2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Towards Click-based Models of Geographic Interests in Web Search

Ziming Zhuang, Cliff Brunk
Yahoo! Inc.
Sunnyvale, CA 94089, USA
{ziming,brunk}@yahoo-inc.com

Prasenjit Mitra, C. Lee Giles
Pennsylvania State University
University Park, PA 16802, USA
{pmitra,giles}@ist.psu.edu

Abstract

With the recent surge in the volume of search queries that explicitly or implicitly express users' geographical interests, to accurately infer users' locality preference becomes an increasingly important yet challenging issue. We study two click-based models of the distribution of such geographical interests by mining the user click stream data in the search engine logs, addressing three important issues in spatial Web search. First, search queries and documents can be classified by the models according to their spatial specificity. Second, the geographic center(s) of interests for queries and documents can be inferred. Finally, the model can be applied to generate relevance features for search ranking. We evaluated our proposals on a large dataset with about 10,000 unique queries sampled from the Yahoo! Search query logs, and about 450 million user clicks on 1.4 million unique Web pages over a six-months period. We report about 90% accuracy and about 3% false positive rate in identifying search queries with or without specific geographical interests, as well as statistically significant improvement in relevance ranking over a strong baseline.

1 Introduction

The volume of *local* search - search queries with a *local flavor* - has been constantly increasing over the past few years. With the advance in the offerings by most commercial search engines to conduct such local search, there has been a growing interests in associating search with an interest in a specific geographical locality, and applying this knowledge to enhance the search experience for users.

The problem of inferring users' geographical interests is non-trivial and challenging. While some queries (e.g. *"auto repair san francisco"*) contain an explicitly location qualifier (*"san francisco"*), other queries may come with spatial ambiguities. For example, a query *"bay bridge"* may refer to the San Francisco-Oakland Bay Bridge in California, the Chesapeake Bay Bridge in Maryland, or the Escambia Bay Bridge in Florida. Similarly, a query *"chicago pizza"* may refer to *chicago style pizza* or *pizza stores in Chicago*, and for the same reason, the fact that a Web page contains both the terms *"chicago"* and *"pizza* doesn't help resolve the ambiguity, either.

In the above and many other similar scenarios, it is clear that the syntactic information alone derived from the queries is not a sufficiently good indicator for users' geographic interests. This observation motivates us to exploit other heuristics such as the user clicks to model their locality preferences.

In contrast to a large body of existing work on modeling users' geographical interests based on syntactic information, we study two click-based models for inferring the geographical interests of the search engine users. We exploit the geographical distributions of user clicks, and empirically demonstrate that such distributions shed lights on the geographical variations in users' interests for a given query or Web page. Specifically, we address the following three problems:

- Given a large number of search queries or Web pages, how do we model their geographical specificity?
- For queries and Web pages that are identified as having a *local flavor*, how do we detect and disambiguate the underlying geographical location(s) of interests?
- How do we improve search relevance ranking given a model of the distribution of the *geographic interests*?

Organization The rest of the paper is organized as follows. We briefly review the related literature in Section 2, and introduce the dataset in Section 3. We visualize the correlation between the geographical distributions of user clicks and their locality preference in Section 4, and study two click-based models that capture such correlation in Section 5. We discuss two applications of these models: classifying queries according to their geo-sensitivity (Section 6) and improving search relevance ranking (Section 7). In Section 8 we present empirical evaluations. We conclude our paper with directions for future work in Section 9.

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.365

2 Related Studies

Prior studies investigated the language of geographical search queries, including the characteristics of the search terms in geographical queries [14, 15], and geographical query expansion and rewriting [5, 6, 16, 21]. Another group of work studied geo-tagging Web resources [2, 4, 13, 18, 20] and providing a visual model of the geo-referenced data on the Web [1, 10, 11, 22].

Of particular interests to this study are several recent studies on identifying the dominant location of search queries [7, 19], website visitors' geographic interests patterns [17], and geo-sensitive ranking [12]. Gravano et al. applied several machine learning methods to classify search queries with regard to their *localness*, based on a number of features derived from the associated search results returned by a search engine [7]. Wang et al. exploited the search results clicked by the users, and extracted location names from these Web pages as potential location candidates for annotating the associated search queries [19]. More recently, Sheng et al. proposed a click-based model to discover the geographical patterns of the visitors' interests for websites, and studied the temporal change of these patterns [17]. Their work showed that the click data contained strong signals of users' various interests for Web documents in different geographic locations. Lee et al. proposed several variations of the PageRank-style link-based ranking algorithms by considering the geographic attributes of the hyperlinks [12].

