You Are Where You Edit: Locating Wikipedia Users Through Edit Histories*

Michael D. Lieberman
Center for Automation Research
Institute for Advanced Computer Studies
Department of Computer Science
University of Maryland
College Park, MD 20742 USA
codepoet@cs.umd.edu

Abstract

Whether knowingly or otherwise, Wikipedia users reveal their interest and expertise through their contribution patterns. An analysis of Wikipedia edit histories shows that it is often possible to associate users with relatively small geographic regions, usually corresponding to where those users were born or where they presently live. Also, for many users, the geographic coordinates of pages to which they contribute are tightly clustered. Results indicate that a wealth of information about Wikipedia users can be gleaned from their edit histories. They illustrate the efficacy of data mining on large, publicly-available datasets and raise potential privacy concerns.

Introduction

Collaboration, end-user involvement, and openness with data are among today's most prevalent Web trends. Web 2.0-style websites such as Facebook, del.ici.ous, and a plethora of extant Wiki projects including Wikipedia all rely on significant contributions from their users that are then shared with the world to achieve a collective user experience unattainable with traditional development methods. In particular, Wikipedia[1] is a collaborative online encyclopedia that grows from article contributions and edits made by its readers. As the quality of Wikipedia articles rivals those of other traditional encyclopedias (Giles 2005), it is perhaps unsurprising that users of Wikipedia tend to contribute information about which they have interest or expertise. Wikipedia has special pages that recognize users with particularly high-quality of large numbers of page edits, providing an important reward for contributing content to the project (Forte and Bruckman 2005). All user page edits are logged and publicly viewable in *edit histories*, which provide a treasure trove of information about the interests and expertise of the users themselves.

Wikipedia users have the option to create personalized user pages that detail information about themselves, such as where they were born, where they live, and their interests. However, even without such pages, users characterize themselves by the number and type of contributions they make. Whether knowingly or otherwise, users reveal their interest and expertise through their edit histories. For example, we might infer that a user with many edits to pages about mountains and mountaineering has a significant interest in that sport. Likewise, a user that contributes significant text to pages about tightly clustered geographic locations, such as College Park, Laurel, and Beltsville (all locales in Prince George's County, Maryland, USA) could then be "located" in that general area. This work demonstrates that by analyzing Wikipedia edit histories, it is often possible to associate users with relatively small geographic regions, usually the areas where those users were born or presently live.

Geography is of special interest because it pervades most topics on Wikipedia, as evidenced by Figure 1, a rendering of the English Wikipedia's geographic coverage, where each point corresponds to a Wikipedia page with geographic coordinates. Even though pages on Wikipedia might not nominally concern specific geographic locations, often pages contain implicit geography that can be used to characterize editors. For example, a number of edits to pages about radio stations in the vicinity of College Park, such as WMUC, WAMU, and WTOP would serve equally well to place the editor near College Park. Pages concerning schools, universities, airports, landmarks, and other notable areas can also serve as markers to associate editors with their implicit geographic locations. We will refer to pages marked with geographic coordinates as *geopages*. Furthermore, we term the minimum region encompassing a user's edits to geopages as the user's *edit area*, which can be computed by taking the convex hull of the geopage coordinates. A small edit area might indicate a general familiarity with the geographic area in question, due to the user being born there, living there, or having an interest in the region.

In this work, we collect a variety of statistics about Wikipedia and its users as they relate to geography. In particular, we examine the geographic coverage of Wikipedia, both in terms of which geographic areas receive the most attention from users and the prevalence of editors of geopages. We also investigate edit patterns and the sizes of edit areas for users who contribute to geopages. Our analysis shows that a significant percentage of users have relatively small edit areas. We identify reasons for this by manually examining users' personal pages, and also find that many users tend to focus their attention on a particular "pet" geopage.

*In partial fulfillment of the requirements for the Master's Degree in Computer Science.

