# Geographic Features in Web Search Retrieval

Rosie Jones
Yahoo! Labs
3333 Empire Ave
Burbank, CA 91504
jonesr@yahoo-inc.com

Ahmed Hassan*
U. Michigan Ann Arbor
Dept. of EECS
2260 Hayward Street
Ann Arbor, MI 48109
hassanam@umich.edu

Fernando Diaz
Yahoo! Labs
1000 Rue de la Gauchetiere,
Suite 2400
Montreal, QC
diazf@yahoo-inc.com

## ABSTRACT

We conduct large-scale search engine relevance experiments, using the 12% of queries that contain placenames, matching the placenames to places in the documents, and examining the impact of geographic features on web retrieval relevance. Specifically we examine distance between query and document place-names mentioned, noting that when a document has multiple places (which we observe in 82% of documents) we must choose a function over those multiple places. We find that the minimum distance between the document locations and query location is the strongest geographical predictor of document relevance, and that combining geographic features with text features gives us a 5% improvement in relevance over using text features alone.

## Categories and Subject Descriptors

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

## General Terms

Experimentation

## Keywords

geographic information retrieval

## 1. INTRODUCTION

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 [8] retrieved documents by using a relevance score which linearly combined textual and geographic similarity. Purves *et al.* [6] 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. [3]) has used geographic term expansion on the queries and documents and then conventional term matching using BM25 on the resulting expanded texts. Our work differs in that we look at geographic features of the document, the query, and the document-query combined, and train boosted decision trees to learn weights combining textual and geographic similarity. We train a relevance model with both BM25 and geo-spatial features as inputs, and use the learned weights to predict relevance and perform ranking.

In data mining, Mei *et al.* use the geography of weblog authors in order to model spatial patterns of news topics [5]. Mehler *et al.* use locations mentioned in documents to construct spatial models of people mentioned in documents [4]. Zhuang *et al.* use click information in order to model the geographic intent of a searcher and query [9].

## 2. DATA

We sampled queries according to how often they were issued to the Yahoo! search engine, then sampled 532 of them that contained a place-name. For each query we obtained 5 or more documents which were returned by the Yahoo! search engine, giving us a total of 10,588 query-document pairs, with a total of 10,394 unique documents since some documents were returned for more than one query. We obtained editorial relevance judgments for query-document pairs as Perfect, Excellent, Good, Fair, or Bad. Our editors are professional and trained for the task, but the judgements are not specifically targeted for geographic relevance. Editors are instructed to consider relevance for the typical user. This means that we do not capture the way relevance may vary based on the user location.

We extracted place-names from queries using a proprietary black-box location extractor, and similarly extracted place-names from the documents. While these extractors are imperfect, our results will give a lower-bound on performance which could be obtained using more accurate place-name identification.

We found that 68% (7033/10394) of documents contained the place-name found in the query. This is not 100%, since documents may mismatch on ambiguous place-names, or may have matched on anchor text, or may be missed by our location extractor. For documents which did not match the place-name in the query, the average distance of the nearest place was 1,850 km, though 387 had the closest place within 50km. A further 17% (1814/10394) did not contain any place-name recognized by our place-name extractor. 82% (8508/10384) contained multiple place-names (which could be the same place-name repeated).

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


*GIR'08,* October 29–30, 2008, Napa Valley, California, USA.

| | FBM25 | GEO | FBM25+GEO |
|---|---|---|---|
| DCG1 | 3.9 | 2.9 | 4.4 |
| DCG2 | 6.0 | 4.9 | 6.5 |
| DCG3 | 7.5 | 6.4 | 8.0 |
| DCG4 | 8.8 | 7.8 | 9.4 |
| DCG5 | 10.1 | 9.1 | 10.6 |

**Table 1: Comparison of DCG using text features (BM25) and text features plus geo features.**

## 3. GEO FEATURES

We used extracted geographic information in the query and retrieved documents. We used this information to construct a set of ranking features. These features fall into three categories: document features, query features, and query-document features. Document features are derived from the locations mentioned in the document, regardless of the query. These include the number of locations in the documents. Query features are derived from the locations mentioned in the query. These include the location mentioned in the query. Finally, query-document features are derived from the interaction between the locations mentioned in the document and those mentioned in the query. In our experiments, we use the maximum distance from the query location to any location mentioned in the document, and the minimum distance from the query location to any location mentioned in the document.

## 4. EXPERIMENTS

We are interested in examining the role of geographic features when used in conjunction keyword based features. We described geographic features in the previous section. For a keyword based feature, we used a version of BM25 suitable for structured HTML documents [7]. These features were combined using a gradient-boosted decision tree regression with relevance grade targets [1]. Documents are ranked by their predicted grade and evaluated using a variant of discounted cumulative gain (DCG) [2]. We define the DCG at rank $k$ for a single query as,

$$\mathrm{DCG}_q = \sum_{r=1}^{k} \frac{g(d_r)}{\log_2(r+1)} \quad (1)$$

where $d_r$ is the url of the document at rank $r$, and $g(d_r)$ is the grade of that document. We average the DCG for all queries in our evaluation.

We present results are shown in Table 4. We see that adding geo features improves results over text features (FBM25) by about 5%.

## 5. DISCUSSION

We found that the significant geographic features in order were (1) number of locations in the document (2) minimum distance between query-location and document location (3) location type (city, state, country, etc) (4) maximum distance between query location and document location.

We find that the introduction of geographic features slightly improves performance for those queries containing geographies. These results, while preliminary, suggest that even the simple geo-sensitivity classifier we use allows us to improve performance on a subset of queries. Our gains can be increased in two ways: by improving the classification of geo-sensitive queries and by using better ranking features. One feature of particular interest would be the searcher's location. This would provide a geographic reference for geo-sensitive queries lacking explicit geographic intent.

## 6. REFERENCES


[1] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[2] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, New York, NY, USA, 2000. ACM.

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

[4] 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.

[5] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 533–542, New York, NY, USA, 2006. ACM.

[6] 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.

[7] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49, New York, NY, USA, 2004. ACM.

[8] 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.

[9] Z. Zhuang, C. Brunk, and C. L. Giles. Modeling and visualizing geo-sensitive queries based on user clicks. In *LOCWEB '08: Proceedings of the first international workshop on Location and the web*, pages 73–76, New York, NY, USA, 2008. ACM.