e-Research: A Journal of Undergraduate Work

Volume 1
Number 2 *Vol 1, No 2 (2010)*

Article 4

September 2014

Computational Methods for Historical Research on Wikipedia's Archives

Jonathan Cohen

e-Research: A Journal of Undergraduate Work, Vol 1, No 2 (2010)

Computational Methods for Historical Research on Wikipedia's Archives

Jonathan Cohen

Abstract

This paper presents a novel study of geographic information implicit in the English Wikipedia archive. This project demonstrates a method to extract data from the archive with data mining, map the global distribution of Wikipedia editors through geocoding in GIS, and proceed with a spatial analysis of Wikipedia use in metropolitan cities.

Keywords: Wikipedia Archive, Data Mining, Geocoding, Spatial Data Analysis

Introduction

Wikipedia is one of the most powerful sources of information on the Internet. The site is ranked number 5 among all Internet websites, far ahead of the New York Times, which is the highest-ranked newspaper at number 97 and Encyclopedia Britannica, which is the highest-ranked encyclopedia at number 2,566.[i] The site received over 300 million unique visitors in December 2009 alone.[ii] Wikipedia's content is widely read and cited, and for many, their primary source of online information.[iii]

Wikipedia's popularity and prominence has made it an emerging issue in technology and education. Founded in 2001 as a free online encyclopedia, the defining feature of Wikipedia is that each user can change any page. Each article displays a link which reads "edit this page" to allow any user to change the article's content. This means that when a user makes a change it will be visible to the next user who visits the page. Anecdotal reports suggest that Wikipedia has become immensely popular among students and the bane of teachers in higher education.[iv]

Previous Scholarship

Wikipedia's accuracy and reliability as a source of information has been called into question by the media and academic researchers.[v] Much of this criticism has focused on issues of accuracy connected to the fact that

Wikipedia has no formal peer-review process. However, this criticism mostly faded once the journal *Nature* published a study that found the accuracy of Wikipedia comparable to that of Encyclopedia Britannica.[vi] The criticism of systemic bias on Wikipedia has proved more dogged. Wikipedia editors tend to contribute information that is important and correct to them, but not important and not correct to users in other countries.[vii] This self-focus bias leads Wikipedia toward the problematic representation of Western knowledge defined as all human knowledge. The purpose of this study is to contribute to existing lines of research on systemic bias inherent in Wikipedia.

Prior research has aimed to develop a greater understanding of the population and demographics of the Wikipedia community of editors. To date the most prominent and comprehensive study of the Wikipedia community of editors was presented in Dr. Felipe Ortega's doctoral thesis "Wikipedia: A quantitative analysis."[viii] Ortega traced the evolution of the Wikipedia community in a comparative study of the top ten language versions of Wikipedia. Ortega demonstrated dynamic population trends among this community. The most heavily publicized findings in this study suggested that the population of Wikipedia editors declined by 4,900 in the first three months of 2008 and by 49,000 in the first three months of 2009.[ix] These population trends deserve further study.

One possible means of explaining demographic trends is to break down the data in terms of location to explore precisely *where* Wikipedia editors are leaving the project. Previous scholarship has explored a number of methods to map the location of Wikipedia editors. Some projects made use of a method known as spatial data mining to show that it was possible to link individual Wikipedia editors to a general geographic region. Lieberman demonstrated that users' edits often contained an implicit geographic focus that provided an indication of that users' general location.[x] For example, edits to the Wikipedia articles for the New York Stock Exchange, Central Park, 5th Avenue, and the United Nations would provide a strong indication that the user was born in New York city or had lived in the area. This type of spatial data mining allows researchers to identify Wikipedia editors with a general geographic region to gain insight into user population trends.

Other scholarship refined this spatial data mining methodology to more accurately identify the geographic locations of Wikipedia users. These projects demonstrated a method to link individual Wikipedia editors to a specific geographic location. Hardy demonstrated that the location of anonymous Wikipedia editors could be identified through the IP addresses associated with each of their edits.[xi] Hardy used IP-based geolocation to estimate editor location based on the street address of their Internet service providers.[xii] IP-based geocoding allows for a significantly more focused analysis of population data.

IP-based geolocation can illuminate earlier studies on the social networks of the Wikipedia community of editors. These studies lacked the necessary geographic information to locate social networks of Wikipedia editors and instead represented these networks with abstract diagrams. Zlatic et. al., for example, studied network characteristics among several language versions of Wikipedia and illustrated the results with abstract web graphs.[xiii] This study clearly demonstrated that the Wikipedia community of editors can be understood and

analyzed as a social network. With the incorporation of geographic information to this analysis, social networks can be derived from users who contribute to articles from similar geographic locations. This allows researchers can analyze and illustrate the evolution of the Wikipedia community of editors through their activity in geographically-situated social networks.

