Search Words and Geography

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

Yu Han
Department of Information Studies
University of Sheffield
Western Bank, Sheffield, UK

ABSTRACT
In this paper, we present a preliminary study of geographic query words, which users' tend to re-use. The categories of the words demonstrate that geographically related words take up the largest proportion of all repeated words. These geo-words refer to a range of spatial areas. In addition, it was found that different geo-word types are re-used in different ways by users.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]:

General Terms
Measurement

Keywords
Geography, re-used query words

1. INTRODUCTION
With the advent of large scale logging of user's activities on search engines, analysis of those logs has produced much valuable information for the information retrieval community. The vast majority of the studies – as typified by the extensive studies of Jansen and Spink (most recently summarised in their 2006 paper) [1] – have focused on the topic of the whole query. Far less focus has been on the study of individual search words.

As part of a study of a large query log gathered using a voluntary opt-in feedback system from users of a large search engine, an analysis of the words that users use in queries was conducted. In particular, the words that users tend to re-use in different queries were examined. The log (covering queries issued to the engine for 4 weeks starting on 9th January, 2006) was composed of 1 million queries with 2.3 million search result clicks. The text within the queries was indexed and all words and phrases occurring more than 100 times in the data set were isolated and the queries issued that contained those words were examined. In order to keep the data to be examined manageable, only those words that were used by more than 100 users in the log were retained.

The 500 most re-used words from this analysis were extracted and analysed by hand to determine subject categories for those words [2]. The categories were assigned by one of the authors of this paper and then re-checked by the other. Details of this analysis are shown in the Table 1. As can be seen from the table, geography related words constitute the largest percentage of all the categories. It was easy to identity such words as they were commonly well known place names and/or abbreviation, or they were postal codes of a US state. This short paper presents an analysis of those geographic words and their re-use in user's queries.

GIR'07, November 9, 2007, Lisbon, Portugal.
ACM 978-1-59593-828-2/07/0011.

2. ANALYSIS
Of the 188 words found to be related to geography. There were 63 words related in some way to countries; 67 were related to smaller areas, such as states, provinces, counties or special areas; 41 words were related to city or city borough.

Note, this distribution of types of geographic words is different from that reported by Zhang et al (2006) [3], where they found city names much more commonly used than country names and country names more commonly than state names. However, the manner in which data was gathered in the two studies was so different that comparisons are hard to make.

The words referring to countries were sub-divided into two classes composed of 33 and 30 words respectively: the two letter codes used in domain names to identify a country (e.g. 'za' for South Africa, 'pt' for Portugal); and those words referred to a country by its full name. As might be expected when working with such data, some of the words (17) were found to be ambiguous, such as the word "in", which might refer to the abbreviation of the state of Indiana in the US, or the country of India, or simply a preposition. Because this part of geo-words was less than 10%, we chose to ignore such words, leaving 171 for the next stage of analysis, see Table 2.

2.2 Different Areas Re-used Word Model
Since the motivation for the work was a study in the repetition of words in queries, the way in which such words were re-used was examined. In particular, an examination of how such words were re-used by users was explored. The key question was do users tend to re-use words within a short space of time, or do they re-use over a longer time frame. To quantify this issue, the percentage of users who re-used a geo-word on a single day only was calculated. This data was tabulated and is shown in Table 3.

The high values in this table indicate that many users simply used the words in one search session and don't re-use them in subsequent queries. However, across the four classes of words highlighted in the data, between 20% and nearly 30% of users studied in the logs did re-use the words on different days and that re-use appeared to vary across the types of geographic word.

Category	Meaning	No.	%
Activity	Words for searching social activities	5	1.0
Adult	Words related to porn	16	3.2
Arts & Humanities	Words related to society, history and culture	14	2.8
Assistant word	Word does not make clear meaning of query by itself, and prep word	56	11.2
Business & Shopping	Words of famous brand, and shopping related	23	4.6
Computer	Computer terms, such as sql	41	8.2
Education	Words related to education	5	1.0
Entertainment & Recreation	Words related to sport, entertainment, games and music	58	11.6
Geography	**Place names, and abbreviation of place names or postal code**	**188**	**37.7**
Healthcare	Words related to disease, drug or body health	15	3.0
People	People names	5	1.0
Science & Technology	Words related to science and technology	14	2.8
Things	Words implied an object or animal or a thing, meaningless itself	31	6.2
URL	The word for search website	18	3.6
Non-English & Unknown	Not English words and words cannot identify the query	10	2.0

Table 1: Categories of the 500 re-used words

Categories	Number	%
State/Province/County/Area	67	35.6
Country (two letters)	33	17.5
Country (full name)	30	16.0
City/Borough	41	21.8
Ambiguous	17	9.0

Table 2: Categories of geography area

	Repeat Single Day
City	80.2%
Country	79.5%
State	75.7%
Country (2 letter)	70.5%

Table 3: Words used on different days

	Different queries per user
City	1.51
Country	1.57
State	1.58
Country (2 letter)	1.77

Table 4: Different queries per user

As can be seen, those words in the query that referred to a country using its two letter code were the most likely to be re-used by users on different days, the common re-use of words also appears to hold true for words associated with sections of a country such as states. In contrast, the results indicate that users use city or borough names and country names in a much more concentrated way: more often using a word across a series of queries on a single day than using that word again at a later date.

Using a 2 sample un-equal variance t-test, it was determined that the difference in use between "City" and "Country" words was significantly different from the use of "Country (2 letter)" and from "State" words. Additionally, the 2 letter country codes were used in a significantly different way from the "State" words.

An additional examination of the data was conducted to explore the average number of unique queries a user would on average use for each of the words types. The results for this examination are shown in Table 4. As with Table 3, it can be seen that most users query on these words only once. However, of those that re-issue queries, it would appear that the 2 letter country codes were those that were re-used in the most different types of queries, whereas words referring to cities were used in somewhat fewer unique queries. Significance tests were applied to the sets of figures from which the Table 4 averages were derived: significance was observed when comparing city-state, city-country and between all pairs with the 2 letter country codes. It would appear from this analysis also that there was a small but significant difference in the way that different geographical words were used in querying.

3. CONCLUSIONS

From past work, it has been shown that geographic terms in queries are common. In this work, we show that when studying words that are used repeatedly in queries, geographic words again are common. What we also determined however, is that geographic words referring to different types of administrative areas appear to be repeated in different ways. Countries names are used in queries in a more 'bursty' manner, with users repeatedly searching for queries related to these areas within a single day. In contrast, two-letter country codes and words referring to the states or counties within countries appear to be more used more often by users in different queries on different days. Quite why this somewhat un-expected quality of the data is occurring is important, but one that will be left for future work.

4. REFERENCES

[1] Jansen, B. J., Spink, A. *How are we searching the World Wide Web? A comparison of nine search engine transaction logs*. Inf. Process Manage. 42(1):248-263, 2006

[2] Pu Hsiao-Tieh, Chuang Shui-Lung, Yang Chyan. *Subject categorization of query terms for exploring Web users' search interests*. JASIST, 53(8):617-630, August 2002

[3] Zhang, V. W., Rey, B., Stipp, E., and Jones., R. *Geomodification in query rewriting*. In Proceeding of the 3rd workshop of Geographic Information Retrieval. Seattle, WA USA, 2006