Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

GeoSiteSearch: A Tool to Map Vietnamese Diaspora by Deducing Geographical Information of Web Pages about Our Lady of LaVang

Madison G. Masten,[1] **Thien-Huong Ninh,**[2] **Nicholas Tran**[3]
[1]Maxar Technologies, Palo Alto, CA 94303
[2]Cosumnes River College, Sacramento, CA 95823
[3]Santa Clara University, Santa Clara, CA 95053

Abstract

We construct a web tool to extract geographical locations from web pages returned by the Google search engine for an arbitrary query and display those locations on an interactive map. The tool was used to track the worldwide Vietnamese diaspora using Our Lady of LaVang as proxy for presence of a Vietnamese community, but it could potentially have other applications.

The Vietnamese Global Diasporic Network

The victory of communist North Vietnam ended the Vietnam War in 1975 and immediately created one of the largest and most tragic forced displacements in history, with at least half of the Vietnamese who fled dying during their escape journeys and the others scattered in different countries throughout the world. Currently the population of this diaspora is estimated to be more than four million. Data about the structure and dynamics of this growing social network remain sparse and come mostly from government statistics with few details about how clusters of overseas Vietnamese have evolved over time.

The recent emergence of data science as a body of new techniques in automatic extraction of patterns and rules from databases allows a new approach to studying the Vietnamese global diasporic network using real-time data available from search engines and online social networking sites. Besides using data independent from those collected by governments, this approach yields a much more dynamic description of the network that may suggest new insights into its evolution and structure and the relationships between its components.

Our Lady of LaVang As Online Proxy

Our Lady of LaVang is a Marian apparition whose appearance was first reported in the late 18^{th} century in Vietnam. Although she was originally depicted with European features (as other Virgin Mary figures are commonly represented), starting around the beginning of the new millennium her statues became Vietnamized, wearing the national costume and a headdress decorated with the stars Vietnamese refugees had used to navigate to freedom. The reinterpreted statues have become common diplomatic gifts exchanged between Vietnamese Catholic communities in different countries. In 2002, six of them were blessed by Pope John Paul II and distributed to Vietnamese Catholic communities on six different continents, and by 2010 Our Lady of LaVang has become a de facto symbol of the global Vietnamese Catholic diaspora (Ninh 2017).

We assume that a reasonably-sized overseas Vietnamese community includes a significant Catholic population (30-40% in the US) whose online presence and activities are likely to include the symbol of Our Lady of LaVang. Thus, retrieving the web pages containing the phrase "Our Lady of LaVang" and deducing their physical addresses could produce a real-time reliable map of the Vietnamese overseas diaspora.

GeoSiteSearch: Geolocating Web Pages

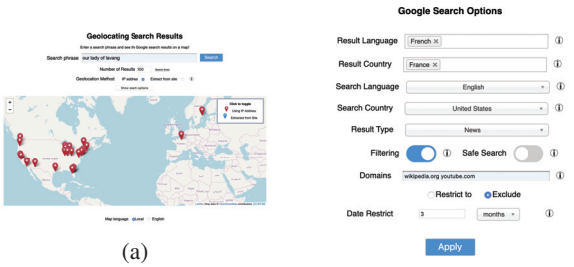
We have constructed a web tool, GeoSiteSearch, to automate this process, and it is available at dh.scu.edu:5000. A Google query is performed on a user-provided phrase ("Our Lady of LaVang" for our purpose), and the geographical locations of the returned web pages are deduced and displayed on a global map.

Although motivated by an application in cultural mapping (Duxbury, Garrett-Petts, and MacLennan 2015), GeoSiteSearch is a general-purpose geographic visualization tool for exploration (Nöllenburg 2007). By attempting to extract addresses directly from web pages, our tool takes a content-based approach to IP geolocation to identify the location of a web page's owner rather than of its host, unlike the usual register-based or measurement-based techniques in recent literature (Center for Applied Internet Data Analysis 2020).

Finer-grained data generated by our tool for cultural mapping is naturally of interests to academics, and it can also potentially be crucial in shaping timely public policies to promote the mapped community's common good. We believe that these benefits outweigh the privacy risk posed by our tool, which is less invasive than the decennial census and other periodic government data collection (The Markkula Center for Applied Ethics 2018).

