A Study of the Tourism Web Coverage in Switzerland

Ramya Venkateswaran

University of Zurich
Department of Geography
Winterthurerstr 190
CH-8057 Zurich, Switzerland
Email: ramya@geo.uzh.ch

1. Introduction

This paper discusses experiments that were performed to understand the geographic and linguistic coverage of web resources focusing on tourism-related themes in Switzerland. The research was prompted by the observation that studies in geographic information retrieval (GIR) and volunteered geographic information (VGI) commonly assume web coverage to be homogenous across geographic space, themes, and languages. There are, however, strong hints that this assumption is unfounded (Pasley et al. 2008).

The goal of studying the geographic web coverage is one of the preliminary steps in generating (geographic) data from the web that can be used as valid information. An idea on how well certain areas are geographically covered by information available on the web, their frequency and patterns that emerge from this data collection help in the decision of preselecting web data for further investigation. For this experiment the language is also considered as coverage varies greatly on the tongue of the place.

Ad hoc tourism information is readily available on the web in the form of pages that contain news, lists, catalogue, reviews, blogs and multimedia content. All this provides a vast playground for tourism as a use case for generating geographic information from the web.

The key questions driving this research are: 1) What is the geographic distribution of web coverage for tourism-related themes? 2) How does language affect web coverage?

2. Related work

This work involves concepts and techniques from geographic information retrieval (GIR) (e.g. Purves et al. 2007, Overell et al. 2008), but due to lack of space the review of related work is limited to references immediately relevant to the research.

Volk (2009) accumulated counts for occurrences of mountain names from the yearbooks of the Swiss Alpine Club for a period of time. The method proposed by Pasley et al. (2008) involves gathering counts from the web using the Ordnance Survey 50K gazetteer for Great Britain and a search engine API. Place names can also be mined from a document using trigger phrases and in turn from the web as described by Twaroch et al. (2008).

The study reported here extends over the work by Volk (2009) by using the web rather than a closed corpus of text documents. It makes use of the methodology proposed by Pasley et al. (2008) but extends over both their work and Volk's work by examining the effects of using different gazetteers as well as languages.

The current work uses web counting as a measure of the coverage. There are many other approaches as well. For example, Hecht and Gergle (2010) measure the diversity for the Wikipedia corpus in 25 different languages by counting the concepts that were included and the ways in which these concepts were described.

3. Experiments and Results

3.1 Datasets and Experiment Design

The experiment consists in examining the geographical web coverage for tourism-related themes by gathering of web counts for a preselected number of places, tourist places, and mountain features. In order to select place names, two datasets containing toponym information were use: SwissNames by Swisstopo and the Points of Interest (POI) dataset by Tele Atlas.

SwissNames is maintained by the Swiss Federal Office of Topography (Swisstopo). The dataset contains 155,571 names and 62 categories. The dataset is at a resolution of 1:25,000 and contains other essential pieces of information like the coordinates, altitude, 'Gemeinde' (town) name and canton name.

Tele Atlas POI dataset contains 54,912 points of interest in 50 categories. The POIs are attributed with important information like name, address and other details.

Switzerland has four official languages; German, French, Italian and Romansh. Language plays an important part as content varies depending on the local tongue of the place. Hence it was decided to also examine all the web counts in four languages; German, English, French and Italian. Romansh was not selected as number of Romansh speakers are few as compared to the other languages. Since it was a tourism based use case, English was also selected as one of the four languages.

The test was to calculate the number of hits given by the Yahoo! Search BOSS API for the places along with phrases in the four languages; The search keywords used were *'Place_name' Schweiz tourismus, 'Place_name' Switzerland tourism, 'Place_name' Suisse tourisme* and *'Place_name' Svizzera turismo*. Place name and country were chosen after looking at the analysis of flickr images by Hollenstein (2009). The search was not restricted to the top level domain '.ch', since many tourism websites are hosted under .com.

