac{d_0}{d})^{\alpha})^2} \\ &\quad + \frac{-C_1 d_0^{-\alpha}(\frac{d_0}{d})^{\alpha} \log^2 \frac{d_0}{d}}{1 - C_1 d_0^{-\alpha_1}(\frac{d_0}{d})^{\alpha}} \end{aligned}$$

$C_1 d_0^{-\alpha}(\frac{d_0}{d_i})^{\alpha}$ is a probability in $[0,1]$, which makes the first term at most 0. The second term is also non-positive for the same reason. The only way that the expression can evaluate to 0 is if C_1 is 0. However, in this case the log-odds will have a $\log 0$ for each $i \in S$. □

In practice, our numerical methods converge quickly to the single maximum value. Our (unoptimized) implementation runs in under 30 seconds on a modern machine.

3. ASSESSING THE MODEL

We now discuss some of the results that can obtained from this model, running on the complete set of query logs. These results can be organized by the kinds of queries for which spatial variation stands out: on the one hand, there are classes of queries that by their nature have a geographic focus (for example, names of schools, newspapers, or sports teams); and on the other hand, there are queries whose geographic content is a priori less apparent. Queries in this latter category can be found effectively by *enumeration* — that is, applying the model to all the frequent queries in the log and identifying those with large exponents, indicating geographic concentration.

We begin with the simplest kinds of examples — those queries for which there is a natural geographic focus. For analysis and evaluation, we consider several classes of such queries here: names of the most populous U.S. cities, names of certain universities, names of high-circulation U.S. newspapers, names of all Major League Baseball teams, names of all U.S. national parks, names of all U.S. Senators, as well as certain companies such as banks, airlines, and cell-phone carriers that have a regional focus. We will refer to these categories as our *basic classes*. Each query in one of the basic classes has an a priori natural geographic center, though the center may be conceptually a "point" (e.g. in the case of a national park or the home city of a sports team) or a broader region (e.g. in the case of a state represented by a Senator or a region served by a cell-phone carrier). In all cases, the model identifies these natural centers with high accuracy, as our more detailed evaluation below demonstrates.

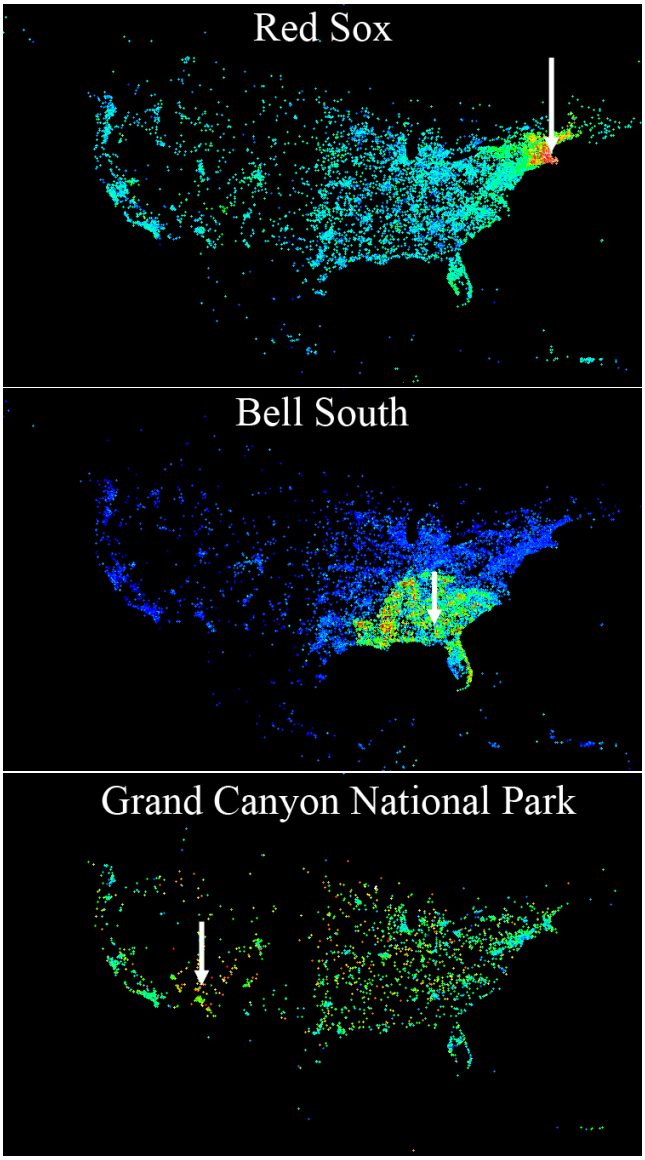

Figure 1: Geolocation of queries "Red Sox," "Grand Canyon National Park," and "Bell South". (The capitalization of queries is reduced to a canonical form in our experiments.) These figures are drawn as *heat maps*, with the color spectrum indicating the query intensity per grid cell (and hence there is value in viewing these, as well as later figures, on a color display or color print-out). The arrows indicate the centers computed using our model.