In our work, we adopt a different methodology to directly model the geographic distribution of users' interests by their past click patterns. The proposed models are applicable to both search queries as well as Web pages.

3 Dataset

We randomly sampled from the anonymized *Yahoo! Search*[1] query log a set of 10,000 queries which were believed to be representative and covering a broad range of search topics. We then collected about 450M user click records for these queries from the click logs between November 2007 and April 2008, which contained about 1.5M unique (query, URL) pairs. Each record consisted of the following fields: query, URL, rank of the URL in the search results, IP address, and the aggregated number of clicks which originated from this address. An example is (*"florida international university"*, *"http://www.stanford.edu/group/ree/reeusa03/posters/FIU.pdf"*, 58, 4ae96544, 1). Note that we collapse multiple clicks from the same search session into one.

We first removed the IP addresses of popular proxy servers and large ISP hubs, and then geo-mapped each of the remaining of the IP addresses onto a geographical location (i.e., latitude, longitude, city/town, county, state, zipcode, and country) using a proprietary IP lookup database. In the previous example *"florida international university"*, the geographical location is (25.727, -80.235, Miami, Miami-Dade, FL, 33133, USA).

[1] http://search.yahoo.com

4 Visualize Geographic Interests Distribution

The *geo-sensitivity* of a query denotes that, to answer the query, Web pages that either have explicit / implicit association with certain geographical location(s) or are considered more relevant to users in certain geographical location(s) will be considered more *relevant*. To make it more concrete, we define four categories of geo-sensitivity for queries in Table 1. For the rest of this paper, we will follow these definitions. If a search query falls into either the *explicit* or the *implicit* sensitivity category, it is referred to as a *Geo-Sensitive Query* (*GSQ*), otherwise a *Non-Geo-Sensitive Query* (*NGSQ*).

Previously we discussed the disadvantage of relying solely on the syntactic information of a query to infer its geo-sensitivity. Assume that a user submits a query and the search engine returns a list of Web pages, he/she is more likely to click on Web pages that appear *more relevant*. For the *GSQ*s, whether a Web page is *geographically* relevant is an important factor of the overall relevance perceived by the user. Thus, given a search query, once we aggregate a significant number of user clicks and conduct a reverse-lookup of the IP addresses to find out where they come from, it is possible to learn about the locality preference of this query by *"following the herd"*. For example, a *GSQ "broward community college"* received the majority of user clicks originated from the Fort Lauderdale, FL area, which could indicate that this query was *more geographically sensitive* to the Fort Lauderdale area than anywhere else. Thus, we examine the following hypothesis:

Hypothesis 1. *Given a search query or a Web page, the geographic distribution of the user clicks it receives approximates the geographic distribution of users' interests in it.*

Hypothesis 2. *The geographical distributions of user clicks for* NGSQ*s and* GSQ*s have different patterns: the distributions for* NGSQ*s tend to resemble the user population distribution, while the distributions for* GSQ*s tend to deviate from the population distribution.*

To examine the above hypothesis, we visualize the aggregated click records in our dataset using *click maps*. The *click map* of a query visualizes the geographical locations of search users who have issued the query and

Table 1: Definitions of Query Geo-Sensitivity

Sensitivity Category	Definition	Examples
Explicit	Queries about local business, organizations, product, and information of a particular location, which explicitly specify a geographical location.	*cleveland plain dealer, nissan dealer in bay area, boston weather*
Implicit	Queries that implicitly specify a location with a popular or famous local business or landmark.	*bridgeport board of education, broward community college, wilkes regional medical center, darlington raceway*
Local	Queries that are not specific to a particular location but local information is implicitly preferred; typically contain the name or type of a business, service, or organization without a specific location.	*car wash, AAA branch, movie showtime*
Non-Sensitive	Generic or navigational queries, and queries about information interesting to users regardless of their physical locations.	*yahoo, myspace, disneyland, new york times*

also clicked on (some of) the results. Each dot on the click map of a query represents the aggregated user clicks originated from the corresponding geographic location.