Methods/Data

This project was based upon a sample from the edit histories in the digital archive of Wikipedia. [xiv] This digital archive represents a complete history of change on the site. It logs the nature of each edit along with a timestamp, a user name, an IP address, and a comment from the user. A computer program was written in the Python programming language to download a sample of 23,741 anonymous edits for processing and analysis.[xv]

Each of the 23,741 anonymous edits in the sample were linked to implicit geographic information.[xvi] The IP address of an anonymous Wikipedia edit can be traced to the street address of an Internet Service Provider. The IP addresses in the sample were converted into a database of latitude and longitude coordinates and represented on a map using Geographic Information Systems software. The resulting map clearly represented the spatial distribution of Wikipedia edits.

The most significant uncertainty inherent in this methodology is connected to the fact that the *wiki_tracker.py* program only collects data about anonymous Wikipedia users. Wiki software masks the IP addresses of registered users, so geographic information is only easily available for anonymous users. Research to date suggests that the editing patterns of anonymous users are comparable to those of registered users.[xvii] However, the Wikimedia Foundation will need to provide a list of IP addresses of registered users for a more comprehensive study.

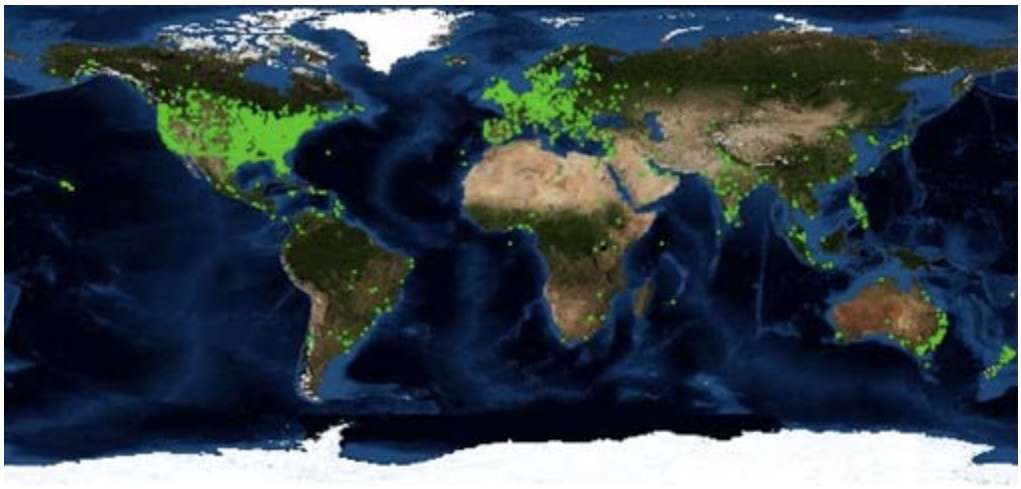

Fig.1 Map of edits to Wikipedia (2001-2009)

Analysis/Discussion

The spatial distribution of Wikipedia editors is clearly represented on the map after geocoding the IP address information. Edit activity is concentrated in North America, Western Europe, and Australia. A small number of countries can account for a majority of the edit activity. About 80% of the edits in the sample originated from the United States (59.5%), United Kingdom (11.8%), Canada (6.6%) and Australia (5.4%). Also, edit activity originating from South America, Africa, and Asia is scarce even when the measures were normalized to account for the relatively lower prevalence of Internet-access in those areas. This suggests that spatial factors do play a role in the population mechanics among the Wikipedia community of users.

Edit activity is concentrated around major metropolitan centers. Major population centers are defined as urban spaces with 100,000 or more Internet users. Examples are New York, Los Angeles, London, Montreal and Melbourne. Spatial analysis of Wikipedia edit activity indicates that 91% of edit activity (21,604 of 23,741 edits) originates within 25km of highly populated urban areas.

There is also a temporal component to the changes in edit activity. Fig. 1 suggests that the population of Wikipedia users grew exponentially from 2002 to 2006. It also suggests that the population of *new* Wikipedia editors started to decline after 2006.[xviii] These data show similar population patterns across the top four countries where Wikipedia is used, suggesting that non-geographic factors are involved in the dynamics of these Wikipedia social networks.