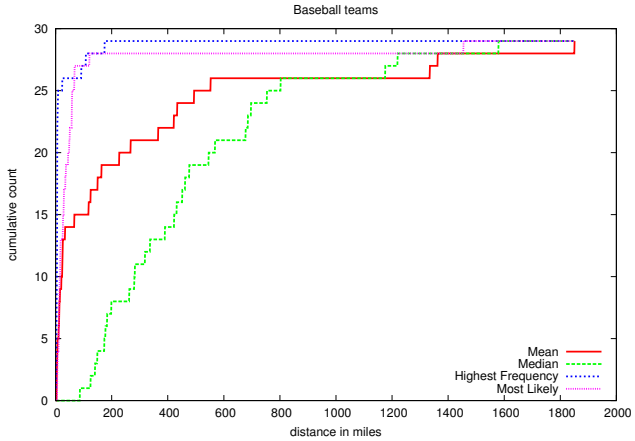

Figure 2: Performance of algorithms on baseball team queries.

By way of example, Figure 1 shows the query distribution for three queries in these classes: "Red Sox," "Bell South," "Grand Canyon National Park." The first two queries are clearly easy to localize: We see that the query "Red Sox" has a conceptual hot-spot in New England, while the hot-spot for "Bell South" closely tracks the boundaries of the states that this company primarily serves. "Grand Canyon National Park" is one instance of a class of examples that is more subtle for several reasons. First, there is relatively little data on this query, even at the scale of complete logs; and as the image makes clear, the location of the park itself is found although the hot-spot is not immediately apparent visually. But beyond this, it should not have been clear in advance that the location of the park should even be a natural "center" for this query: the fact that it emerges as the center suggests that there is a hot-spot in query activity coming from people who are already at the physical location, rather than people from nearby population centers planning future vacations. We find this in general with geographic destinations in less-populated areas — despite the fact that a large fraction of visitors come from some distance away, the center is generally at the location itself.

In addition to the basic classes of queries, applying the model to all frequent queries turns up geographic information about queries that have no a priori "home." As one illustration, the model uncovers the oft-discussed observation that different social-networking sites have particular concentration in different regions. For example, despite the enormous penetration of Facebook, a hot-spot is still uncovered in Southern Ontario — a fact about the Facebook user demographics that the company has remarked on in its own publicity. Similar observations are easily found for other social-networking sites as well.

We now turn from these examples to a more systematic evaluation of the model's effectiveness at localizing centers.

3.1 Evaluation

We begin by formulating a framework within which to evaluate the model. The premise is that for a basic class in which the a priori natural centers have precise coordinates (e.g. the home city of a sports team), we define a_q to be

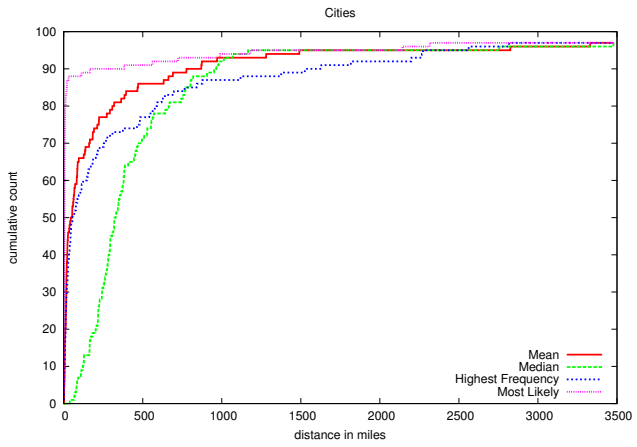

Figure 3: Performance of algorithms on high-population U.S. cities.

Accuracy	Mean	Median	Local density	Our model
	51	12	80	90

Table 1: Accuracy of algorithms for localizing senators inside their respective states.

this natural center, and we define b_q to be the center computed by the model. Evaluating the distance between them, $d(a_q, b_q)$, gives an indication of how accurate the model's localization is.

To compare our model with simpler baselines, we also determine the distance from a_q to centers computed by other means; specifically:

- n_q, the weighted center of gravity of all instances of query q;
- m_q, the point at the median latitude and median longitude of all instances of query q; and
- ℓ_q, the point with the highest local density of instances of q — that is, with the lowest likelihood relative to the overall base-rate for query q.

Note that the first two methods are geometric, while the third is probabilistic but much simpler than our model. In Figure 2, we compare all these methods at localizing all Major League Baseball team names to their home cities — in particular, depicting the cumulative distribution of distances to a_q over all teams. We see that our model's center b_q and the optimum log-odds center ℓ_q greatly outperform the geometric methods, with both b_q and ℓ_q localizing almost all team names to within 60 miles of their respective home cities. This is in a sense somewhat surprising, given the multiple meanings (other than baseball-related ones) that many baseball team names have. Also, recall that our model, in addition to producing the center b_q, is also estimating dispersion in the form of the exponent α, whereas one gets only a center from the baseline ℓ_q. Due in part to the need to fit the full query distribution via this exponent, our model is less exact in its localization (compared to ℓ_q) for distances significantly under 60 miles.