Figure 1(a) shows a click map of all the randomly sampled search queries, in which the geographical distribution of user clicks mostly follows the population distribution in the U.S. Comparing Figure 1(a) with Figure 1(b) which shows a click map of the query *"Google"*, the two resemble each other, indicating that the query *"Google"* is not *geographically sensitive* to a particular location.

In contrast, Figure 1(c), 1(d), 1(e), and 1(f) show the click maps of four *geo-sensitive* queries: *"Aspen Grove Shopping Center"*, *"Sun Country Airlines"*, *"93.7 Houston"*, and *"University of San Francisco"*, respectively. The spatial distributions of the user clicks for these queries all display a strong deviation from the population distribution, and have unusually high density in some regions. For example, *"Sun Country Airlines"* received more than 30% of its clicks coming from Minneapolis, MN (a hub of the airline), and *"93.7 Houston"* – a local radio-station in Houston – received more than 80% of its clicks from Houston, TX. As we examined the click maps of more search queries, it was consistent that the geographical distribution of user clicks were usually quite different between *GSQs* and *NGSQs*.

5 Modeling the Geographic Interests

In this section we study two models of the geographical distribution of users' interests based on their clicks.

5.1 Vector model

We first define the set of geographical regions as $\mathbb{R} \doteq \{\mathbf{r}\}$, where $\mathbf{r}$ denotes a physical location of an appropriate granularity (i.e. how large the region is). The granularity can be either iteratively tuned or chosen arbitrarily.

Next, we define the set of queries as $\mathbb{Q} \doteq \{\mathbf{q}\}$, where $\mathbf{q} = \langle c_1, c_2, \cdots, c_k \rangle$ denotes a query q. $\mathbf{q}$ is a k-dimensional vector, and each dimension c_i represents *the probability of the query receiving user clicks from a geographical region* $r_i \in \mathbb{R}$. Here c_i can be defined as

$$c_i = \frac{c_{r_i}}{\sum_{r_t \in \mathbb{R}} c_{r_t}} \cdot (1 + \log \frac{|\mathbb{Q}|}{|\{\mathbf{q} : c_{r_i} \neq 0\}|}), \qquad (1)$$

where c_{r_i} is the number of user clicks from r_i for $\mathbf{q}$.

The same model can also be applied to the set of Web pages as $\mathbb{D} \doteq \{\mathbf{d}\}$, where $\mathbf{d} = \langle c'_1, c'_2, \cdots, c'_k \rangle$ denotes a Web page d, and each dimension c'_i denotes *the probability of the Web page receiving user clicks from a geographical region* $r_i \in \mathbb{R}$, with a definition of c'_i similar to c_i.

5.2 Maximum-likelihood model

Backstrom et al. presented a generative model to derive the geographic center and spatial dispersion of search queries [3]. The model was based on the observation of the *view* event in the search logs: a user issues a query from an IP address which is mapped to a physical location.

In this paper, we extend this model to apply it on the *click* events recorded in the search logs: a user issues a search query and then clicks on a certain search result Web page. The intuition is that compared with the action of issuing a query, user clicks represent stronger signals of endorsement with regard to the locality preference of the user.

Let $\mathbb{Q} \doteq \{\mathbf{q}\}$ denote the set of queries in the query logs. Each query $\mathbf{q}$ has a geographic center of "interests" $\mathcal{C}$, from where most of the user clicks for this query are observed. And the observed number of user clicks from anywhere else is in reverse proportion to its distance d away from $\mathcal{C}$. Now the probability of observing a click by a random user at distance d from $\mathcal{C}$ is

$$p = C \cdot d^{-\alpha}, \qquad (2)$$

where C is the observed number of user clicks originated from $\mathcal{C}$, and the exponent α is a spreading factor which specifies how fast the number of clicks decreases further away from $\mathcal{C}$.