[1] http://wikipedia.org/

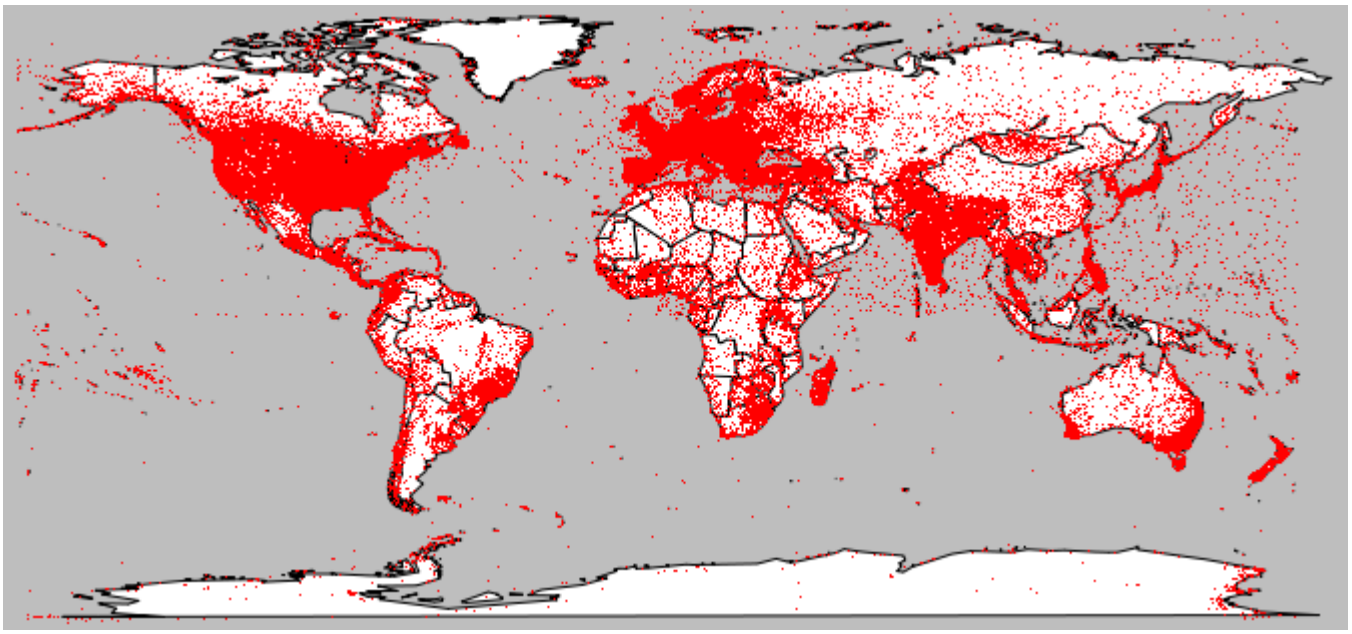


Figure 1: Geographic coverage of Wikipedia pages. Each point represents a latitude/longitude pair found on a Wikipedia geopage. The coverage is uneven, with most geopages placed in the United States and various countries of Europe.

Results show that edit histories provide a wealth of evidence for associating Wikipedia users with geographic regions.

Data mining from user edit histories has a variety of applications, such as psychographic and geographic market segmentation (Lesser and Hughes 1986). It also raises privacy concerns, as users might not intend or want to reveal this information about themselves. Further concern is warranted when this information is joined with data gleaned from other online sources to assemble accurate, multi-faceted profiles of users. This work illustrates the efficacy of data mining on large publicly-available datasets, and highlights the extent to which private information may be inferred from seemingly-innocuous digital footprints.

Related Work

Recently, Wikipedia has been at the center of research in a variety of fields. Researchers have examined general trends in Wikipedia's growth, in terms of number of users and contributions. Voss (2005) examined edit histories of pages and measured several statistics, such as the growth in number of pages and users over time, number of edits per article and user, and page link structure, comparing results between the various language versions of Wikipedia. Roth, Taraborelli, and Gilbert (2008) looked for correlations between administrative policies and growth rates, for a variety of Wiki projects including Wikipedia. Almeida, Mozafari, and Cho (2007) characterize Wikipedia's evolution over time in terms of contributor behavior. They also identify a user behavior where in a single Wikipedia *session*, or time window of 3 hours, users tend to focus their attention on editing a single Wikipedia page. We show that this "pet" page phenomenon holds true for geopages as well, regardless of time window.

Several studies highlight the social qualities of Wikipedia and its users. In interviews, Forte and Bruckman (2005) found that most Wikipedia contributors are motivated by recognition and acknowledgment by their peers. Wikipedia edit histories have also provided a data source for examining how individuals collaborate and resolve conflict in a distributed fashion (Kittur et al. 2007b). Visualizations have have been helpful in this respect (Viégas, Wattenberg, and Dave 2004; Suh et al. 2007). Kittur et al. (2007a) classify users as "elite" or "common" based on administrator status or number of edits, and found that most contributions were initially made by elite users, but gradually shifted to common users over the history of Wikipedia. Ortega and Gonzalez-Barahona (2007) extend the work of Kittur et al. to additional languages, and further examine groups of edits by month, rather than over Wikipedia's entire history. Burke and Kraut (2008) identify characteristics of Wikipedia administrators, then try to predict based on these characteristics whether a given user has a chance of being promoted to administrator status.

Aside from more general measures of Wikipedia's content, a number of researchers focus specifically on geopages in Wikipedia as a source of *volunteered geographic information* (Goodchild 2007), and automated methods of using Wikipedia's geographic content to various ends. Toral and Munoz (2006) examine the utility of Wikipedia pages in creating gazetteers for *named-entity recognition* (Borthwick 1999). Other researchers (Buscaldi, Rosso, and García 2006; Popescu, Grefenstette, and Moëllic 2008) integrated Wikipedia's geographic content with a number of other sources to create *gazetteers*, databases of geographic locations and associated metadata. In a similar vein, Lim et

al. (2006) integrate content mined from Wikipedia geopages into an online digital library. To aid tourists and educators, Hecht et al. (2007) designed a visualization for mobile devices that dynamically places Wikipedia content on a map, using location information sent from the device.

