Analyzing geographic queries

Mark Sanderson
Department of Information Studies
University of Sheffield
Western Bank, Sheffield, UK
+44 114 22 22648
m.sanderson@shef.ac.uk

Janet Kohler
Department of Information Studies
University of Sheffield
Western Bank, Sheffield, UK

ABSTRACT

The aim of this study was to analyze the 2001 Excite query log to investigate the extent and variation of Web queries containing geographic terms. In particular, an investigation into what people search for when they use geographic terms, the ways in which they describe a geographic location, the terminology used to find geographically related information and the structure of users' queries when looking for geographically related information on the Web. This study also attempted to determine how geographically related queries differ from other queries. Geographically related queries formed nearly one fifth of all queries submitted to Excite, the terms occurring most frequently being place names. Geographic queries were also shown to be longer than average and the association of two or more terms within geographic queries was found to be high.

1. INTRODUCTION

Web users searching for information about locations, institutions and many other topics often require information that is geographically specific. It has been suggested [1] that users will focus their Web query by using geographic terminology such as place names and spatial prepositions (e.g. "near", "between" and "north of") to associate a topic with a location. When the name of a place is typed into a typical search engine, Web pages that include that name in the text will be retrieved but most likely, not places that are within or close to that specified place. In order to understand the potential of improving the functionality of search engines in relation to geographic search, it is necessary to understand what people search for and how they structure their queries. There are an increasing number of studies available of how people formulate Web queries and how they modify those queries during a search. However, to the best of our knowledge, no study exists on the use of geographic terminology within Web search queries. It is acknowledged by [3] that the potential benefit of Web query log studies to IR system developers, users, and Web site classifiers and designers could be considerable. It is therefore of interest to assess the manner in which users formulate their queries to find information on geographic and related topics, in order to gauge whether the interpretation of queries by search engines could be improved.

The aim of this initial study was to analyze the 2001 Excite query log to investigate the extent and variation of Web queries containing geographic terms. In particular it is an investigation into what people search for when they use geographic terms, the ways in which they describe a geographic location, the terminology used to find geographically related information and the structure of users' queries when looking for geographically related information on the Web. This study also attempts to determine how geographically related queries differ from other queries.

This poster describes the methodology and results of the query log study followed by conclusions and possible future work. Due to space limitations a review of past work is omitted, the interested reader is directed to the dissertation work of the 2nd author that this poster was derived from [2].

2. METHODOLOGY

A log of 1,025,910 queries was made available for this study from the Excite search engine. It contains an anonymised user ID, a time stamp, and the query text itself. It only contains queries where client machines accepted cookies from the Excite server. For the purpose of this study a *geographic query* was defined as a query which included at least one of the following types of *geographic terms*: place names e.g. Houston, Texas, US; other locators e.g. postcode, ZIP code; adjectives of place e.g. American, international, western; terms descriptive of location e.g. state, county, city, site, street; geographic features e.g. island, lake; and directions e.g. north, south. Owing to the size of the query corpus, to better enable human analysis of the data, a random sample of the corpus was taken.

3. RESULTS

A random sample of 2,500 queries was extracted from the table of unique queries. These were analyzed by a human classification method for place names and other geographic terms and a new data set formed of these random geographic queries for further analysis. The results are shown in the table below.

3.1 Numbers of geo-queries

Of the 2,500 queries, 18.6% contained a geographic term; and 14.8% held a place name. This compares with the findings of [4] who identified 19.7% of their random sample of 2,453 unique queries as containing "people, places or things". Although these figures are not directly comparable, the proportion of "places" found in the current study does appear to be consistent with the figures of [4]. Of the 464 terms identified as including a geographic term, nearly 80% of the queries that had geographic content had a place name as one of the terms in the query.

Variables	No. of qrys	% of the geo qrys	% of full sample
Queries with place names	369	79.5%	14.8%
Queries with other geo terms	189	40.7%	7.6%
Queries with any geo term	464	100.0%	18.6%

3.2 Categorizing geo-queries

A classification of queries was conducted using headings adapted from [4] (see table below). Searches about places were

the top form of geo-query with business related searches the next most common. Due to the adaptation of the categories, direct comparison between this and the past study was unfortunately hampered. The emphasis of "place queries" was un-surprisingly greater than that of the previous study of general Web queries. Comparing the rest of the categories with [4] work revealed relatively small differences between numbers of queries across each of the categories.