Newspaper	α
The Wall Street Journal	0.111327
USA Today	0.263173
The New York Times	0.304889
New York Post	0.459145
The Daily News	0.601810
Washington Post	0.719161
Los Angeles Times	0.782538
The Star Ledger	0.998462
Detroit Freepress	1.068055
San Francisco Chronicle	1.091030
Chicago Tribue	1.102554
Philadelphia Inquirer	1.140618
Chicago Sun Times	1.165482
The Boston Globe	1.171179
The Arizona Republic	1.284957
Dallas Morning News	1.286526
Houston Chronicle	1.289576
Star Tribune	1.337356

Table 2: Estimation of exponents α for high-circulation U.S. newspapers.

City	α
New York	0.396527
Chicago	0.528589
Phoenix	0.551841
Dallas	0.588299
Houston	0.608562
Los Angeles	0.615746
San Antonio	0.763223
Philadelphia	0.783850
Detroit	0.786158
San Jose	0.850962

Table 3: Estimation of exponents α for the 10 most populous U.S. cities.

We perform an analogous evaluation for the names of all U.S. Senators, in which the natural center is no longer a point but a region (namely, their home state.) We evaluate, for our model and the three baseline methods, how many of the 100 Senators are localized to a center within the state they represent. Table 1 shows these results; our model outperforms all the baselines, with mean and median performing particularly poorly. (Certain queries in this class illustrate additional qualitative contrasts between the models; for example, our method localizes the query "Lisa Murkowski" to her home state of Alaska, while the three baseline methods all put the center in the continental U.S.)

It is also natural to evaluate our model against state-of-the-art commercial services, which employ features other than usage, for inferring whether a query is "local." In particular, we use the service WhereOnEarth, a leading exemplar of this type of application. Our first finding is that query log data reveals strong spatial variation for much broader ranges of queries than services such as WhereOnEarth pick up. As a result, direct comparison is a bit difficult, since many of even our basic classes above are not considered localizable by these services. For example, WhereOnEarth does not consider the names of any U.S. Senators

School	α
Harvard	0.386832
Caltech	0.423631
Columbia	0.441880
MIT	0.457628
Princeton	0.497590
Yale	0.514267
Stanford	0.627069
U. Penn	0.729556
Duke	0.741114
U. Chicago	1.097012

Table 4: Estimation of exponents α for the 10 highest-ranked U.S. universities according to U.S. News & World Report.

or Major League Baseball teams to be local queries for which a center can be inferred, despite the fact that our model finds correct centers for almost all from usage data.

For other basic classes, such as high-population U.S. cities, major U.S. universities, and U.S. national parks, WhereOnEarth can determine exact values for almost all by table look-up. Our model does well in all these cases too, despite having no comparable access to hard-coded data; and it outperforms the three baselines n_q, m_q, ℓ_q in all these cases. Figure 3 shows the performance for U.S. cities; note that our model significantly outperforms the other three approaches, and the the center of gravity n_q outperforms the simple probabilistic baseline ℓ_q in this case.

3.2 Exponents and Dispersion

Thus far we have been considering the centers computed by the model, but there is additional value in the exponent as well. This provides the measure of *dispersion* mentioned in the introduction; a large exponent indicates rapid decay away from the center, and hence strong geographic concentration, while a smaller exponent indicates interest over a broader region.

Thus, particularly when we compare exponents for queries from the same basic class, we can place items on a spectrum ranging from local appeal to more national appeal. For example, Tables 2-4 show the exponents for the 10 most populous U.S. cities, for the 10 highest-ranked U.S. universities according to U.S. News & World Report, and for a collection of high-circulation U.S. newspapers.

Ranking each of these lists by exponent places them on a spectrum from local to national appeal. For example, the Wall Street Journal and USA Today are the two newspapers with the lowest exponents, indicating national interest, with the New York Times close behind. Other high-circulation newspapers are regional in their appeal, with exponents that are much higher. We also see that the spatial variation in queries for city names does not directly correspond to the populations of the cities; for example, Los Angeles has a comparatively large exponent, while the second-lowest exponent among large U.S. cities belongs to one that is not in the top 10: Las Vegas, with an exponent of .482. While we omit the list of national parks due to space limitations, there is significant variation in exponents here too, with Death Valley, the Grand Canyon, and the Everglades having the lowest values (and hence the most national reach in queries).

Figure 4: The path of Hurricane Dean's storm center, moving west through the Caribbean, alongside the smoothed path of query centers for "Hurricane Dean."