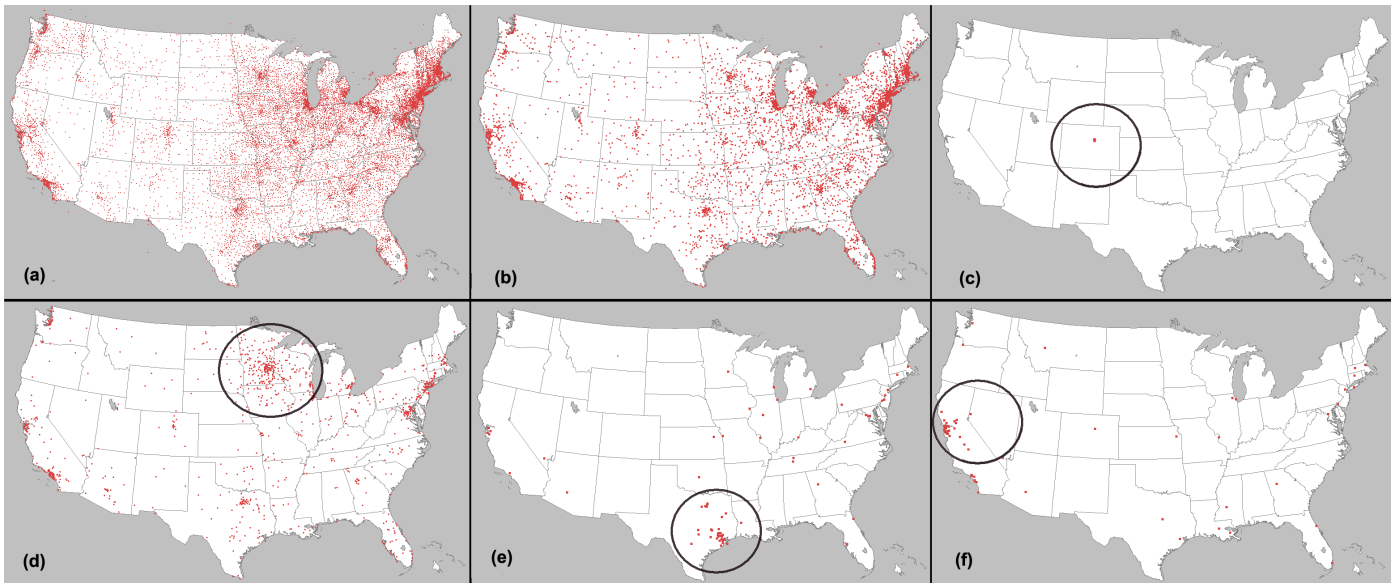

Figure 1: (a) The click map of all the randomly sampled search queries. (b) Query *"Google"*. (c) Query *"Aspen Grove Shopping Center"*. (d) Query *"Sun Country Airlines"*. (e) Query *"93.7 Houston"*. (f) Query *"University of San Francisco"*. Regions with unusually high density of clicks for the *geo-sensitive* queries are marked by the black circle.

Therefore, given a query $\mathbf{q}$ and the past user clicks for this query, we can infer the center of interests as well as the corresponding spreading factor in a maximum-likelihood fashion,

$$\arg\max p^m(1-p)^n, \tag{3}$$

where m is the number of observations of a click by the user at distance d from $\mathcal{C}$, and n is the number of observations otherwise. Similarly, we can apply this model for each Web page $\mathbf{d} \in \mathbb{D}$ that has registered user clicks in the search logs.

For more details on this model, we refer readers to the original paper [3].

6 Classify queries based on geo-sensitivity

Witnessing the distinctive patterns of geographical click distributions of *GSQs*, we discuss how to algorithmically derive the geo-sensitivity for queries under a classification scheme. Accepting Hypothesis 2, can we differentiate *GSQs* from *NGSQs*, and infer the locality preference of a query based on its geographical click distribution?

Without loss of generality, the task to identify the *geo-sensitivity* of a search query can be cast into a binary classification problem, formally defined as follows:

Definition *Given a set of queries $\mathcal{Q} = \{q\}$, classify q into one of the two classes: (1) a class of* geo-sensitive queries *GSQ ($GSQ \subseteq \mathcal{Q}$), and (2) a class of* non-geo-sensitive queries *$NGSQ$ ($NGSQ = \mathcal{Q} - GSQ$).*