The work most related to ours is that of Hardy (2008), wherein he collects statistics related to Wikipedia geopages. He classifies users as registered, anonymous, or robot, and describes the relative amount of work done by each group, as well as a normalized "locality" measure indicating the relative sizes of edit areas between groups (though not absolute edit area sizes). He found that anonymous authors' edit areas tended to be smaller than those of the other groups. He also presents statistics across the different language versions of Wikipedia. However, Hardy does not attempt to distinguish between geography with and without extent (e.g., countries versus cities), though he does mention the distinction. Also, while he computes locality of edit areas, he does not explain the meaning of his locality measure in sufficient detail, nor does he delve into reasons for users having small edit areas.

Data Sources

The main data source used in our analysis is the English Wikipedia XML dump[2]. The dumps are updated every several months, and several forms of it are available depending on how it is to be used. In addition to complete page content being available, all previous versions of pages are also stored and available, along with complete page edit histories. For each edit made to a page, the user that made the edit and the edit's timestamp are recorded. In Wikipedia, users have the option of either logging in with a username and password to make edits, or editing anonymously. For users that have logged in to edit, their username is stored in the edit history. Anonymous users have their IP address recorded in the edit history. We used the English Wikipedia page history dump of 8 Oct 2008. The dump file totals 8.4GB compressed, and 61.7GB uncompressed.

When saving an edit, named (i.e., non-anonymous) users have the option of marking their edit as "minor". The minor flag is intended to distinguish between true contributions to a page's content and simple changes, such as spelling or grammar correction, or formatting changes. In addition, a number of robots used to make mass changes to a large collection of pages also use the minor flag. From these observations, we made the simplifying assumption in our analysis that a minor edit to a geopage did not serve as evidence that the user making the edit was in any way related to that geographic location. We therefore excluded minor edits from the analysis, as they would tend to skew correlations between users with legitimate contributions, and page geography. Note that it is entirely subjective whether or not to mark an edit as minor. However, in practice, we found that most geopages tended to have many edits that were marked as minor, with only a few users making significant contributions to a given page. The minor flag thus served as a useful indicator of true knowledge or experience with a geopage and its corresponding geographic location.

We excluded anonymous edits from our analysis for several reasons. While IP addresses serve as a valuable source of location information (Padmanabhan and Subramanian 2001), several problems deter a meaningful analysis of anonymous user edits. For example, when editing Wikipedia, anonymous editors do not have the option to mark page edits as minor, due to the potential for abuse. Therefore, it would be difficult to distinguish significant page edits from typos and spelling corrections. Also, several informal studies[3] show that anonymous editors are responsible for the majority of Wikipedia article vandalism, which should not be considered legitimate evidence of geographic locality. Another problem when using IP addresses is the inherent inability to correlate a single IP address to a single human editor, since they might be assigned to Internet users dynamically. Furthermore, a single IP address might be used by several humans simultaneously, as in the case of proxy servers for local area networks.

Identifying Geography

Finding Wikipedia pages tagged with geographic coordinates, while seemingly simple, tends to be difficult for several reasons. In general, geopages have the relevant geographic coordinates present somewhere in the page's content. However, page content is written in a constantly evolving Wiki markup language, which makes it difficult to parse. The problem is exacerbated by the large number of ways that users express geographic coordinates within page content. Editors often create parameterized templates that can be reused on many pages, to avoid duplicate work by editors and allow for uniformity across pages. However, the templates themselves constantly evolve, and templates follow trends of use and disuse. At this time there exist at least 20 distinct forms of template parameters, all of which serve the same basic purpose of annotating a Wikipedia page with geographic coordinates. For example, separate template parameter sets exist depending on the type of annotated object, such as country, administrative division, city, or spot feature. Various forms of geographic coordinates (e.g., degrees-minutes-seconds (DMS) and decimal degrees (Clarke 1995)) can be used as well, and it is left to the editor to decide which form is most appropriate.

To avoid the messy task of extracting geographic coordinates from raw Wiki markup, we integrated data from DBpedia (Auer et al. 2007), a community project that aims to extract semantic relationships mined from Wikipedia. Along with many other types of semantic information, DBpedia features a table of geographic coordinates mined from Wikipedia's many geographic coordinate templates. This table amounts to a primitive *gazetteer* (Hill 2000), or database of geographic locations and associated metadata. The DBpedia gazetteer thus provides links between Wikipedia pages and geographic coordinates. Its uncompressed size is 166MB.

Another complication that an analysis of Wikipedia geography entails is accounting for the geography of features with significant extent (Clarke 1995), such as regions (e.g.,