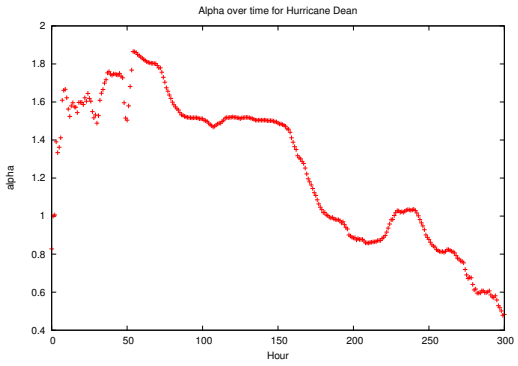

Figure 5: Change in the exponent for "Hurricane Dean" by hour, as interest in the topic shifted from local to national.

4. EXTENSIONS: TEMPORAL VARIATION AND MULTIPLE CENTERS

4.1 Temporal Aspects

While most localizable queries maintain relatively stable centers and dispersions over time, it is easy to find queries which vary in both of these dimensions. A local news story might start with limited dispersion as only people in the region are aware of it. If the story then gains national attention, the center may stay the same, but the exponent α can decrease as query traffic increases from farther away.

In other cases, the center may move as well, and a good source of examples for this comes from large weather phenomena. For instance, as a hurricane moves over time, the people who are next in its path at any given moment tend to search for it with the highest intensity, and thus we might expect the query center to roughly track the storm's center.

We can observe this in the query-log data by considering a sequence of 24-hour time slices, at offsets of one hour from

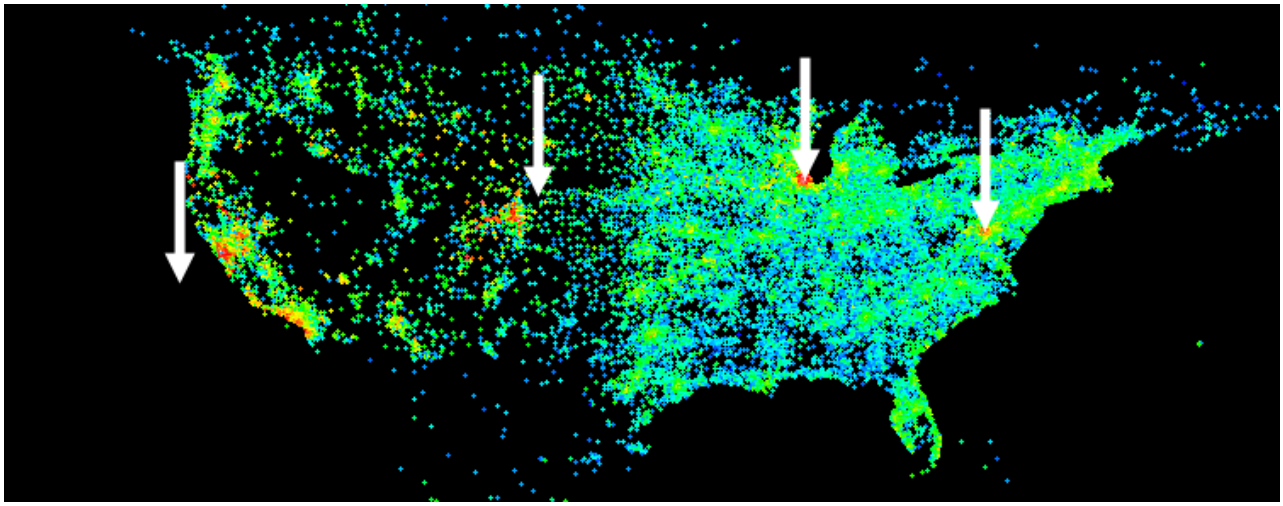

Figure 6: Multiple centers for the query "United Airlines."

each other (i.e. considering the 24 hours starting at midnight, then the 24 hours starting at 1 AM, and so forth). We can then calculate the center for each of these 24 hour periods. By using a sliding window in this way, we are able to fit our parameters to more data, making them more reliable. Employing 24-hour periods has the useful effect of mitigating diurnal variation, since all times of day are represented in each period.

In the case of a major recent hurricane, Hurricane Dean, we observe a clear westerly trend, with centers starting far out in the Caribbean, and tending to move westward towards Southeast Texas as the hurricane does so. There is also a clear trend towards decreasing α as the hurricane gains national attention.

For a number of reasons, however, the sequence of query centers computed for each time slice in isolation does not move very smoothly. In large part, this is because the amount of query-log data for a single hurricane, even a major one, is relatively small, especially before it approaches mainland North America. Thus, we improve the results using a more complex method: we couple the computation of centers across the different time-slices by a natural and efficient algorithm, obtaining a smooth query-center path that tracks the true path of the storm's center with surprising fidelity (Figure 4).