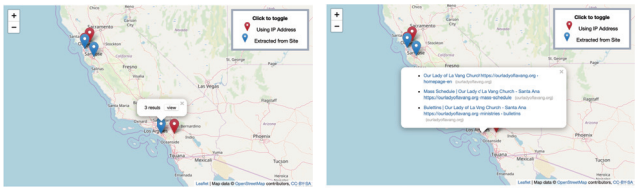
The Graphical User Interface

GeoSiteSearch functions by taking in a search query from the user, as well as the number of results to display, the method of web page geolocation to use, and a number of optional search parameters. The underlying search engine used is the Google search engine, and the additional parameters are a subset of those provided in advanced Google search settings. These options are: result language and country, search language and country, result type, filtering (removes duplicate results), safe search, inclusion and exclusion of specific domains, and date restriction. The user interface for these options is more accessible and flexible than the one provided by Google, which should allow researchers to tailor their queries easily and with very little technical background (see Figure 1). The two available methods for web page geolocation are IP geolocation and extraction of physical addresses from web page text; the color-coded results from one or both of these methods are displayed to the user on an interactive global map. Each marker then allows easy scroll-over access to the results themselves (see Figure 2). Map locations can be displayed in English or in the languages local to each region.



(a) (b)

Figure 1: (a) Layout of the user interface. (b) Interface to Google search options, with sample inputs.



(a) (b)

Figure 2: (a) Display of color-coded results from both geolocation methods, as well as scroll-over interface to result quantities. (b) Expandable popup with full result details, including titles, domains, and links.

Technologies and Data Sources

GeoSiteSearch uses Flask (Pallets Projects 2019) and Python on the backend and vanilla Javascript on the frontend. These technologies were chosen for their simplicity. Interactive mapping is done using the Leaflet Javascript library (Agafonkin and other contributors 2019), and map tiles are provided by OpenStreetMap (local language tiles) and Esri (English tiles) (OpenStreetMap Contributors 2019). The application interfaces to Google search using a modified version of the `googlesearch` Python package (Hseb 2017). For the IP geolocation method, data is provided by extreme-ip-lookup.com, whereas the data used in the address extraction method of geolocation is provided by www.geonames.org. The Python library `BeautifulSoup4` (Richardson 2019) was used for scraping the physical addresses.

Web Page Geolocation Methods

The first method of web page geolocation that we implemented is based on the IP address of the server hosting the result. By using the Python `socket` package to obtain an IP address, we are able to use the extreme-ip-lookup API to obtain a latitude and longitude for mapping. This method returns a coordinate for every result and is fairly time efficient. However, there are a number of reasons that this approach might not be ideal for research applications in the humanities—most significantly, the location of the server hosting a modern web page has very little connection to the location of the people who maintain or use the web page. Modern cloud computing and security services like AWS and Cloudflare contribute to and complicate this issue.

Therefore, we developed an alternative method of geolocation that attempts to extract a physical mailing address from a web page. This approach is limited to sites for organizations that publish a physical address; sites that return no result are displayed to the user in a pop-up box on the map. However, it proved successful for our application, because the churches and other institutions associated with Our Lady of LaVang publish their addresses and also serve as a great indicator for Vietnamese populations.

After scraping the text from a web page, regular expressions (patterned after typical addressing formats) are used to capture the address, and the postal code is then translated into a latitude and longitude using the GeoNames data. If the web page links to a "contact" page, that page is searched first. Since addressing formats vary by country, the regular expressions are country-specific; to try to identify the country of origin of the page, various pieces of information are used, such as page language, top-level domain, and search/result country parameters. In order to cut down on false captures of addresses, the region (state, province, etc.) of the address is checked against a list of valid possibilities. This method of geolocation is currently supported for addresses in the United States, Canada, Australia, and Malaysia, and there is moderate support for web pages in languages other than English.