3.2 Settlements from SwissNames

The first experiment was carried out for 757 settlements in Switzerland. These places were selected based on their size in terms of the number of inhabitants as given in Volk (2009) and the metadata by Swisstopo. Places with fewer than 2000 inhabitants were not selected, resulting in 757 settlements. Out of these, 28 places were eliminated because of geo/non-geo ambiguities that arose causing the counts to be artificially high. Figure 1 shows the resulting web counts of ***Swiss settlements*** on a logarithmic scale, sorted on all language web counts.

3.3 Tourist destinations from Tele Atlas POI

For this experiment a list of 787 tourist destinations was made using towns or points of interest that were explicitly marked "Important Tourist Attraction" in the Tele Atlas database. A similar experiment like the one above was then performed. Almost all the entries were deemed to be genuine as no ambiguities like in the previous experiment were discovered. Figure 1 shows the resulting web counts of ***important tourist attractions in Switzerland*** on a logarithmic scale, sorted on all language web counts.

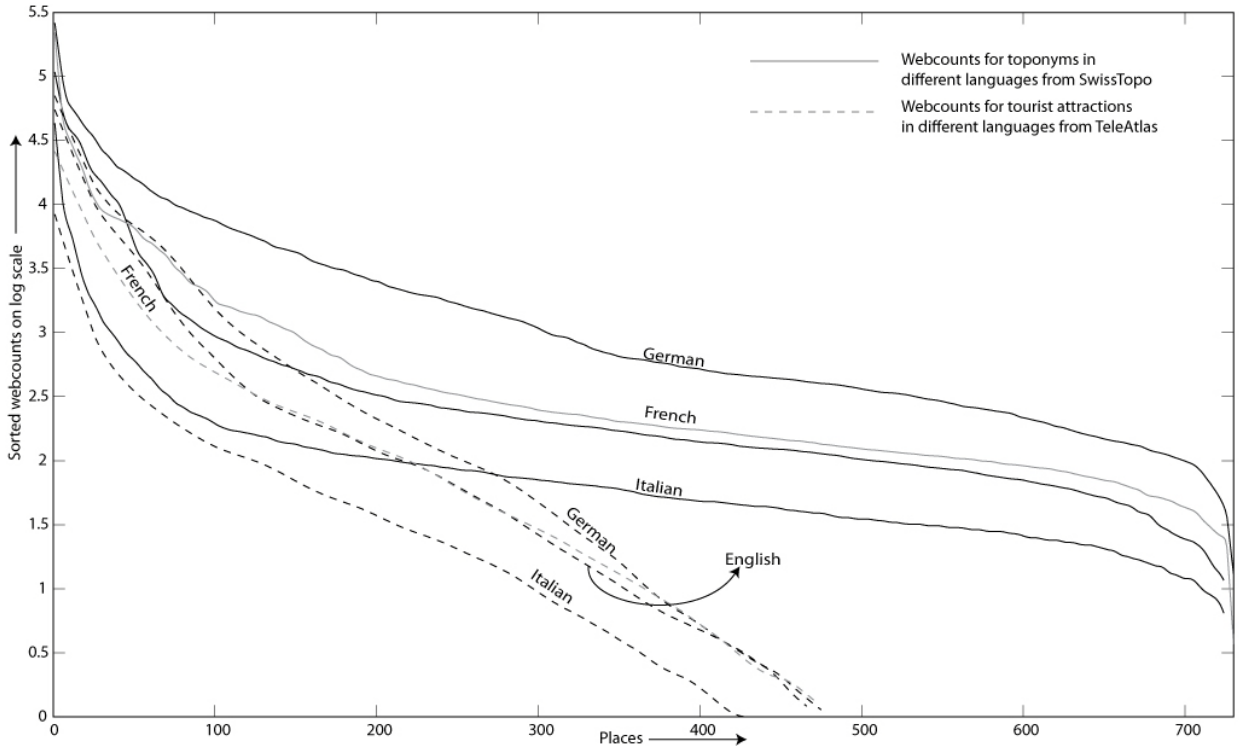

Figure 1. Plot showing linguistic differences for 729 toponyms and tourist attractions from SwissNames and Tele Atlas POI data sources. The lighter lines show the difference between French and English counts.