The coupled computation works as follows. For each 24-hour period, and each latitude and longitude, we compute the cost of the center at that location as the negative log-probability for the optimal choice of C and α. In order to account for difference in total volume between different 24-hour periods, we normalize by dividing the cost at a point A by the the minimum cost over all possible centers for that same 24-hour period. Thus, we have a normalized cost for every coordinate for every time window. We now define a cost for moving the center from point A to point B as $\gamma|A - B|^2$. Thus paths which jump around a lot are penalized, while smooth paths which move at a constant rate have relatively low cost. The goal now is to find a sequence of centers for the sequence of 24-hour windows (each offset by one hour from the previous one) that minimizes the sum of the costs from the placement of the centers and the costs for the movement from each center to the next.

It is easy to find the lowest-cost sequence of centers for various constants γ using dynamic programming. Once we have done this, we can examine the smoothed paths taken by the center of the query for different γ, one of which is shown in Figure 4. We see a striking similarity between the actual path of the storm's center and the smoothed path taken by the computed center.

In addition to tracking the storm's motion through query logs, we can also watch how the query's dispersion changes over time (Figure 5). By examining the optimal choices of α over time for the smoothed path, we can see that the hurricane started out as an event of very local interest, with its center near the Lesser Antilles. As the storm moved west and intensified, more and more people started taking notice, and it eventually became a major news story, as it was the one of the most intense hurricanes to ever reach land.

4.2 Multiple Centers

While the simple, single-center model describes many queries fairly well, some queries are clearly better modeled with multiple centers. For example, major airlines typically have three or four hub cities, and it is clear from the query-log data that the regions around each of these cities have high query frequency for their respective airlines.

To model this sort of spatial variation we extend our generative model by placing multiple centers, each with its own C and α parameters. For each point, we use the probability given by the center which yields the highest probability to that point. Thus, we can easily calculate the log-odds for a choice of multiple centers with different parameters.

This, however, makes the maximum-likelihood optimization problem much harder, and so we use a heuristic based on the K-means clustering algorithm. We start with K centers placed at random, for some constant K. We then optimize each of these centers, treating each one as if it were the only center being used (our previous algorithm). After we do this for every center, we look at each geographic point and determine which of the K centers gives that point the highest probability, according to the polynomial-decay prob-