Rnk	Category	Prop.	Rnk	Category	Prop.
1	Places	15.9%	9=	Unknown	5.0%
2	Commerce & services	14.7%	11	Computers & internet	3.9%
3	Rec. & sport	8.8%	12	Health	2.6%
4	Education	7.8%	13	News & media	2.2%
5	Tourism	7.3%	14=	Society & culture	1.9%
6	Travel	6.7%	14=	Sci. & tech.	1.9%
7	Government	6.0%	16	Entertainment	1.7%
8	Arts	5.6%	17=	Employment	1.5%
9=	Sex & porn.	5.0%	17=	People	1.5%

3.3 Length of geo-queries

The following table shows the results of the analysis of the 464 geographic queries for length. It is evident that the number of geographic queries having only one term is barely one third of the number of all the unique queries having one term: 9.4% compared to 26.9%. The number of queries containing two terms is nearly the same, but the number of queries using three or more terms is almost 50% higher for geographic queries than for all queries. The combination of fewer one term queries and more queries containing three or more terms results in the average length of the geographic query being 25% higher than the average length of all unique queries, at 3.3 terms per query against 2.6 terms per query.

Variables	Place names	Non-place names	Any geo-term	All queries
Average terms per query	3.4	3.3	3.3	2.6
1 term	9.5%	7.9%	9.4%	26.9%
2 terms	27.6%	28.0%	30.4%	30.5%
3 terms	24.7%	30.2%	25.4%	22.6%
4 terms	16.8%	15.3%	15.7%	8.3%
5+ terms	21.4%	18.5%	19.0%	11.7%

That geographic queries appear to be longer than average can be attributed to a range of factors: geographic queries often take the format "object in place name" or "Where is…."; place names are often composed of two words e.g. "Las Vegas"; sometimes a region is specified in addition to the name of a place, especially where more than one place with the same name exists; if a spatial term is used it is almost always associated with a place name.

3.4 Spatial relationships

The one million queries were searched for terms indicating a spatial relationship. The complete analysis is described in [2]; the most frequent relationship indicators are described here.

Of the 9,960 queries (0.96% of the total data set) containing the word "in", 5,725 also contained a place name. In most of the queries "in" directly preceded (modified) a place name. There were 821 queries containing "at", of which 274 modified a place name. The spatial term "from" occurred 217 (out of 749) with a place name: generally used in the sense of something originating from, e.g. "famous people from philadelphia" or "flights from denver".

The directions "north", "south", "east" and "west" were mainly used as parts of place names, companies or institutions. It was striking that counter to the speculation of [1] these terms were not used in a directional sense, for example "north of Las Vegas". These directions were occasionally used to specify part of a larger area, for example a county or a state. Just over one hundred queries specified "near" or "surrounding" relationships looking for something close to a place, where the place is a center of population, political region, country, institution, or the like. Of these, sixty modified a place name.

4. CONCLUSIONS & FUTURE WORK

The results indicate that geographically related queries are a significant sub-set of the queries submitted to a search engine. The topic area covered by such queries appeared only slightly different to those areas in standard Web queries. Geo-queries are noticeably longer than the notoriously short typical Web query.

A more extensive study of a wider query corpus is planned examining geographic query refinement. In addition, the small number of spatial relationships entered will be studied: examining if the low number is due to a lack of user need for searching on such relationships; or on a user perception that search engines are incapable of dealing with such queries.

5. ACKNOWLEDGMENTS

Thanks to Amanda Spink for providing the 2001 Excite query log and to Excite.com for making the data set available in the first place. The work was partially supported by the SPIRIT project, funded by the EU.

6. REFERENCES

[1] Jones, C., Purves, R., Sanderson, M. (2002) Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT Project in the *proceedings of 25th ACM Conference of the Special Interest Group in Information Retrieval*

[2] Kohler, J. (2003) Analysing search engine queries for the use of geographic terms. Masters Dissertation, *University of Sheffield*

[3] Spink, A., Wolfram, D., Jansen, B.J. & Saracevic, T. (2001). Searching the Web: The Public and Their Queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.

[4] Spink, A., Jansen, B.J., Wolfram, D. & Saracevic, T. (2002). From E-sex to E-Commerce: Web Search Changes. *IEEE Computer*, 35(3), 107-109.

Columns on Last Page Should Be Made As Close As Possible to Equal Length