Vector Model. Let vector $\mathbf{p} = \langle d_1^p, d_2^p, \cdots, d_k^p \rangle$ denote the user population, with each dimension d_i^p represents the population in a geographical region r_i same as in $\mathbf{q}$. By accepting Hypothesis 2, we chose to approximate the population distribution using the aggregated click probability distribution of all the queries in $NGSQ$, i.e., the vector summation of all vectors $\mathbf{q}' \in NGSQ$ and normalized:

$$\mathbf{p} = \frac{\sum \mathbf{q}'}{|NGSQ|}, \forall \mathbf{q}' \in NGSQ \tag{4}$$

Given a query q, we can then measure the distance $\mathcal{D}$ between its geographical click probability distribution $\mathbf{q}$ and the user population distribution $\mathbf{p}$ approximated by Equation 4:

$$\mathcal{D} = \mathbf{dist}(\mathbf{p}, \mathbf{q}) \tag{5}$$

where $\mathbf{dist}$ can be an appropriate standard distance measure of choice, e.g. *chi-squared* χ^2, *KL Divergence*. We can then classify q as GSQ or $NGSQ$ based on $\mathcal{D}$.

Maximum-likelihood Model. The spreading factor α is an indicator of how *local* a query or a Web page is: a larger α indicates a query or a Web page of more *local* interests. Thus, given a labeled training set of *NGSQs*, we can calculate the *"norm"* (e.g. mean, mode) of α for such queries, and then use it as a threshold to determine if a given query is a *GSQ* (if it has a larger-than-threshold α) or a *NGSQ* (otherwise).

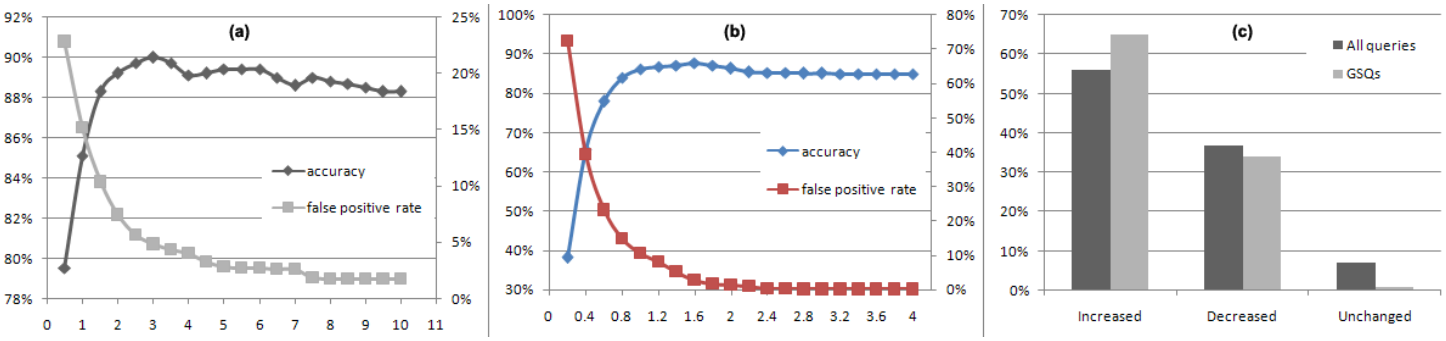

Figure 2: (a): Classification accuracy and false positive rate of the vector model classifier. (b): Classification accuracy and false positive rate of the maximum-likelihood classifier. (c): Percentage of all queries and *GSQ*s with increased, decreased, and unchanged DCG scores. It is not surprising to see that there is more relevance improvement when we apply the model on a set of queries comprised entirely of *GSQ*s.

7 Improve relevance ranking for geo-sensitive queries

We discussed how to algorithmically identify geo-sensitive queries in the previous section. In this section, we propose to improve the quality of search relevance ranking for the geo-sensitive queries by taking into account the geographic interests of the users.

We see in Section 4 that for a search query, the pattern of its geographical click distribution could indicate the locality preference of the users. For example, the query *"amc mercado 20"* receives more user clicks from Santa Clara, CA than any other area, which may indicate that this query about a local movie theater is indeed more *interesting* to search users in the same geographical region. Thus, the click map of this query will likely show high density of clicks in the Santa Clara region. For the same reason, Web pages that are relevant to the AMC Mercado 20 movie theater will more likely register clicks from local search users than users from other locations (e.g. search users in New York). Thus, the geographic distribution pattern of the aggregated user clicks on these pages may also show a concentration of clicks in the Santa Clara area. This observation suggests us study the correlation between the geographic distribution of interests of the query and of the corresponding Web pages as a measure of their *relevance*.