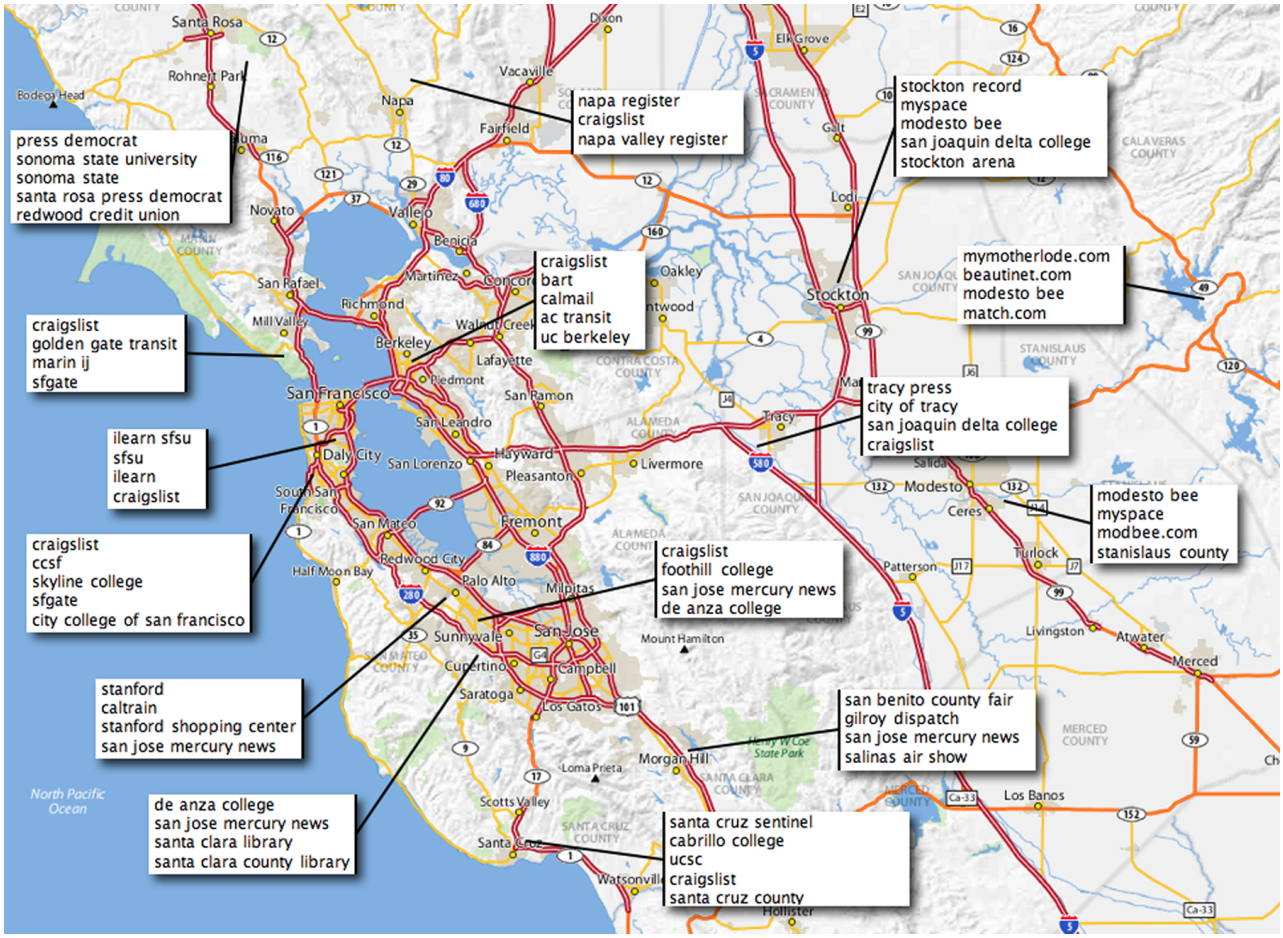

Figure 7: Distinctive queries for locations in the San Francisco Bay Area.

ability function. We now say that each point is associated with the center giving it the highest probability. We then reoptimize each center independently, but considering only those points that were associated with it during the previous iteration. This process is repeated a number of times until it converges. The algorithm is sensitive to the starting locations of the centers, but by running it many times and choosing the best outcome, we achieve good results.

As an illustration, Figure 6 shows the results of this algorithm for the query "United Airlines." United's largest hub is in Chicago, and it also has hubs in Denver, Washington DC, San Francisco, and Los Angeles. The algorithm places centers in Chicago, Washington, and near Denver. It places a fourth center off the coast of California, which has the effect of hitting both San Francisco and Los Angeles somewhat equally. (Note that it is a natural consequence of the probabilistic model, even with one center, that there may be low query density at the exact point corresponding to the center itself.) We see similar results for other airlines.

The multiple-center model is also useful for queries with two distinct geographic meanings. For example, on the query "Washington," with two centers, the algorithm places one center in DC and the other in Washington state. For the query "Cardinals," the algorithm places one center in St. Louis (the home of the baseball team) and the other in Arizona (the home of the football team).

5. ENUMERATING MULTIPLE QUERIES ON A SHARED MAP

5.1 Distinctive Queries for all Locations

Given a way to assess the spatial variation of individual queries, we can enumerate all the queries in the log and — for each location on earth — find the queries that are the most "unusual" or "distinctive" for that location. In this section we describe the results of such a computation, leading to an annotated world map of which the image in Figure 7 is a tiny portion.

We define locations as before, using tenth-of-a-degree grid cells. (For reference, such a cell has a side length of less than 10 miles .) We define "distinctiveness" at a given cell x using a variant of the probabilistic model from Section 2. For each query q, let p denote the fraction of all entries in the log corresponding to users issuing q. Let t_x be the total number of log entries from x, and let s_x be the number of log entries from x corresponding to users issuing q. Assuming a simple independence-based model, the probability of this observed data given the background probability p is