[2] http://download.wikimedia.org/enwiki/

[3] http://wikipedia.org/wiki/WP:WPVS

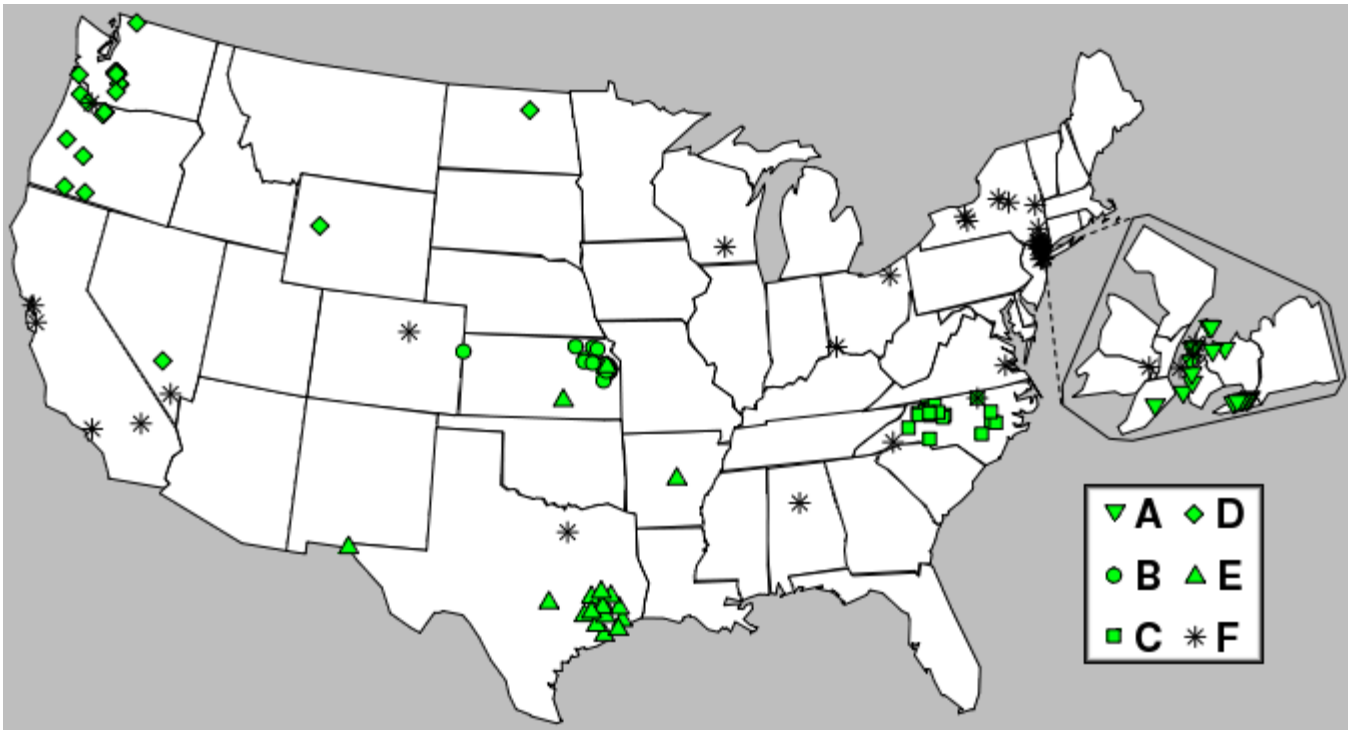


Figure 2: A variety of edit patterns in the USA that lead to distinct edit areas. Each letter refers to a different user, and each point corresponds to an edit to a geopage tagged to those coordinates. Notice that in many cases, a user's edits to geopages are tightly clustered (e.g., A), but might have one or several edits that are geographically distant (e.g., B, E).

countries, administrative divisions, lakes) and linear features (e.g., roads, rivers, canals). In Wikipedia, all geographic features, including those with extent, are annotated with a single lat/lon point. The point chosen is particular to the type of feature. For example, political regions like countries and administrative divisions (e.g., states, counties, boroughs) are tagged with the geographic coordinates of their capital or home office, while linear features are generally tagged with their midpoint or an end point (e.g., for rivers, the mouth or source of the river). Ideally, features with extent would be tagged in a distinct manner from point features, but for the moment, geographic tagging projects on Wikipedia favor uniformity over representational accuracy. Incorporating the tagged coordinates of features with extent is problematic because geographic coordinates can only capture distance relationships between points, but not other spatial relationships such as overlap and containment, which might reveal additional connections between page edits. For example, a user with several edits to College Park, Laurel, and Beltsville, as well as Maryland, would indicate a strong association with the three initial localities, since they are all in Maryland and are geographically proximate. However, examining the coordinates of the corresponding Wikipedia pages might indicate otherwise, because Maryland would be tagged with the coordinates of its capital city, Annapolis, which could be geographically distant from the remaining localities.

However, on Wikipedia, the precision of tagged geographic coordinates serves as a hint for the feature's size, which in turn reveals whether the location in question has significant extent. For example, the Maryland Wikipedia page is tagged with decimal coordinates (39, -76.7), which is in fact the coordinates of its capital, Annapolis (38.972945, -76.501157) but expressed with less precision. In contrast, the College Park, Maryland article is tagged with coordinates (38.99656, -76.927509) which indicates a much higher degree of precision, and hence smaller extent. We therefore marked those pages with fewer than 2 digits in the fractional part of the decimal coordinates as being features with extent.

Like all of Wikipedia's content, the precision of tagged geographic coordinates is subject to human error. For some geopages, we found that tagged coordinates were entered with too little or too much precision, especially for those pages that received little attention from editors. However, we found that geopages corresponding to features with extent were in general correctly tagged, as they tended to have multiple revisions by different users. In our analysis, we tested the effects of both including and excluding features with extent on user edit area sizes.