Our Lady of LaVang Locations

As a first application, we mapped locations of Our Lady of LaVang using 30 Google search results (see Figure 3). The markers (red: IP addresses, blue: extracted postal addresses) closely correspond to the largest Vietnamese overseas communities in the US, including Orange County, CA, the San Francisco Bay Area, Texas, Northern Virginia and Florida. Although the map also contains spurious sites (web pages

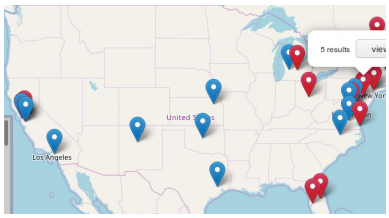
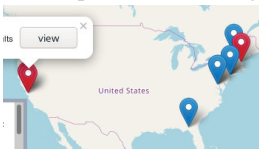


Figure 3: Locations of “Our Lady of LaVang”.

(a) Search phrase: Joe Biden

(b) Search phrase: Pete Buttigieg



(c) Search phrase: Bernie Sanders

(d) Search phrase: Elizabeth Warren

(e) Search phrase: Andrew Yang

Figure 4: Locations of top 30 Google search results for five major 2020 Democratic presidential candidates in alphabetical order obtained in mid-January 2020.

sometimes list addresses that do not correspond to a relevant location), it provides qualitatively useful visual information that suffices for our purpose.

Google Search Results on Democratic Hopefuls

Our tool could potentially be useful for other applications. As another experiment, we mapped the locations of the top 30 Google search results for the five major 2020 Democratic presidential candidates (see Figure 4). Maps for Sanders and Warren contain a marker in France, not shown here. It is injudicious to extrapolate from a few data points; however, the maps do not seem to contradict conventional wisdom of Biden’s strength in the Beltway, Buttigieg’s in the San Francisco Bay Area, Sanders’ in Florida, and Warren’s in New England. The restriction to 30 top Google search results was imposed to obtain each map in a few minutes; the current implementation of physical address extraction is a bottleneck due to the time-intensive web scraping and processing involved. We believe that a faster implementation of our tool using more search results would again produce qualitatively useful visual information for applications such as this.

Conclusions

We proposed studying the Vietnamese global diasporic network using real-time data available from search engines and online social networking sites. To do so, we produced a web tool (publicly accessible at dh.scu.edu:5000) to create an interactive map of the locations of Google search results for the phrase “Our Lady of LaVang,” a Marian apparition that has become the de facto symbol of Catholic Vietnamese overseas communities. Our tool’s generated map closely corresponds to the largest Vietnamese communities in the US. Because our tool allows arbitrary search phrases from the user, it could potentially be used for other applications; as an example we presented the results generated by our tool for major 2020 Democratic presidential candidates, which appear consistent with their acknowledged strengths. We plan to work on improving the speed of our tool and on mapping image queries as well.

References

Agafonkin, V., and other contributors. 2019. Leaflet 1.6.0. https://leafletjs.com/.

Center for Applied Internet Data Analysis. 2020. Internet protocol address (IP) geolocation bibliography. https://www.caida.org/projects/cybersecurity/geolocation/bib.

Duxbury, N.; Garrett-Petts, W.; and MacLennan, D., eds. 2015. *Cultural Mapping as Cultural Inquiry*. Routledge.

Hseb, A. 2017. google-search 1.0.2. https://pypi.org/project/google-search/.

Ninh, T.-H. 2017. *Race, Gender, and Religion in the Vietnamese Diaspora: The New Chosen People*. Palgrave Macmillan.

Nöllenburg, M. 2007. Geographic visualization. In Kerren, A.; Ebert, A.; and Meyer, J., eds., *Human-Centered Visualization Environments*, volume 4417 of *LNCS*. Springer Berlin Heidelberg. chapter 6, 257–294.

OpenStreetMap Contributors. 2019. Openstreetmap. https://www.openstreetmap.org/.

Pallets Projects. 2019. Flask 1.1.1. http://flask.palletsprojects.com/en/1.1.x/.

Richardson, L. 2019. Beautifulsoup4 4.8.2. https://pypi.org/project/beautifulsoup4/.

The Markkula Center for Applied Ethics. 2018. A framework for ethical decision making. https://www.scu.edu/media/ethics-center/ethical-decision-making/A-Framework-for-Ethical-Decision-Making.pdf.