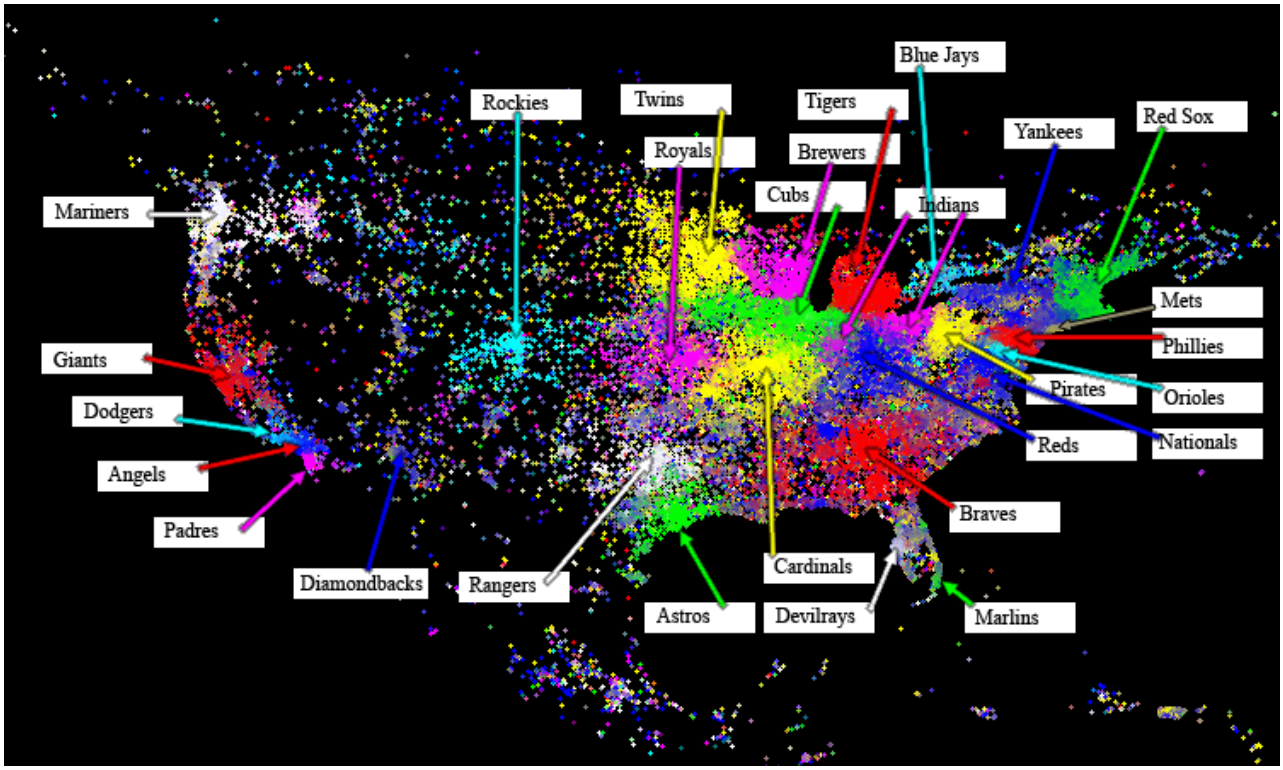

Figure 8: Spheres of influence of baseball teams.

$\binom{t_x}{s_x} p^{s_x}(1-p)^{t_x - s_x}$. We choose the queries q for which this probability is lowest to serve as the most distinctive queries for location x — they are the ones that deviate the most significantly at x from their global background rate.

We perform the computation on queries issued during a week. To have sufficient sample size, we only consider locations with at least 5000 queries during the week. This yields 3154 locations worldwide with 2643 locations in the continental US; for each, we find the most distinctive queries.

We illustrate some of the results of this in Figure 7; for ease of presentation, we only display the San Francisco Bay Area, and only a subset of the locations there. While one might have expected a region this small to have roughly the same distinctive queries in each cell (things like "San Francisco," "San Jose," and so forth), in fact we see significant and meaningful differences between locations that are only a few miles apart (see for example, the distinction between queries being issued in Palo Alto and Sunnyvale.)

5.2 Spheres of Influence

If we consider many of the basic classes of queries from Section 3, they represent entities that are at least implicitly in competition. (Consider, for example, baseball teams, universities, or newspapers.) By representing all the queries from a single class on a shared map, we can try understanding and visualizing their respective "spheres of influence" — the regions in which each is dominant. In Figure 8, we depict such regions of influence for all Major League Baseball teams on a map of the U.S.: thus the team names are the queries, and each grid cell is colored according to the distribution of queries for baseball teams issued from that cell.

We now discuss the methodology underlying images such as this, then make some observations about this image itself.

To define regions of influence, we need a way to represent the queries that are dominant at different grid cells. The simplest way would be just to see which query produces the most log entries for each cell, and color the cell with this most abundant query. However, due to the sparsity of the data, this produces a fairly noisy representation, with adjacent cells generally differing in color, and it doesn't capture the difference between minor dominance and strong dominance in a given cell.

A better way to represent the regions of influence, then, is to imagine that each query from a particular cell acts as a vote for that query. The pixels are then colored by blending colors according to the voting. Done in a straightforward way, this now has the opposite problem — most regions are too blended in the image. To strike a balance, we produce the image in Figure 8 by counting a query with N log entries in a cell as having N^c votes for a small constant $c > 1$. Varying c lets us bring out the dominant color, but still allowing blending when there are close competitors.

The first observation about Figure 8, of course, is that despite the highly uneven national volumes of queries for different teams, we see a very clean geographic breakdown of who follows which teams. It is also somewhat surprising to see the extent to which team dominance breaks down along state lines — a number of the colored regions follow state boundaries very clearly. For instance, in Michigan, across the lake from Chicago but far from Detroit, it is the Tigers, not the Cubs who have the largest following. It is also interesting to note the regions which do not have a clear-

cut winner. For instance, in the Carolinas and Louisiana there are many queries for baseball teams, but there is no one team that stands out above the rest.

6. RELATED WORK

Prior work related to this paper can be grouped roughly into four high-level areas: geolocation of Web content, geolocation of search queries, efficient query processing with geographic information, and spatial hot-spot models.