3.4 Coverage maps

Figure 2 shows the tourism web coverage on the map of Switzerland with mean counts for four languages. Jenks' natural breaks classification method has been used to plot the points on the map.

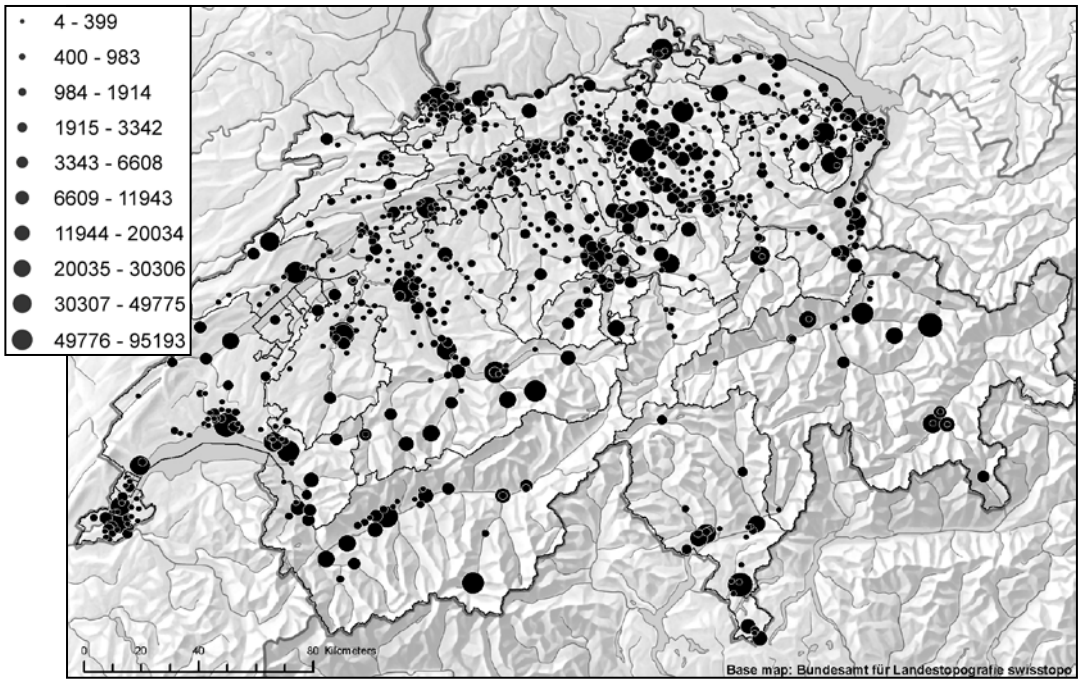

Figure 2. Map showing the Tourism web coverage of Switzerland along with counts for the toponyms from SwissNames (mean counts; experiment 1)

From the coverage map (Figure 2) it is clear that the coverage is not homogeneous across geographic space. The counts are not only affected by the population density, but also by the place's touristic importance. In some cases the mountains also have high counting owing to tourist and sport activities.

4. Results and Discussion

Looking at Figure 1, German (which is the most frequently spoken language in Switzerland) seems to clearly top the list of the most number of counts (Table 1) for many places, as compared to English, French and Italian. French turns out on second place, which might be surprising as English is more frequently used on a global level. The English counts, however, could be skewed mainly because of the use of English transliteration and transcription (e.g. using 'Zurich' for 'Zürich').

Table 1. SwissNames and Tele Atlas list of places with the counts of highest frequency. (The highest counts are the sum of the number of times the web counts for a particular language was the highest among the other three.)

No. of highest counts	Keywords in different languages for SwissNames	No. of highest counts	Keywords in different languages for Tele Atlas
555	German	290	German
144	French	112	French
16	English	78	English
14	Italian	22	Italian
		185	0 counts
729	**Total**	**787**	**Total**

Table 2. SwissNames list of places with top 10 counts.

