A Case Study of Using Geographic Cues to Predict Query News Intent

Ahmed Hassan*
U. Michigan Ann Arbor
EECS Dept.
Ann Arbor, MI 48109
hassanam@umich.edu

Rosie Jones
Yahoo! Labs
4 Cambridge Center
Cambridge, MA 02142
jonesr@yahoo-inc.com

Fernando Diaz
Yahoo! Labs
4401 Great America Parkway
Santa Clara, CA 95054
diazf@yahoo-inc.com

ABSTRACT

Geographic information retrieval encompasses important tasks including finding the location of a user, and locations relevant to their search queries. Web-based search engines receive queries from numerous users located in very different parts of the world. A typical way for people to find news is through a general web search engine, which makes it important for search engines to recognize queries with news intent. An important question for geographic information retrieval is how we can benefit from geographic cues to predict the intent of users. This work presents a case study of an application using geographic features to improve the quality of an important web search task, involving predicting which queries have news intent and hence are likely to receive clicks on news search results. Our case study suggests that information derived from geographic features can help the task. The information we consider includes cues derived from the location of the user, from the IP address, the location relevant to the query, automatically extracted from the query string, and the relation between the two locations. We build a classifier that uses geographical cues to predict whether a query will result in a news click or not. We compare our classifier to a strong baseline that use non-geographic click-based features and we show that our classifier outperforms the baseline for geographic queries.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Selection process

General Terms

Experimentation

Keywords

geographic information retrieval, local news search

*This work carried out while this author was at Yahoo! Labs

1. INTRODUCTION

News is available in print, television and radio, for local, national and international news. Each of these media provide their own websites, but a typical way for people to find information online is through web search. Web searchers may be interested in learning about local and international news, and if we can customize web search results to the user based on their location, we can greatly improve user satisfaction. We analyzed usage logs from searches on a general purpose search engine to see how geography affected user interest in news, as measured by clicks on news results. We analyzed data from over two million searches with news results over a period of two weeks in February 2008.

Our contributions include demonstrating that geo-spatial information helps predict when a user is likely to click on a news search result, based on the population density of the user's location, the distance of the user from the search topic, and the actual location in the search query. In addition we show that this information is complementary to other factors predicting news clicks, and can be combined to produce improved performance.

In Section 2 we describe related work on geographical properties of search and news. In Section 3 we give a detailed overview of the search and click information on web news that form the basis of our experiments. In Section 4 we describe briefly the black-box location identification system we use to pre-process our data to extract locations from queries and infer locations from IP addresses. In Section 5 we describe geographical and lexical features that we derive from the data, and give overviews of their distributions. In Section 6 we describe the task of predicting which queries are likely to lead to clicks on news search results, and our experiments using geographical features to improve performance at this task.

2. RELATED WORK

News corpora play an important role in the history of information retrieval. Much seminal research in information retrieval was conducted using the TREC news corpora [24]. In the 2004 High Accuracy Retrieval from Documents (HARD) task, the geography of retrieved documents was one of the attributes used to define the relevance of documents [1]. Like the HARD task, we focus on geographic aspects of the query. However, we study the relationship between the geographic intent and news intent manifested in a set of queries. While the HARD task studied 50 queries, we study several millions.

Previous work analyzing the presence of geographic terms in web search queries has shown that 13-15% of queries contain locations [20, 11, 8]; these may be for many topics, including *local search* for goods and services, as well as *news search* for up-to-date news information. Gan et al [8] showed that 35% of queries containing a place-name are for locating goods-and-services, the Locate category using Rose and Levinson's taxonomy of query types [19]. They also showed that around 6% of geo-specified queries are for media, including news, radio, papers and magazines. In Section 5 we give an estimate of the proportion of queries containing placenames which could be explicitly satisfied by showing news results.

Jones *et al.* [11] showed that users' search-interests in topics has a geographic component, with users interested in restaurants and day-care near their locations, and maps for locations that are far from them. This kind of result may follow from and validate Christaller's Central Place Theory [4].