Typical Edit Patterns

To clarify the preceding discussion, we now present several examples of real user edit areas from the English Wikipedia. Figure 2 shows six users whose edit areas lie mostly in the United States. Each letter corresponds to a Wikipedia user. The outlined region at the extreme right containing user A's edits is an enlargement of New York City and surrounding

counties. These editors were selected because they have a sizable number of geopage edits, and they exhibit a wide range of edit area sizes. In addition, while one user posted biographical information on a Wikipedia user page, the rest did not, and thus might be surprised that information about their geographic origins and interests could be gleaned from their edit histories. We also found these edit patterns to be representative of a large portion of Wikipedia editors that modified geopages.

In the figure, users with small edit areas include user A, editing many geopages in New York City, New York, and B, with many edits in Douglas County, Kansas. These editors have minimal edit areas of under $1\,\mathrm{deg}^2$. The collection of geopages edited by A include the page about New York City, as well as a number of smaller pages about subway stops in New York City's various boroughs. As a result, the edits are tightly clustered, with no geographic outliers. It would be safe to say that user A is familiar with New York City, and likely lives there. It might even be possible to pin user A to specific neighborhoods by examining which subway stops were edited most by content or number of edits. Similarly, user B's edits include many small townships in Douglas County, in the northeastern part of Kansas, and other nearby cities. B's edits include one outlier on the border of Kansas and Colorado. In our analysis, we discounted a small percentage of geographic outliers in determining users' edit areas, to account for cases like these.

Medium-sized edit areas can be attributed to users C, D, and E, with edit area sizes ranging between about $3\,\mathrm{deg}^2$ (C) and $71\,\mathrm{deg}^2$ (E). Most of user C's edits are to geopages about various populated places in North Carolina. However, one of these geopages is actually that of a local television station which was tagged with the geographic coordinates of its transmission antenna. In a similar vein, user D's edit area includes several different types of geographic features in Washington State and Oregon, including villages, glacier sites, rivers, and mountains, as well as a number of outlier edits. These users demonstrate that articles about many types of geographic features can contribute to characterizing a user's edit area. User E's edit area is somewhat larger, mainly focused on large cities and counties in Texas, but also including edits to articles with coordinates in nearby states. Again, we account for these outliers in our analysis.

Finally, the largest edit area belongs to user F, with a total area encompassing over $1000\,\mathrm{deg}^2$, and including edits to geopages situated all across the United States, with a sizable number of edits in New York State. The types of geopages edited by user F are greatly varied, including the usual populated places, but also bridges, hotels, and the sites of several plane crashes. Several edits are to geopages placed outside the United States and are not shown.

Analysis

We first present basic statistics about the Wikipedia and DBpedia dumps used in our analysis in Table 1. The `Total` column indicates the total number of objects in the dump, while the `Geo` and `Geo%` columns give the number and percentage of objects that contain geographic information. The statistics show that a considerable number of pages are geopages and are marked with geographic coordinates. Also, a nontrivial number of users (14.6%) have made at least one non-minor edit to a geopage. In addition, most (58.7%) edits to geopages are non-minor edits.

Statistic		Total	Geo	Geo%
Pages		14915993	328393	2.2%
Users		2440777	356693	14.6%
Edits	Total	168901990	10822130	6.4%
	Non-Minor	114844836	6357558	5.5%

Table 1: Wikipedia/DBpedia dump statistics. A considerable number of pages are tagged with geographic coordinates, and most edits are marked as non-minor edits.

Country	Count	Country	Count
United States	83971	Russia	10964
France	37730	Canada	8970
United Kingdom	26651	Italy	8772
Poland	16050	Spain	6603
Germany	15939	India	5683

Table 2: Top page counts, aggregated by country. As might be expected for the English Wikipedia, the majority of geopages edited lie in the United States and various countries of Europe.

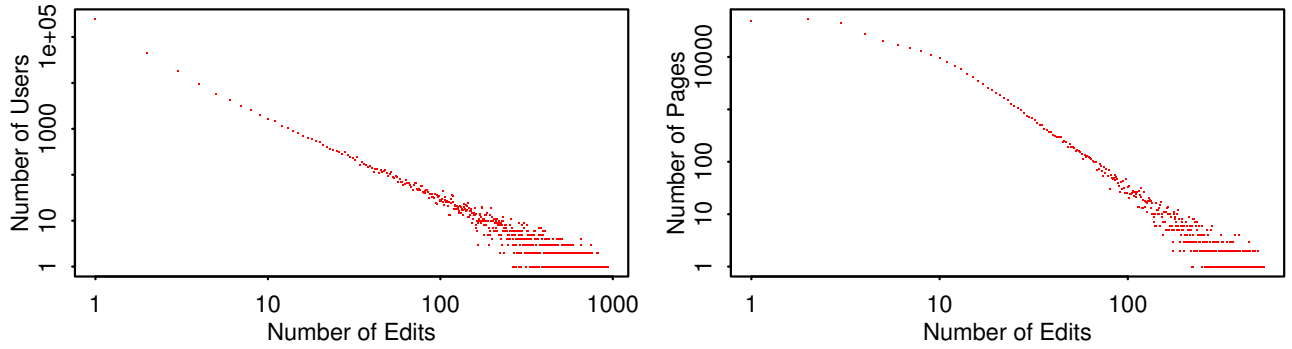
Table 2 contains the top ten page counts, aggregated by country. These page counts were determined by assigning each page's coordinates to the country that contains it, thus determining the countries containing the largest number of pages. As can be seen in the table and Figure 1, the vast majority of Wikipedia's geographic coverage lies in the United States and various countries of Europe. This uneven coverage might reflect the geographic distribution of contributors due to limiting out analysis to the English Wikipedia. If we examined Wikipedia dumps of other languages, a different bias would likely be found.