Top 10 counts for English	Top 10 counts for German	Top 10 counts for French	Top 10 counts for Italian
Bern	Bern	Genève	Lugano
Lausanne	Luzern	Neuchâtel	Locarno
Davos	Freiburg	Lausanne	Bellinzona
Lugano	Zürich	Montreux	Chiasso
Basel	Basel	La Chaux-de-Fonds	Mendrisio
Zermatt	Schaffhausen	Fribourg	St. Moritz
Grindelwald	Winterthur	Bern	Zürich
Sion	Solothurn	Nyon	Zermatt
Interlaken	Chur	Vevey	Davos
St. Moritz	Appenzell	Yverdon-les-Bains	Ascona

Note that normalization has not been applied in this study. This caused, for instance, the number of counts to drastically increase to 111,139 for 'Zurich', as compared to 28,227 for 'Zürich'.

Looking at Figure 1, the results for Tele Atlas POI looked very similar (Table 1) to the above, almost like a validation of the first experiment. In this case also German clearly

dominated the number of web counts but contrary to the previous experiment, English did much better. There were many 0 counts and upon further investigation of the dataset, this result could be expected as most tourist places are in the local language (i.e. not English). For example many of the entries in the POI dataset relate to transportation tourist attractions like cable car, mountain train, ropeway etc. These names are given in the local language.

5. Conclusion and outlook

The calculation of web coverage for tourism in Switzerland is part of the bigger master plan to tap tourism information from the web for mobile, location-based services. From the above experiments and discussion, however, it is clear that web coverage cannot be assumed to be homogenous across geographic space, theme and language, as it is often assumed in VGI and GIR studies. For instance, the web counts vastly differ for different languages. German is very well covered, but not so well for Italian, corresponding to the frequency of language representation in the Swiss population. The coverage figure also suggest that more investigation could be done on other variables, such as city size and its link to tourism web content.

These results have not yet been cross-validated with another toponym dataset as the most dense dataset by far is SwissNames, which was already part of the experiment. One option would be to test this with the Openstreet map dataset (switzerland.osm) or with the geonames dataset (geonames.org).

The current experiments were performed only with phrases in different languages. It will be useful to know how many of these are in the local language. It would be a short-sighted approach not to look at pages in these languages.

Acknowledgements

I would like to acknowledge the comments of Robert Weibel and Ross Purves. I would also like to acknowledge Pia Bereuter's help with respect to looking for ambiguities in four languages and Ronald Schmidt for helping me with the maps. The work reported represents part of my PhD project. I am grateful for funding by the Swiss National Science Foundation through project no. 200020-120256.

References

Bundesamt für Landestopografie swisstopo http://www.swisstopo.admin.ch

Hecht, B. and Gergle, D. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *CHI 2010: Proceedings of the 28th international conference on Human factors in computing systems.* ACM, New York, NY, 291-300.

Hollenstein, L. 2009. Capturing Vernacular Geography from Georeferenced Tags. MSc Thesis, Department of Geography, University of Zurich.

Overell S, Rüger S. 2008. Using co-occurrence models for placename disambiguation. *International Journal of Geographic Information Science*. 22(3): 265-287.

Pasley R, Clough P, Purves RS, Twaroch FA. 2008. Mapping geographic coverage of the web. In: *GIS '08: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. New York, NY, USA: ACM; :1-9.

Purves RS, Clough P, Jones CB, et al. 2007 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.

Tele Atlas BV 2010 http://www.teleatlas.com/index.htm

Twaroch FA, Smart PD, Jones CB. 2008. Mining the web to detect place names. In: *GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval*. New York, NY, USA: ACM; 2008:43-44.

Volk M. 2009. *How many Mountains are there in Switzerland? Explorations of the SwissTopo Name List.* In: Simon Clematide, Manfred Klenner, Martin Volk (eds.): Searching Answers. A Festschrift for Michael Hess on the Occasion of his 60th Birthday. MV-Verlag.

Yahoo! Search BOSS API http://developer.yahoo.com/search/boss