Backstrom *et al.* [2] showed that we can identify a center and a distribution of interest for a query, by examining the IP addresses of users who issue the query. Zhuang *et al.* [25] look at the geographic distribution of search engine queries receiving a click. Their work shows that the location of a searcher, inferred from the IP address, carries some semantics about the geography of the search term itself. In our work we examine whether the distance of the user (inferred from their IP) and the places specified in their query impact their propensity to click on news.

Sheng *et al.* [21] give an algorithm for biasing search results based on geographic distributions; for example the continent a searcher is from may influence the types of sports that user is interested in.

Geographic information also exists in the news corpus itself. Mehler *et al.* analyze the spatial distribution of people mentioned in news articles using a named entity recognizer [15]. Mei *et al.* conduct similar spatiotemporal analysis of topics using a weblog corpus [16]. Liu and Birnbaum [14] show that the geographical source of a news item can impact the perspective on the story given in the article. Teitler et al. [22] describe a system for extracting locations from news, clustering them, and presenting the results on a map.

Our work focuses on relating the geographic information of the *user* and the *query* which is appropriate in a streaming enviroment like news where we may not have time to run analysis on the documents. The use of geographic information in documents has been explored in both information retrieval and data mining. In geographic information retrieval, van Kreveld et al [23] retrieved documents by using a relevance score which linearly combined textual and geographic similarity. Purves *et al.* [17] extract location information from documents and linearly interpolate geographic and text-based retrieval scores in the context of free text ranking. Other work in geoCLEF (eg. [12]) has used geographic term expansion on the queries and documents and then conventional term matching using BM25 on the resulting expanded texts. Liberman et al [13] describe the architecture of a spatio-textual search engine.

Diaz [5] described a method for predicting when web searchers will be interested in clicking on news, based on changes in query and click frequency, and the distribution of documents in the news corpus.

Our novel contribution is examining geographic information in the context of millions of web search queries, and showing that we can use that geographic information to predict when a news search result will be clicked on.

3. DATA

Searches on a general purpose major search engine represent situations where there is no *explicit* intent for news documents. The user may be interested in the website of a company or information about a product. One way to determine news intent is to look at the results clicked on by the user. We hypothesize that a user interested in news is more likely to click on a news result than on a non-news result. (We validate this hypothesis in a small experiment in Section 3.1 below).

To gather data with possible news intent we conducted the following experiment. For a subset of all traffic, we presented a small box containing up to three news articles above the top-ranked web result *if the query retrieved any documents in the news index.* There may be a news intent if users, in response to this display, clicked on a news article, and our task will be to predict which queries lead users to click on news articles. Our data set consists of triplets of $\langle \text{IP address}, \text{query}, \text{click} \rangle$. User location is inferred from the IP address, the search query is a case-normalized query, and click is a binary indicator of whether the user clicked or not.

We gathered data from February 13-26, 2008 on a sample of web traffic to a major search engine. This resulted in 2,205,155 triples. All data was processed in accordance with the search engine's privacy policy, and no attempt was made to connect different queries from the user or identify the individual who issued the query.

3.1 Manual Labeling of Queries with Placenames

Our data consists of a set of queries which match a document in the news corpus. While some of these may be intended as news queries, others may be intended as informational queries, or queries for local shopping options. We manually labeled 300 queries, restricting the sample to those which contain placenames, to identify what proportion of them are unambiguously for news. This will give us a lower-bound on the proportion of these queries which can be satisfied with news results, since ambiguous queries (such as a single placename like "afghanistan" or "new zealand") may occasionally also be intended as news queries.

In Table 1 we see that when 300 queries that have matches in the news index are manually labeled, about 21% are unambiguously for news (such as "illinois university shootings") while 26% are a location with no other information, such as "los angeles", "jericho" or "pakistan". In some cases these may be intended as news queries, while in others they may be informational. 12% of queries were for sports information, which may be sports news, such as "chicago cubs" and

type	count	%	norm. ctr
news	62	21	4.25
news source	15	5	1.00
sports	37	12	2.32
locationOnly	77	26	1.68
other	109	36	1.27

Table 1: About 21% of queries which contain a place-name and match a news index are explicitly for news, when manually labeled, and about 5 percent are for news sources. 27% are queries which contain only a place-names, which are ambiguous in intent. Normalized news-result click-through rate (norm. ctr) is highest for queries which are unambiguously news-related, but sports and location-only queries also have high news result click-through rate.