Vector Model. We denote a search query q by a k-dimensional probability distribution $\mathbf{q} = \langle d_1, d_2, \cdots, d_k \rangle$, and denote a Web page w by $\mathbf{w} = \langle d_1^w, d_2^w, \cdots, d_k^w \rangle$, where each dimension d_i or d_i^w represents *the probability of registering user clicks from a geographical region* r_i. Thus, we can measure the *similarity* in geographical click distributions of the query and the Web page as the *cosine similarity* of the two vectors:

$$\mathbf{D}(\mathbf{q}, \mathbf{w}) = \frac{\mathbf{q} \cdot \mathbf{w}}{\|\mathbf{q}\|\|\mathbf{w}\|} \tag{6}$$

where $\cdot$ denotes the dot product operation and $\|\cdot\|$ denotes the magnitude. Note that other vector distance metrics can also be applied.

Maximum-likelihood Model. Given a query $\mathbf{q}$ and a Web page $\mathbf{w}$, let $p_q = C_q \cdot d_q^{-\alpha_q}$ and $p_w = C_w \cdot d_w^{-\alpha_w}$ denote the click probability at a point at distance d_q and d_w away from the centers of interests $\mathcal{C}_q$ and $\mathcal{C}_w$, correspondingly. We can measure the proximity of two centers of interests with regard to the corresponding spreading factors as

$$\mathbf{P}(\mathbf{q}, \mathbf{w}) = C_q d_c^{-\alpha_q} \cdot C_w d_c^{-\alpha_w}, \tag{7}$$

where d_c is the distance between $\mathcal{C}_q$ and $\mathcal{C}_w$. Intuitively, $\mathbf{P}(\mathbf{q}, \mathbf{w})$ becomes maximal when d_c approaches 0, i.e. when the centers of interests for the search query and the Web page converge.

8 Experimentation and Discussions

8.1 Query Geo-sensitivity Classification

To establish the ground-truth for evaluating the classification accuracy, a test-set of 1,000 queries were randomly sampled from the search logs, and each query in the test-set was reviewed by a team of professional editors and classified as either *GSQ* or *NGSQ* according to the *geo-sensitivity definitions* outlined in Table 1.

We limited the scale of evaluation to geographic locations within the contiguous United States. Based on the classification results in pilot tests comparing various granularity settings, we fixed the region granularity as a *State*. Thus, $\mathbf{q}$ is a 48-dimensional vector, and each dimension represents the probability of user clicks from the corresponding State. Same for $\mathbf{p}$. We chose χ^2 as the distance metric.

We first calculated $\mathbf{p}$ using the 10K queries training dataset, and then applied the proposed classification methods on the 1K queries test set. Evaluated against

label	vector model	maximum-likelihood model
GSQ	*metro community college, utah power, dixon middle school, kcci, idaho lottery, alaska weather, salt lake tribune newspaper, oklahoma county assessor, SLED, job openings in alabama, utah jazz, CCSU, iowa dmv, oregonian, green bay press gazette*	*MJR Waterford Cinema 16 Waterford MI, whittemore speedway, mercedes benz san antonio, berkeley nj, job openings in alabama, speaker of baseball, deaf smith county, idaho lottery, maui vacation rental, city of brawley, south plains college, wtvm, utah power, kcci, puerto vallarta hotel*
$NGSQ$	*online dictionary, internet explorer, adobe, winzip, famous quotes, microsoft word, the notebook, ipod shuffle, mcafee, spanish translator, xbox, us postal service, poker, matrix, colors*	*wwe, elton john, meditation, geri halliwell, salary, taxi, charlie sheen, supreme court, Uma Thurman, eliza dushku, toledo spain, torrent search, the producers, baseball, mcafee*

Table 2: Top 15 queries classified as *geo-sensitive* or *non-sensitive* by the vector model and the maximum-likelihood model.