Figure 3 shows the distributions of edits to geopages across users and pages. Both distributions follow a general power-law curve. That is, a tiny number of users and geopages have very large edit counts, and the number of edits rapidly falls as the number of users and pages increase.

To determine whether users might be surprised by the information revealed through their edits, we checked what percentage of geopage editors also have user pages. If a user has a user page, we assume that the user is willing to share at least some information about themselves, and is more heavily involved in Wikipedia. Of 356693 users with at least one edit to a geopage, only 102271 (28.7%) also have user pages. Also, for the 93195 users with at least five edits to geopages, only 47623 users (51.1%) also have user pages.

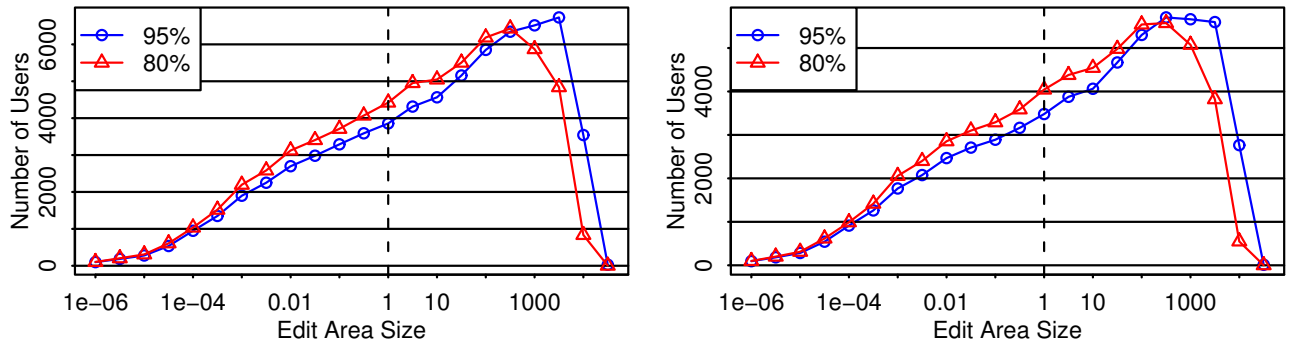
Locality of Edit Areas

We next analyzed the tightness of users' edit areas. For each user, we computed the convex hull of the geographic coordinates of the pages that the user edited, then computed the area of the polygon defined by the convex hull. A smaller



(a) (b)

Figure 3: Distributions of geographic edits across (a) users and (b) pages. The number of edits and editors follow power-law distributions.



(a) (b)

Figure 4: Geographic locality of user edit areas with features with extent (a) included and (b) excluded. A large number of users — approximately 30–35% of all users with edits to geopages — have edit areas smaller than $1\,\mathrm{deg}^2$, indicated by the dashed vertical line. Using a smaller fraction of user edits shifts edit areas significantly across the $1\,\mathrm{deg}^2$ boundary.

edit area thus indicates more geographically clustered edits. However, this simple computation does not adequately account for geographic outliers. A single edit to a page with coordinates located very far from a tight cluster of edit locations would expand the convex hull's area to a very large value, despite most locations being tightly clustered. To account for these outliers and ensure a more meaningful analysis, we removed a fraction (5% and 20%) of problematic edits from each user's set of edits and computed edit area based on the remaining points. Furthermore, we only considered those users with at least three edits. As an example, in Figure 2, removing 20% of user E's 18 edits leaves only the tight cluster of edits in southeast Texas.

Figure 4a shows the number of users with a given edit area, in deg^2, using 95% and 80% of edited geopages for each user. Of 67638 users plotted, 20737 (30.7%) and 23544 (34.8%) of users' edit areas cover less than a $1\,\mathrm{deg}^2$ region with 95% and 80% confidence, respectively. An area of $1\,\mathrm{deg}^2$ approximately corresponds to a 100x100km region, or the size of a typical metropolitan region. Furthermore, of users with under 5 edited geopages, which account for 37820 of the total number of users, 17813 (47.1%) and 19633 (51.9%) have their edit areas constrained to a $1\,\mathrm{deg}^2$ region with 95% and 80% confidence. Using 80% rather than 95% significantly shifts user edit areas toward smaller values, especially across the $1\,\mathrm{deg}^2$ boundary. These figures and statistics indicate that a significant portion of users' edits are restricted to relatively small geographic areas.