"manchester united". Queries for news sources, such as "india times", represent 5% of our manually labeled sample, while the remainder were informational queries, such as "new york state tax" or queries with placenames used as part of an organization name, as as "malaysia airlines".

We then calculated the news-result click-through rate in each of these manually labeled classes. We see that news-result click-through rate is highest for the queries identified unambiguously as news queries, while sports queries also have quite a high click-through rate on news stories - web searchers who query for sports topics are often interested in news results, and indeed most news sources have a sports section. Queries consisting of just a location, had on average higher click-through rate than other queries, suggesting that location queries such as "lebanon" can sometimes be intended as news queries. We will examine this more in section 5. Queries for news sources (eg. "new york times") had the lowest news result click-through rate - these are navigational queries for users are looking for a specific website rather than general news from any source about the place-name mentioned.

4. EXTRACTING GEOGRAPHIC INFORMATION FROM DATA

We use state-of-the art tools for extracting geographic information from our data. Our research does not focus on the details or performance of those extraction techniques, which is an important research area in its own right, but rather on how much benefit can be obtained from applying current geographic information extraction technology to an information retrieval task, when it is used as a black-box.

In this section we give a brief overview of the tools and techniques we use to extract geographic information from our data. More specifically, we will describe how we identify place-names in queries, identify user location from IP address, and measure the distance between any two given locations.

We use the system described in [18] to identify place-names in queries. The system identifies which words in a search query denote place-names. It uses both context-dependent and context-independent features to identify place-names in the query. After identifying a place-name in the query, the system also tries to map the location to a large database of place-names to find out which particular location the user has in mind.

The IP address can be utilized to associate a location with the user issuing the query. We use information supplied from the Regional Internet Registries (RIRs). An RIR is a governing body that is responsible for the administration of Internet addresses in a specific geographic region. The RIR databases keeps track of IP addresses, Internet Service Providers (ISPs), and general geographic location. Using this information, we can identify the state and city of the user IP address.

To measure the distance between two given locations, we map each location to a longitude and latitude, and then calculate the standard spherical distance between the longitudes and latitudes. When the place-name is a general area, such as a state or country, we map it to a bounding box and measure the distance to/from the center of that bounding box.

For more details about the geographic information extraction process, please refer to [18].

5. GEOGRAPHIC PROPERTIES OF QUERIES

In this section we describe a set of geographic features that may be used to predict whether a query has news intent. We propose a large set of features some of which are query related, others are user related, and the rest consider the relation between the locations associated with the query and the user location. We describe each feature and measure the correlation between it and the probability that a query receives a news click in the following subsections.

5.1 Query Related

5.1.1 Query Location Confidence

The system we use for identifying place-names in queries assigns a score to each candidate place-name. This score acts as an estimate of how confident the system is in the fact that the candidate place-name is indeed a place-name given the context. We only consider candidate place-names if this score exceeds 0.5. Each query is assigned a value between 0.5 and 1. This value is the location confidence of the location in the query. If the query contains more than one location, we use the maximum location confidence of all locations in the query.

Our hypothesis is that queries containing ambiguous place-names may be less likely to be related to news, while queries with unambiguous place-names may be more likely to be related to news.

Figure 1 shows how the probability of receiving a news click for a given query is affected by our level of confidence in the location in the query.[1] The horizontal axis represents different levels of confidence in the place-name in the query. The vertical axis represents the probability of receiving a news click for each confidence interval.

[1]Absolute click probabilities have been normalized to hide sensitive information.

We notice that queries with high location confidence are more likely to receive news clicks. On the other hand, queries with low location probabilities are less likely to receive news clicks.

We see a drop in the bin 0.9 - 1.0 confidence. This is because this bin contains some very confidently recognized location queries, such as "japan" and "australia" which have low news click-through rates, as well as some others such as "iran" and "pakistan" which have very high news click-through rates.

This motivates an analysis of the place-names themselves - names of locations which tend to appear in news-stories may be a better indicator of how likely it is the user is thinking of news when they issue the query.