	United States	United Kingdom	Canada	Australia
2002	94	17	16	4
2003	220	41	29	21
2004	619	109	77	49
2005	2924	558	375	236
2006	5122	1197	559	479
2007	2622	449	243	253
2008	1536	225	148	180
2009	964	140	118	71

Fig.2 Yearly edits by anonymous users for top four countries (2002-2009)

Geographic information helps provide a better understanding of how the demographics and population of the Wikipedia community of authors changes over time. Existing lines of research can benefit from this method to map the geographic location of Wikipedia users and to analyze spatial relationships between them.

Conclusion

This paper has demonstrated a method for data mining, geocoding, and spatial analysis of data from the English Wikipedia archive. While earlier Wikipedia research into the population and demographics of Wikipedia editors

has focused on the activity of individual authors, data mining and geocoding can generate new information about the Wikipedia community of authors. This paper uses those methods for spatial analysis to discover how geographic location factors into the population and demographic changes of Wikipedia editors embedded in metropolitan social networks.

The data suggests that geographic is a significant factor in the population and demographic patterns of Wikipedia editors. Most Wikipedia editors originate from a small number of English-speaking countries with high degrees of Internet access. Furthermore, the vast majority of Wikipedia editors live in close proximity to a major metropolitan center. Future research will explore the unique contributions that each metropolitan social network makes to Wikipedia.

References

[i] Alexa Internet Traffic Rankings. January 2, 2010.

[ii] Wikipedia: Statistics. http://www.en.wikipedia.org/wiki/Wikipedia:Statistics.

[iii] Roy Rosensweig, "Can History Be Open Source? *Wikipedia* and the Future of the Past," *Journal of American History* (93), 1-3.

[iv] This issue is discussed in most news reports on the topic of Wikipedia. For a visual representation of Wikipedia's popularity among students see Google Trends: http://www.google.com/trends?q=wikipedia. From 2005 to present, Wikipedia use peaks during the times of year when schools and universities are in session and plummets during periods of summer and winter vacation.

[v] Andrew Keen. *The Cult of the Amateur* (Doubleday/Currency, 2007).

[vi] Jim Giles. "Internet encyclopedias go head to head." *Nature* 438: 900-901.

[vii] Darren Hardy. "Discovering behavioral patterns in collective authorship of place-based information." Paper presented at *Internet Research 9.0: Rethinking Community, Rethinking Place*, 2008, 9.

[viii] Felipe Ortega. "Wikipedia: A Quantitative Analysis." (Ph.D diss., Universidad Rey Juan Carlos, 2009).

[ix] See Julia Angwin and Geoffrey A. Fowler. "Volunteers Log Off as Wikipedia Ages," *Wall Street Journal*, November 27, 2009. Also see: "Wikipedia denies mass exodus of editors," *BBC News*, November 27, 2009.

[x] Michael D. Lieberman and J. Lin. "You are where you edit: Locating Wikipedia users through edit histories," in *Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM'09)*, 106-113, San Jose, CA, May 2009.

[xi] Darren Hardy. "Discovering behavioral patterns in collective authorship of place-based information," in *9th International Conference of the Association of Internet Researchers (IR9: Rethinking community, Rethinking Place)*, Copenhagen, Denmark, October 15-18, 2008.

[xii] Geographic information obtained via this method has weaknesses. Users who access the Internet via a proxy connection would be mis-placed on the map. Also, IP address information is collected only for anonymous users, not registered users.

[xiii] V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet. "Wikipedias: Collaborative web-based encyclopedias as complex networks." *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics* 74 (2006). Specifically, this study traced degree distributions, growth, topology, reciprocity, clustering, assortativity, path lengths, and triad significance profiles.

[xiv] Wikipedia XML database dumps are regularly made available at http://download.wikipedia.org/enwiki/.

[xv] See Appendix A for complete program: *wiki_tracker.py*.

[xvi] Sample derived from the top-level pages and taken to be representative of activity on the site as a whole. Top-level pages used were "Main Page," "Arts," "Biography," "Geography," "History," "Mathematics," "Science," "Society," and "Technology."

[xvii] Darren Hardy. "Discovering behavioral patterns in collective authorship of place-based information," in *9th International Conference of the Association of Internet Researchers (IR9: Rethinking community, Rethinking Place)*, Copenhagen, Denmark, October 15-18, 2008. 7, 10.

[xviii] Data in Fig. 2 is best interpreted as an indicator of the amount of new users arriving to the site because this data is not comprehensive. Fig. 1 only includes edits from anonymous Wikipedia editors. Users tend to register and operate under a single user name after some period of anonymous activity.