We also identified geopages that correspond to regions with extent, and investigated the effects of their removal on users' edit areas. Figure 4b shows our results. Of 60045 users, 18917 (31.5%) and 21531 (35.9%) of users' edit areas are smaller than $1\,\mathrm{deg}^2$ using 95% and 80% of users' total edited geopages. Also, of the 33385 users with less than 5 edited geopages, 16094 (48.2%) and 17809 (53.3%) have edit areas smaller than $1\,\mathrm{deg}^2$. Excluding regions with extent thus results in about a 1% drop in edit area sizes across the $1\,\mathrm{deg}^2$ boundary. The main effects were on users with initially large edit areas when taking 95% of edits to geopages, which are shown in the extreme right of the graphs. This indicates that few users make many edits to geopages corresponding to large geographic features. Instead, user edit areas mostly consist of small features, which better aid in tying users to specific geographic areas.

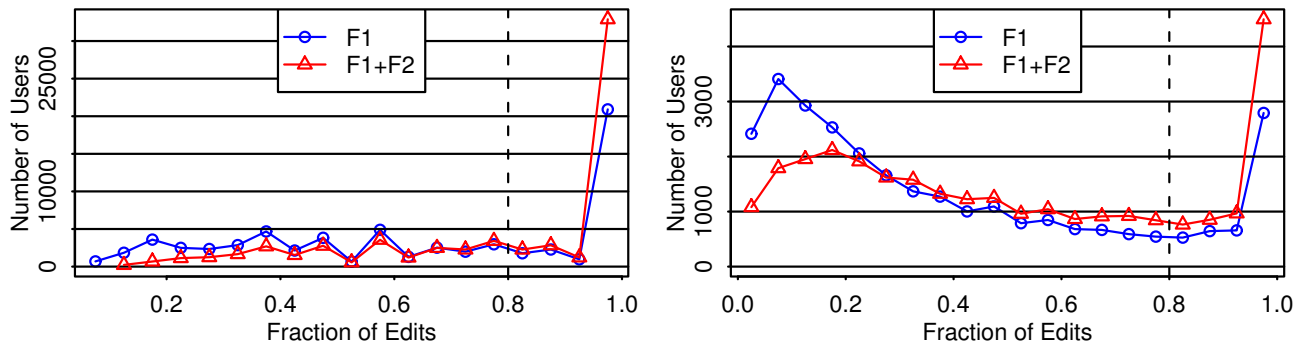


Figure 5: Frequency statistics revealing the prevalence of pet geopages among users with (a) 5–20 and (b) over 20 edits to geopages. Significant numbers of users have a large percentage of their edits confined to one or two geopages.

Pet Geopages

We next looked for users keeping "pet" geopages — those users who concentrate their edits on one or two geopages. For each user u, we checked the number of edits to each geopage edited by u, and determined u's most-edited geopages. Figures 5a and 5b show our results for users with 5–20 and over 20 edits to geopages, respectively. In the figures, F1 and F2 refer to the frequencies of the most- and second-most edited geopage. Of the 93195 users with 5–20 edits to geopages, 32899 (35.3%) have at least 80% of their edits confined to a single geopage, and 48969 (52.5%) have over 80% of their edits for two geopages. Also, for the 28475 users with over 20 edits to geopages, 4689 (16.5%) and 7186 (25.2%) have at least 80% edits constrained to one and two geopages, respectively. Pet geopages thus are a common occurrence, for both casual and regular editors.

Reasons for Small Edit Areas

We then tried to determine the reasons for users having especially small edit areas, based on the information made public by users on their user pages. We randomly selected 100 users with at least 10 edited geopages, having edit area sizes of less than $1\,\mathrm{deg}^2$, and having user pages. Then, for each user u, we concurrently viewed u's user page and set of geopages edited, to determine these reasons. Table 3 lists our findings. As we expected, users with small edit areas tended to either be born in or living in the region defined by their edit area, with over half of users stating so explicitly. The remaining users did not state a geographic interest or expressed general or special interest for some local features of their edit areas, such as local businesses, schools, and railways. Note that only reasons stated explicitly on user pages were included in our counts, but it is reasonable to assume certain relationships with edit areas even if not stated. For example, the users with interests in local schools most likely were born or live in the area as well.

Interest	Count	Interest	Count
Living there	56	General	5
Unknown	24	Local schools	5
Born there	19	Local businesses	3
Local railways	9	Local history	1

Table 3: Reasons for users having especially small edit areas (under $1\,\mathrm{deg}^2$), determined by voluntary information gleaned from user pages.

Future Work

We have shown that a significant group of Wikipedia users exhibits selectivity and geographic locality in the geopages that they edit. However, more Wikipedia information could be used to identify edit areas for a larger portion of Wikipedia users. For example, we used the presence or absence of a minor flag to determine edit importance, but the minor flag is manually set at the time of editing, and in some cases might be missed. Alternative measures can serve as more accurate indicators of the importance of individual edits, such as the page size difference before and after the edit, and whether the page was reverted to an earlier version by another user. Also, rather than ignoring minor edits, they might be used as an additional source of evidence for determining edit areas. A large number of minor edits to geopages in a small geographic area could indicate interest in that area, even if few significant contributions were made to those pages.