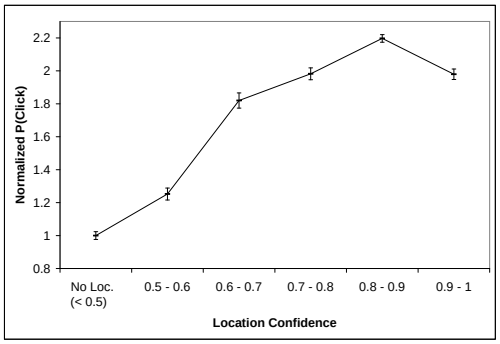

Figure 1: Normalized probability of receiving a news click for a given query for different bins of location confidence.

Table 2 shows a set of examples for locations with both high and low confidences. Those with high location confidence are locations such as "pakistan" and "chicago" which are also likely to be the site of news stories. Those with low confidence are those containing ambiguous terms such as "rogers" - which are also less likely to obtain a news click.

Overall we found that queries containing a place-name are up to twice as likely to receive a click on news results than queries which do not contain a place-name. This is important because it suggests that focusing on queries with geographic properties is a promising direction for predicting query news intent.

5.1.2 Location Type

Each place-name identified in a query is classified into a set of pre-determined location types. For example, the place-name might be a country name, a state name, a town name, etc. We use the location type as a categorical feature with a number representing each possible category.

Figure 2 shows how the probability of receiving a news click for a given query is affected by the location type. The horizontal axis represents different types of locations. The vertical axis represents the probability of receiving a news click for each location type.

We notice from the figure that countries and states are more likely to be associated with queries that received news clicks. This suggests that users tend to use country and state names more often when they are looking for news. On the other hand, towns are more likely to be associated with queries that did not receive news clicks. This indicated that users specify town names more often when they are looking for non-news results like services and businesses.

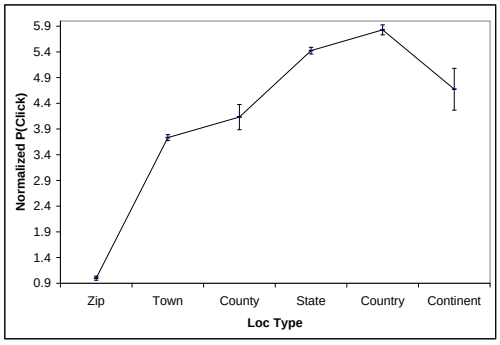

Figure 2: The normalized probability of receiving a news click for a given query for different location types.

5.1.3 Location Word/Click Probability

Certain place-names are more likely to be news related than other place-names, perhaps because more news-worthy events happen there. For example, a user query with a place-name like "kosovo" or "pakistan" is more likely to lead to a news click than a query with a place-name like "baltimore", or "utah". To capture this information, we build a large probability table that specifies, for each place-name, the probability that a query contain this place-name would receive a news click. We only consider place-names that occurred more than 20 times in the training set. The probability associated with each place-name is estimated as follows:

$$P(p) = \frac{N_{click}(p)}{N_{click}(p) + N_{noclick}(p)} \quad (1)$$

where $N_{click}(p)$ is the number of queries containing place-name p that received news clicks, and $N_{noclick}(p)$ is the number of queries containing place-name p that did not receive news clicks.

Each query is assigned a value depending on the place-names it contains. This value is calculated as:

$$L(q) = d\frac{N_{clicks}}{N_{click} + N_{noclick}} + (1-d)max_p P(p) \quad (2)$$

where N_{click} is the total number of queries in the training set that received news clicks, $N_{noclick}$ is the total number of queries in the training set that did not receive news clicks, and d is a damping factor, which is typically chosen in the interval $[0.1, 0.2]$.