There is a significant line of work on inferring geographic locations for Web pages and other on-line content. Buyukkokten et al. [2] use geographic entities and network IP address information to geolocate Web pages. McCurley [12] proposed a spatial browsing of Web data; his approach was to use a rich set of geographic features, including telephone numbers, to infer geospatial context. Amitay et al. [1] describe a system, Web-a-Where, that assigns each page a geographic focus. Further work includes [11, 14, 18]. Ding et al. [4] introduced the idea of the *power* and *spread* of a Web page, which are analogous to the center and dispersion parameters in our model. However, their approach is not based on a probabilistic model or corresponding optimization criterion for the parameters. Some applications of Web content geolocation include mining spatio-temporal themes [13] and geographically focused collaborative crawling [6].

In contrast, much less work has been done on geolocating Web queries, our focus here. Gravano et al. [7] performed an early investigation of this issue, using machine learning techniques to classify search queries as either local or global. Closer to our work, the paper of Wang et al. [19] searches for the "dominant location" of a query in the context of a system for exploiting query localization to improve retrieval performance. They include power and spread among their features, and again the approach is quite different, and does not include a specific model for spatial variation.

Query processing with geographic constraints is an active research area. Much work here constructs and processes spatial representations to enable efficient query processing. Some recent work in this area is by Chen et al. [3], Tezuka et al. [17], and Schockaert and De Cock [16].

Finally, spatial hot-spot models have been extensively studied in statistics. For a good account of the work in this area, ranging from the ad hoc to the model-based, see the discussion in Neill et al. [15], as well as Kulldorff [9] and the book edited by Lawson and Denison [10]. There are also connections to temporal hot-spot models, which use one-dimensional analogues of these computations [5, 8].

7. CONCLUSIONS

We have seen that large-scale query-log data contains enough information to build effective models of spatial variation. In addition to finding centers for queries with an accuracy that outperforms competing baselines, and extracting geographic information for a broader range of queries than is accessible to commercial systems, our models also form the basis for algorithms that incorporate temporal processes, as well as methods to analyze variation for many queries simultaneously at a global level.

There are a number of directions in which this work could be extended. It would be interesting to consider our analysis of simultaneous spatial and temporal variation (as in Section 4.1) in the context of further probabilistic models, potentially exploring connections with the methodology in [13]. It would also be interesting to incorporate more complex models of user behavior into a framework that explicitly took spatial variation into account, potentially resulting in more accurate kinds of localization for broader classes of queries. Ultimately, as the local applications of search continue to broaden, we can expect to see questions of this sort arise increasingly from the rich interaction between Web information, user interests, and the geographic and spatial frames of reference in which they are embedded.

8. REFERENCES

[1] E. Amitay, N. Har'El, R. Sivan, A. Soffer. Web-a-where: Geotagging Web content. In *SIGIR*, pages 273–280, 2004.

[2] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting geographical location information of Web pages. In *WebDB (Informal Proceedings)*, pages 91–96, 1999.

[3] Y.-Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic Web search engines. In *SIGMOD*, pages 277–288, 2006.

[4] J. Ding, L. Gravano, N. Shivakumar. Computing geographical scopes of Web resources. In *VLDB*, pages 545–556, 2000.

[5] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, A. Tomkins. Visualizing tags over time. In *WWW*, pages 193–202, 2006.

[6] W. Gao, H. C. Lee, Y. Miao. Geographically focused collaborative crawling. In *WWW*, pages 287–296, 2006.

[7] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing Web queries according to geographical locality. In *CIKM*, pages 325–333, 2003.

[8] J. Kleinberg. Temporal dynamics of online information streams. In M. Garofalakis, J. Gehrke, R. Rastogi (eds.) *Data Stream Management*. Springer, 2008.

[9] M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6):1481–1496, 1997.

[10] B. Lawson and D. G. T. Denison, editors. *Spatial Cluster Modelling*. Chapman & Hall, 2002.

[11] B. Martins, M. S. Chaves, M. J. Silva. Assigning geographical scopes to Web pages. In *ECIR*, pages 564–567, 2005.

[12] K. S. McCurley. Geospatial mapping and navigation of the Web. In *WWW*, pages 221–229, 2001.

[13] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, pages 533–542, 2006.

[14] Y. Morimoto, M. Aono, M. E. Houle, and K. S. McCurley. Extracting spatial knowledge from the Web. In *SAINT*, pages 326–333, 2003.

[15] D. B. Neill, A. W. Moore, and G. F. Cooper. A Bayesian spatial scan statistic. In *NIPS*, 2005.

[16] S. Schockaert and M. D. Cock. Neighborhood restrictions in geographic IR. In *SIGIR*, pages 167–174, 2007.

[17] T. Tezuka, T. Kurashima, and K. Tanaka. Toward tighter integration of Web search with a geographic information system. In *WWW*, pages 277–286, 2006.

[18] C. Wang, X. Xie, L. Wang, Y. Lu, W.-Y. Ma. Detecting geographic locations from Web resources. In *GIR*, pages 17–24, 2005.

[19] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, and Y. Li. Detecting dominant locations from search queries. In *SIGIR*, pages 424–431, 2005.