the ground-truth editorial labels, the two models showed quite promising accuracy. In Figure 2(a) and 2(b), the X-axis represents $\mathcal{D}$ for the vector model and α for the maximum-likelihood model; the Y-axis on the left applies to the accuracy curve, and the Y-axis on the right applies to the false positive rate curve. Here the *classification accuracy* is the ratio of queries that were correctly classified as either *GSQ* or *NGSQ*, and the *false positive rate* is the ratio of *NGSQ*s that were incorrectly labeled as *GSQ*s. The vector model achieved 90% accuracy and 4.8% false positive rate when $\mathcal{D} = 3.0$ (see Table 3 for the confusion matrix), while the maximum-likelihood model achieved 87.8% accuracy and 2.8% false positive rate when $\alpha = 1.6$. For most queries, the vector model was an order of magnitude faster to compute than the maximum-likelihood model.

	GSQ	***NGSQ***
Labeled as *GSQ*	91	41
Labeled as *NGSQ*	59	809

Table 3: The confusion matrix of the classification results using the vector model with $\mathcal{D} = 3.0$.

Table 2 shows the top 15 queries that were classified as either *GSQ* or *NGSQ*, order by $\mathcal{D}$. There are several *GSQ*s that are worth mentioning. For example, *"kcci"* is a local TV station in Des Moines, IA. *"SLED"* is the acronym for the *South Carolina Law Enforcement Division*. *"CCSU"* is the shorthand for *Central Connecticut State University*. And *"wtvm"* is a television station in Columbus, Georgia. All these queries would probably not have been easily identified as *geographically sensitive* had we relied only on the syntactic information of the queries. It is also worth noting that the top queries identified using the two models did not have much overlap. In particular, the maximum-likelihood model picked up much more celebrities' names as *NGSQ*s (five out of the top 15) than the vector model. We plan to further investigate this interesting difference.

8.2 Relevance Ranking Improvement

To measure the improvements with regard to search relevance ranking, we used a popular metric Discounted Cumulative Gain (DCG) [8, 9]. DCG allows us to attach more importance to documents that are ranked higher in the list, and to differentiate various levels of subjective relevance judgment made by the human evaluators. For a given query q, DCG is defined as

$$DCG(q, N) = \sum_{d=1}^{N} \frac{2^{R(d)} - 1}{\ln(1 + d)} \tag{8}$$

where $R(d)$ is the editorial rating of the $d-$th Web page in the top N ranked search results. Intuitively, a higher DCG score indicates a better relevance ranking.

We evaluated the relevance improvement on approximately 8,900 search queries sampled from the query logs of Yahoo! Search. For each of these queries, we calculated the DCG scores for a baseline ranking algorithm, and an alternative ranking algorithm that took into account the geographic distribution of users' interests. The baseline ranking algorithm was a proprietary decision-tree-based ranking algorithm trained on a very large number of features, and generated the same search ranking as Yahoo! Search. On the other hand, the alternative ranking algorithm learned to utilize $\mathbf{D}(\mathbf{q}, \mathbf{w})$ in Equation 6 as an important ranking feature, in addition to the existing features being used by the baseline algorithm. We then compared the DCG scores of the alternative algorithm to the baseline.

Due to confidentiality, we are not allowed to report the absolute DCG values on this data set, and we will quantify the performance changes in relative terms. On average, the alternative ranking algorithm outperformed the baseline algorithm by 1%, statistically significant with $\alpha = 0.05$. See Figure 2(c) for a breakdown by query class and Table 4 for a sample of the geo-sensitive queries with increased DCG scores.

9 Conclusion and Future Work

We studied two click-based models of the geographic distribution of users' interests. Given a search query or a

query	DCG@1	DCG@10
greeks pizza muncie	+2.7%	+3.7%
st louis massage school	+1.0%	+13.9%
el paso texas flooding	+ 2.2%	+27.1%

Table 4: Examples of geo-sensitive queries with DCG improvement.

Web page, by mining the geographic distribution of past user clicks, we can indirectly model the distribution of users' interests. We then discussed two applications of the click-based models. We first applied the models to classify search queries based on their locality preference and identify users' *distribution of interests* for such queries. We then used the model to generate meaningful metrics for measuring the proximity between a search query and a Web page with regard to users' locality preference, with the goal to improve search ranking relevance. We presented the results from our empirical experiment on a large dataset, and both the two models demonstrated promising results.