The parameter d is included to smooth the values assigned to

(a) high confidence candidates

query	location	interpretation
new york city news	new york city	New York City/NY/US
pakistan election	pakistan	Pakistan
china economy	china	China
michigan lottery	michigan	MI/US
london fire	london	London/UK
earthquake in mexico	mexico	Mexico
chicago auto show	chicago	Chicago/IL/US
wisconsin primary	wisconsin	WI/US

(b) low confidence candidates

query	location	interpretation
wells fargo	Fargo	Company vs. Fargo/ND/US
jfk conspiracy	jfk	Person name vs. airport
timbaland	timbaland	Person name vs. Timbaland/Scotland
baldwin	baldwin	Person name vs. Baldwin/Canada
victoria beckham	victoria	Person Name vs. Victoria/Australia
paris hilton	paris	Person Name vs. Paris/France
rogers	rogers	Person Name vs. Rogers/AR/US
southwest	southwest	Airline vs. Southwest/IN/US

Table 2: Examples of high/low confidence candidate locations

High Prob. of News Click	Low Prob. of News Click
kosovo	bali
manila	guam
serbia	nashville
pakistan	lincoln
afghanistan	hampton
lebanon	napa
iran	alaska

Table 3: An example of a set of words that had high/low news click probabilities

queries. If we assign the probability of the place-name with the maximum probability to the query, queries with unseen place-names would always be assigned 0. Introducing the damping factor d makes sure that any query will get a non-zero probability even if the place-names in the query have never been seen before.

Table 3 shows some examples of the place-names that are more likely to be included in a news related query and some other place-names that are not. We notice from the table that some place-names with high probability are pretty stable and would receive high probabilities regardless of time, while others are more related to specific events and their probability would probably change with time. This suggests that updating the place-name probability table periodically to capture those variations. Our dataset is from a two-week period of time, so conducting experiments to capture temporal variation is outside the scope of this paper, but the interested reader may look at related work on how query frequencies vary (and correlate) over time (eg. [3], [9]).

Figure 3 shows how the probability a news related link will receive a click for a given query is affected by the query word/click probability score. The horizontal axis represent different bins of the query location word/click probability score. The vertical axis represents the probability of receiving a news click for each bin.

We notice that there is a very strong correlation between the query score and the probability that it will receive a news click. Queries with high scores are mostly likely to receive a news result click. While queries with low scores are mostly unlikely to receive news clicks. The figure shows how powerful this property can be if used as a feature for predicting news result clicks. In particular, it shows that the location in a query alone is a strong predictor of whether that query will receive a news click.

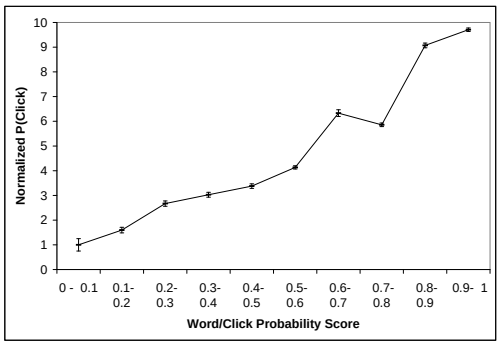

Figure 3: The probability of receiving a news click for a given query for different bins of query location word/click probability scores. The queries come from the test data. The word/click probabilities were calculated using the training data.

5.2 User Related - Population Density

We hypothesized that properties of the user's location might affect the user's interest in news. For example, whether peo-

Bin	Range
Very Low	0 - 2000
Low	2000 - 4000
Intermediate	4000 - 6000
High	6000 - 8000
Very High	8000 - 10000+

Table 4: Population density bins - in $person/km^2$

ple are from rural or urban locations (for which we took population density as a proxy) might affect their level of education or interest in local or international news. We used data from The United States Census Bureau to study the effect of population density on the user's news search behavior. We identify the user location for each query using the user's IP address as described above. We use that to attach a population density to each query. We then divide queries into several subsets based on their population density and estimate the probability that a given query from each subset would result in a news result click. The population density ranged from 30 $person/km^2$, for towns like Pahrump/NV, Buckeye/AZ, to 10,000+ $person/km^2$, for big cities like New York City and San Francisco.

Figure 4 shows how the probability of receiving a news click for a given query is affected by the population density of the user's location. The horizontal axis represent different bins of population density. The range of each bin is shown in Table 4. The vertical axis represents the probability of receiving a news click for each bin.

We notice from the figure that population density has an effect on the probability of receiving a news click. The figure shows that users from areas with high population density are more interested in news and are 20% more likely to click on news results.