Alternatively, more extensive information mining can be done using other freely available data sources to enhance the gazetteer. Instead of using the limited DBpedia gazetteer, another gazetteer, such as the GNIS/GNS[4] or GeoNames[5], could be used to aid analysis. Doing so would allow other gazetteer features, such as population, hierarchy or containment relationships, and feature classes, to aid in identifying features with extent and generally enhancing relationships between geopages in users' edit areas. For example, it might be of interest to examine correlations between geographic location population and the number of edits to the corresponding Wikipedia geopage.

[4] http://geonames.usgs.gov/
[5] http://geonames.org/

Conclusion

This work provides a case study on the efficacy of data mining on large, publicly-available data sets. Active users of projects like Wikipedia should be aware that their contributions can increasingly be used to find information about them that they might not want revealed. Furthermore, for Wikipedia, edit histories permanently associate users with the pages they edit, including non-geopages and future contributions. Additional concern is warranted when multiple datasets are joined to construct accurate multi-faceted user profiles. As the Internet advances toward more interactive and open applications, users should become more savvy in their decisions on making personal information public. However, as we have shown, these are very difficult decisions, considering the wealth of information that can be gleaned from seemingly-innocuous digital footprints.

Acknowledgements

The author wishes to thank Jimmy Lin for his invaluable comments and feedback.

References

Almeida, R. B.; Mozafari, B.; and Cho, J. 2007. On the evolution of Wikipedia. In *Proceedings of the 1st International AAAI Conference on Weblogs and Social Media*.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference*.

Borthwick, A. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. Dissertation, New York University.

Burke, M., and Kraut, R. 2008. Taking up the mop: Identifying future Wikipedia administrators. In *CHI'08 Extended Abstracts*, 3441–3446.

Buscaldi, D.; Rosso, P.; and García, P. P. 2006. Inferring geographical ontologies from multiple resources for geographical information retrieval. In *Proceedings of the SIGIR'06 Workshop on Geographic Information Retrieval*.

Clarke, K. C. 1995. *Analytical and Computer Cartography*. Englewood Cliffs, NJ: Prentice-Hall, second edition.

Forte, A., and Bruckman, A. 2005. Why do people write for Wikipedia? incentives to contribute to open-content publishing. In *Proceedings of the GROUP'05 Workshop on Sustaining Community*.

Giles, J. 2005. Internet encyclopaedias go head to head. *Nature* 438:900–901.

Goodchild, M. F. 2007. Citizens as sensors: The world of volunteered geography. *GeoJournal* 69(4):211–221.

Hardy, D. 2008. Discovering behavioral patterns in collective authorship of place-based information. In *Proceedings of the 9th International Conference of the Association of Internet Researchers (Internet Research 9.0)*.

Hecht, B.; Rohs, M.; Schöning, J.; and Krüger, A. 2007. WikEye — using magic lenses to explore georeferenced Wikipedia content. In *Proceedings of the 3rd International Workshop on Pervasive Mobile Interaction Devices*, 6–10.

Hill, L. L. 2000. Core elements of digital gazetteers: Placenames, categories, and footprints. *Lecture Notes in Computer Science* 1923:280–290.

Kittur, A.; Chi, E.; Pendleton, B. A.; Suh, B.; and Mytkowicz, T. 2007a. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proceedings of Alt.CHI'07*.

Kittur, A.; Suh, B.; Pendleton, B. A.; and Chi, E. H. 2007b. He says, she says: Conflict and coordination in Wikipedia. In *Proceedings of the 2007 SIGCHI Conference on Human Factors in Computing Systems*, 453–462.

Lesser, J. A., and Hughes, M. A. 1986. The generalizability of psychographic market segments across geographic locations. *Journal of Marketing* 50(1):18–27.

Lim, E.-P.; Wang, Z.; Sadeli, D.; Li, Y.; Chang, C.-H.; Chatterjea, K.; Goh, D. H.-L.; Theng, Y.-L.; Zhang, J.; and Sun, A. 2006. Integration of Wikipedia and a geography digital library. *Lecture Notes in Computer Science* 4312:449–458.

Ortega, F., and Gonzalez-Barahona, J. M. 2007. Quantitative analysis of the Wikipedia community of users. In *Proceedings of the 2007 International Wiki Symposium*.

Padmanabhan, V. N., and Subramanian, L. 2001. An investigation of geographic mapping techniques for internet hosts. In *Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, 173–185.

Popescu, A.; Grefenstette, G.; and Moëllic, P.-A. 2008. Gazetiki: Automatic creation of a geographical gazetteer. In *Proceedings of the 8th Conference on Digital Libraries*, 85–93.

Roth, C.; Taraborelli, D.; and Gilbert, N. 2008. Measuring wiki viability: An empirical assessment of the social dynamics of a large sample of wikis. In *Proceedings of the 2008 International Wiki Symposium*.

Suh, B.; Chi, E. H.; Pendleton, B. A.; and Kittur, A. 2007. Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. In *Proceedings of the 2007 Symposium on Visual Analytics Science and Technology*, 163–170.

Toral, A., and Munoz, R. 2006. A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. In *Proceedings of the EACL'06 Workshop on New Text*.

Viégas, F. B.; Wattenberg, M.; and Dave, K. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the 2004 SIGCHI Conference on Human Factors in Computing Systems*, 575–582.

Voss, J. 2005. Measuring Wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*.