The maximum-likelihood model is computationally expensive, but nicely complements the vector model due to its ability to express proximity. The vector model, on the other hand, is computationally efficient but does not consider the proximity among dimensions due to the arbitrarily-defined boundaries. We plan to further evaluate the two models and investigate their utilities for relevance ranking.

Acknowledgments We gratefully acknowledge Ravi Kumar, Jasmine Novak, and Lars Backstrom for their valuable contributions, including an implementation of the maximum-likelihood model. We also thank the anonymous reviewers for the constructive feedbacks.

References

[1] S. Ahern, M. Naaman, R. Nair, and J. Yang. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *Proc. of the Joint Conference on Digital Libraries*, pages 1–10, 2007.

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

[3] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *Proc. of the 17th Intl. Conference on World Wide Web*, pages 357–366, 2008.

[4] K. A. V. Borges, A. H. F. Laender, C. B. Medeiros, and J. Clodoveu A. Davis. Discovering geographic locations in web pages using urban addresses. In *Proc. of the 4th ACM SIGIR Workshop on GIR*, pages 31–36, 2007.

[5] N. Cardoso and M. J. Silva. Query expansion through geographical feature types. In *Proc. of the 4th ACM SIGIR Workshop on GIR*, pages 55–60, 2007.

[6] G. Fu, C. B. Jones, and A. B. Abdelmoty. Ontology-based spatial query expansion in information retrieval. In *Lecture Notes in Computer Science*, pages 1466–1482, 2005.

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

[8] K. Jarvelin and J. Kekalainen. Ir evaluation methods for retrieving highly relevant documents. In *Proc. of the 23rd Annual ACM Conference on Research and Development in Information Retrieval*, pages 41–48, 2000.

[9] K. Jarvelin and J. Kekalainen. Cummulated gain-based evaluation of ir techniques. In *ACM Transactions on Information Systems*, pages 422–446, 2002.

[10] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *Proc. of the 15th Intl. Conference on Multimedia*, pages 631–640, 2007.

[11] L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proc. of the 17th Intl. Conference on World Wide Web*, pages 297–306, 2008.

[12] H. C. Lee, H. Liu, and R. J. Miller. Geographically-sensitive link analysis. In *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 628–634, 2007.

[13] Y. Peng, D. He, and M. Mao. Geographic named entity disambiguation with automatic profile generation. In *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 522–525, 2006.

[14] M. Sanderson and Y. Han. Search words and geography. In *Proc. of the 4th ACM SIGIR Workshop on GIR*, pages 13–14, 2007.

[15] M. Sanderson and J. Kohler. Analyzing geographic queries. In *Proc. of 1st ACM SIGIR Workshop on GIR*, 2004.

[16] V. Sengar, T. Joshi, J. Joy, S. Prakash, and K. Toyama. Robust location search from text queries. In *Proc. of the 15th ACM Intl. Symposium on Advances in Geographic Information Systems*, pages 1–8, 2007.

[17] C. Sheng, W. Hsu, and M.-L. Lee. Discovering geographical-specific interests from web click data. In *Proc. of the 1st ACM Intl. Workshop on Location and the Web*, pages 41–48, April 2008.

[18] C. Wang, X. Xie, L. Wang, Y. Lu, and W.-Y. Ma. Web resource geographic location classification and detection. In *Proc. of the 14th Intl. Conference on World Wide Web*, pages 1138–1139, 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 *Proc. of the 28th ACM Conference on Research and Development in Information Retrieval*, pages 424–431, 2005.

[20] Q. Zhang, X. Xie, L. Wang, L. Yue, and W.-Y. Ma. Detecting geographical serving area of web resources. In *Proc. of the 3th ACM SIGIR Workshop on GIR*, 2006.

[21] V. W. Zhang, B. Rey, E. Stipp, and R. Jones. Geomodification in query rewriting. In *Proc. of the 3th ACM SIGIR Workshop on GIR*, 2006.

[22] M. Zhizhin, E. Kihn, V. Lyutsarev, S. Berezin, A. Poyda, D. Mishin, D. Medvedev, and D. Voitsekhovsky. Environmental scenario search and visualization. In *Proc. of the 15th ACM Intl. Symposium on Advances in Geographic Information Systems*, pages 1–10, 2007.