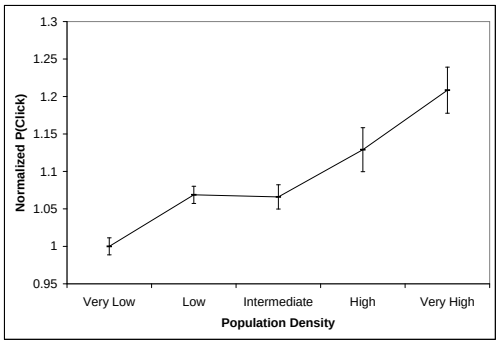

Figure 4: The probability of receiving a news click for a given query for different bins of population density.

5.3 User/Query Related - Distance

We hypothesized that distance might have an effect on the likelihood of a click on a news story. For example, it could be the case that people are most interested in news that takes place close to them, and less interested in stories about places further away. However in the absence of a news query classifier we could see a second phenomenon dominating: queries for locations nearby may commonly be for local businesses and resources, while queries for distant locations may be dominated by news intent.

For our experiments, we look at the distances between the locations specified in queries, and the locations of the IP addresses they are issued from. We try to find out the effect of the distance on whether the query receives a news click or not.

Figure 5 shows how the probability of receiving a news click for a given query is affected by the distance between the user location and the location specified in the query. The horizontal axis represent different bins of user location / query location distance. The range of each bin is shown in Table 5. The vertical axis represents the probability of receiving a news click for each distance bin.

We notice that queries are less likely to receive news clicks when the distance is small. As the distance increases, queries become more likely to receive news clicks. We notice from the figure that the difference between the click probabilities for short and long distances is that very long distances between user and query location lead to increases of about 40% in likelihood of clicking on a news result. Very close locations have the lowest probability of resulting in a news click. This may be because of a the confounding effects of local business search and interest in nearby news. A reliable news classifier could help us tease out these effects in future work.

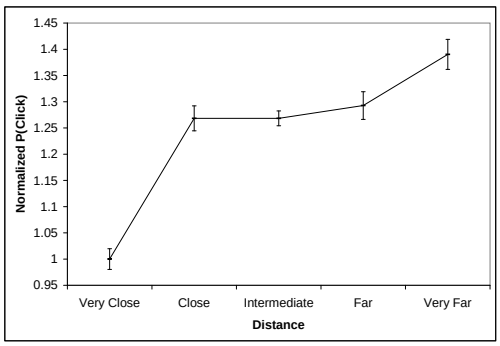

Figure 5: The probability of receiving a news click for a given query for different bins of user location / query location distance.

6. PREDICTING NEWS CLICKS

We have presented descriptive statistics relating individual predictors and the probability of click. In this section, we propose modeling the relationship between clicks and sets of geographic features pertaining to both the query and the user. This analysis allows us to discover important conjunctions of geographic features which are difficult to detect

Bin	Range
Very Close	< 10
close	10 - 500
Intermediate	500 - 5000
Far	5000 - 10000
Very Far	> 10000

Table 5: Distance bins in km

when inspecting predictors individually.

In order to consider several features, we model the experiment as a pattern classification task. Specifically, we would like to predict whether a particular query issued by a particular user is likely to result in a news click or not. We treat the set of geographic features as input and the binary click signal as the target.

In addition to allowing us to identify relationships between geographic features and clicks, this analysis could be used in the decision-making component of a search engine. For example, if our model suggests that a particular query is likely to receive a news click, the search engine may include more news results.

6.1 Classifiers

We use gradient tree boosting (Treenet) [6, 7] , and support vector machines (SVM) [10] for building our supervised learning models.

We use 60% of the data as training data, and 40% as testing data. Queries in training and testing sets were sampled from different periods of time. 50% of the data represent queries that received news clicks and the other 50% represent queries that did not receive news clicks. We train different supervised learning classifiers on the training data, and use the resulting models to predict news clicks for queries in the test data.

The model assigns a value in $[-1, 1]$ to each query. The closer the value to 1, the more likely the query will receive a news click. The closer the value to -1, the less likely the query will receive a news click.

6.2 Baselines and Geographic Features

We compare our results to two baselines. The first is a random baseline that predicts news clicks randomly with probability p, where p equals the total number of queries that received news clicks in the data set divided by the total number of queries in the data set.

The second baseline uses the supervised learning classifiers described above, and a set of non-geographic historical search and click related features to predict news clicks. These features are motivated by previous work on learning to predict when to integrate news content into web results [5] and represent a strong baseline.

The baseline features are shown in Table 6, while the geographic features are shown in Table 7.

Feature	Importance Score
loc_word_click_score	100.0
loc_type	40.9
loc_conf	33.1
dist	24.7
pop_density	18.6
same_country	5.9
same_state	5.4

Table 8: Geo features ranked by importance

6.3 Results

We evaluate our results using precision and recall.

Figure 6 shows the recall vs. precision for random classifier, baseline with non-geographic features, treenet with geographic features only, and treenet with baseline and geographic features. We see that the geographic features are better at predicting news clicks than the baseline features. When we combine geographic and baseline features, we have a model which is highly accurate at predicting clicks on news results.

Figure 7 compares treenet and SVM classifiers. We see that the choice of machine learning algorithm does not have a significant impact on results, but boosted decision trees are slightly more accurate in the high-precision low-recall range.

An important question when building a model combining different features is how much of a contribution each one makes. Table 8 shows the geographic features ranked according to importance. We see that the click probability of individual place-names is the most important feature, suggesting that identifying individual places as newsworthy is important. In addition, the type of location is also important, and we saw in Section 5 that a query containing a country, for example, is much more likely to result in a news click than a query containing a zip code or a town name. Population density is also predictive, suggesting that the demographics of a user's location has some impact on their likelihood to click on news results, even when we have already taken the type of location and distance into account.

7. CONCLUSIONS

We have presented an analysis of the relationship between geographic information—exhibited by the searcher's location and the searcher's query—and the likelihood of clicking news results. Queries containing a location were up to twice as likely to lead to a click on news results. When studying predictors individually, we found that the type of the geographic entity searched for correlated strongly with the probability of click. Predictors based on the user's geography seemed to provide some correlation with the probability of click. These results were confirmed when we experimented with combining geographic predictors. Our results suggest that search engines serving news results should take into account geography; the user's location, the query location, and the distance between them all have an impact on the likelihood of a user to click on a news search result. While our study was restricted to considering queries leading to clicks on news results, it provides evidence that the user and query

Feature Name	Description
news_searches_crnt	number of news searches for this query on the current day
web_searches_crnt	number of web searches for this query on the current day
news_web_crnt	news searches/web searches for the current day
news_searches_prv	number of news searches for this query on the previous day
web_searches_prv	number of web searches for this query on the previous day
news_web_prv	news searches/web searches for this query on the previous day
word_news	Whether the query contains the word "news"

Table 6: Baseline non-geographic features

Feature Name	Description
loc_conf	Location confidence
word_click_score	Location Word/Click probability score
loc_type	Location type
pop_density	Population density of user's location
dist	Distance between the user's location and the location specified in query
same_country	User and query location in the same country or not
same_state	User and query location in the same state or not

Table 7: Geo features used for training the news click prediction model

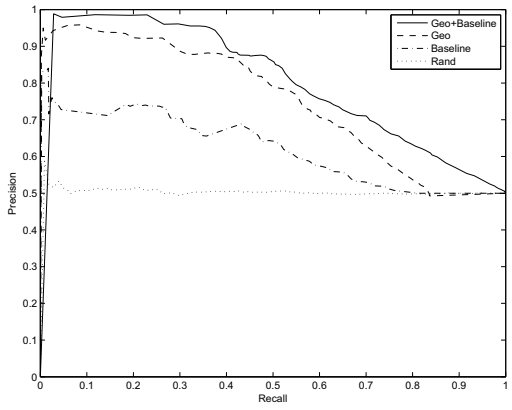

Figure 6: Recall vs. Precision for random classifier, baseline with non-geographic features, treenet with geographic features only, and treenet with baseline and geographic features.

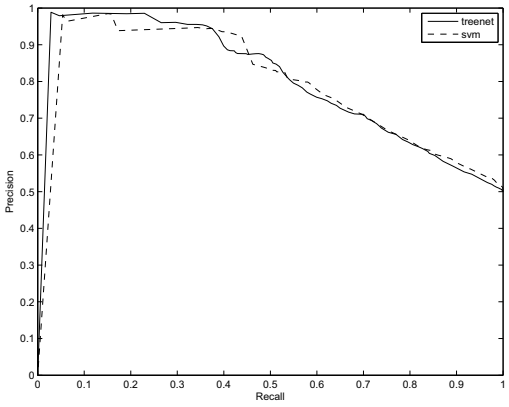

Figure 7: A comparison between treenet and SVM.

location can be important factors in assessing user intent, and this may also be applicable to other web search tasks.

8. ACKNOWLEDGMENTS

Hughes Bouchard, Jean-François Crespo, Remi Kwan, Rajesh Parekh, Jignashu Parikh and Pavel Berkhin originally proposed a news vertical selection algorithm and influenced the features used in the baseline algorithm in this paper; patent under review ("Predicting Newsworthy Queries Using Combined Online and Offlien Models", Attorney Docket No. YAH1P140, Y04164US00).

9. REFERENCES

[1] J. Allan. Hard track overview in trec 2004: High accuracy retrieval from documents. In *The Thirteenth Text Retrieval Conference*, 2004.

[2] L. Backstrom, J. M. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *WWW*, pages 357–366, 2008.

[3] S. Chien and N. Immorlica. Semantic similarity between search engine queries using temporal correlation. In A. Ellis and T. Hagino, editors, *WWW*, pages 2–11. ACM, 2005.

[4] W. Christaller. *Die zentralen Orte in Süddeutschland.* Gustav Fischer, Jena, 1933.

[5] F. Diaz. Integration of news content into web results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 2009.

[6] J. Friedman. Greedy function approximation: the gradient boosting machine, 1999. Technical report,

Stanford University.

[7] J. Friedman. Stochastic gradient boosting, 1999. Technical report, Stanford University.

[8] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel. Analysis of geographic queries in a search engine log. In *LocWeb*, pages 49–56, 2008.

[9] GoogleTrends. http://www.google.com/trends, 2008.

[10] T. Joachims. Making large-scale svm learning practical, 1999. Advances in Kernel Methods - Support Vector Learning MIT-Press.

[11] R. Jones, W. V. Zhang, B. Rey, P. Jhala, and E. Stipp. Geographic intention and modification in web search. *International Journal of Geographical Information Science (IJGIS*, 22(3):229–246, 2008.

[12] Y. Li, N. Stokes, L. Cavedon, and A. Moffat. Nicta i2d2 group at geoclef 2006. In *CLEF*, pages 938–945, 2006.

[13] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. Steward: architecture of a spatio-textual search engine. In H. Samet, C. Shahabi, and M. Schneider, editors, *GIS*, page 25. ACM, 2007.

[14] J. Liu and L. Birnbaum. Localsavvy: aggregating local points of view about news issues. In *LocWeb*, pages 33–40, 2008.

[15] A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):765–772, 2006.

[16] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *World Wide Web Conference (WWW 2006)*, 2006.

[17] R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang. The design and implementation of spirit: a spatially aware search engine for information retrieval on the internet. *International Journal of Geographical Information Science*, 21(7):717–745, 2007.

[18] S. Riise, D. Patel, and E. Stipp. Geographical location extraction. *US Patent 7,257,570*, 2007.

[19] D. Rose and D. Levinson. Understanding user goals in web search. In *WWW 2004*, 2004.

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

[21] C. Sheng, W. Hsu, and M.-L. Lee. Discovering geographical-specific interests from web click data. In *LocWeb*, pages 41–48, 2008.

[22] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: a new view on news. In W. G. Aref, M. F. Mokbel, and M. Schneider, editors, *GIS*, page 18. ACM, 2008.

[23] M. J. van Kreveld, I. Reinbacher, A. Arampatzis, and R. van Zwol. Multi-dimensional scattered ranking methods for geographic information retrieval. *GeoInformatica*, 9(1):61–84, 2005.

[24] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[25] Z. Zhuang, C. Brunk, and C. L. Giles. Modeling and visualizing geo-sensitive queries based on user clicks. In *LocWeb*, pages 73–76, 2008.