What Kobe Bryant and Britney Spears Have in Common: Mining Wikipedia for Characteristics of Notable Individuals

Pauline C. Ng
Independent Researcher
pauline@paulinepi.com

Abstract

This paper proposes a statistical methodology for mining Wikipedia to discover characteristics associated with life outcomes. The methodology is demonstrated using first names and childhood environment. By comparing over 35,000 Wikipedia biographies against spatially and temporally matched census data, we show that individuals with rare names are twice as likely to appear in Wikipedia (RR=2.43 for females; RR=2.30 for males). This result is supported by past studies. Furthermore, birth location also plays a role in success: individuals born in New York and California are ~2x more likely to become entertainers, and those born in the South are ~1.5x more likely to become athletes. These results validate the proposed methodology of using Wikipedia to study life outcomes.

Introduction

What makes a person successful in life? Previous studies have used surveys to determine that family income, neighborhood, and name are some factors that have been shown to be correlated with life outcomes (Davis-Kean 2005, Leventhal and Brooks-Gunn 2000, Zweigenhaft 1977). However, surveys tend to be limited in size (hundreds to low thousands of samples) due to budgetary constraints.

Wikipedia is an open-source online encyclopedia and contains over 500,000 articles on living persons. Clearly, Wikipedia is a substantial resource for information on notable or famous individuals. Because Wikipedia's biographies often contain information on a person's early life, Wikipedia could be mined for characteristics associated with notability and positive outcomes in life. This begs the question: can a person's presence in Wikipedia be analytically correlated with 'success'? If so, then mining the biographies in Wikipedia could be a faster and less expensive alternative to conducting surveys.

Wikipedia's criterion for inclusion of an individual is based on a person's notability: he/she must be the subject of multiple, verifiable sources (e.g. newspaper or magazine articles). The person does not necessarily have to be wealthy, but we make the assumption that notability is correlated with 'success' of some form.[1] Wikipedia's editorial standards attempt to ensure unbiased and independent confirmation of notability by discouraging self-reporting through the requirement of verifiable third-party citations. This is in contrast to self-reported surveys, where people tend to report positively on their own abilities (Cook and Campbell, 1979).

In this study, we present a methodology for discovering trends among Wikipedia individuals. As a proof-of-principle, we analyze two characteristics, first names and birth locations, for correlations among notable individuals. These characteristics are chosen because 1) these are among the first choices that new parents make, and 2) the existence of benchmark data.

Methods

Data Collection

Name and birth location were extracted from the Wikipedia entries for people born in the United States from 1940-1989. On Dec 17, 2010, Wikipedia pages for people belonging to the category "<year> births" where <year> ranged from 1940-1989, and any category that contained the word "American" (e.g. "American architecture writers") were downloaded. The restriction for Americans born between 1940-1989 is necessary in order to properly match to the control population, as explained below.

[1] Wikipedia does contain criminals. These individuals were not eliminated from the study as it is assumed they take up a small number of entries. Furthermore, some individuals may be considered successful yet have served jail time (e.g. Martha Stewart, Bernie Madoff).

This yielded 40,250 biographies for people born in the United States between 1940 and 1989. In the analysis on first names, 35,537 names were found in the U.S. Census data and had corresponding control frequencies; the rest were discarded. For the second analysis on birth states, the 40,250 Wikipedia pages were parsed to retrieve the birth state of a person by either recognizing the Wikipedia infobox, or a sentence within the article containing the word "born", the birth year, and a U.S. state. A total of 35,604 articles with birth states were recognized. Retrieval of first names and birth locations were validated at 93% (93/100 entries) and 97% (97/100) respectively. The 7% error rate for first names is due to not capturing some nicknames. In the 100 entries, we did not find evidence of name changes, suggesting these events are infrequent.

Calculating Relative Risks For Characteristics in Wikipedia

The relative risk for characteristic *k* appearing in Wikipedia is calculated as $RR_k = p_k^{wiki} / p_k^{U.S.}$ where p_k^{wiki} is the frequency of characteristic *k* in Wikipedia, and $p_k^{U.S.}$ is the frequency of characteristic *k* in the U.S. population. When $RR_k = 1$, this means that the frequency of *k* in Wikipedia matches what is observed in the control population and this characteristic is not associated with notability. If $RR_k > 1$ this signifies that the characteristic *k* is observed more frequently in Wikipedia than expected and is associated with increased notability.

p_k^{wiki} is the frequency of the characteristic *k* in Wikipedia summed over all years:

$$p_k^{wiki} = \frac{\sum_{year} Wikipedia\_count_{k,year}}{\sum_{year}\sum_{k} Wikipedia\_count_{k,year}}$$

where $Wikipedia\_count_{k,year}$ is the number of Wikipedia biographies that have characteristic *k* and are born in *year.*

The denominator $p_k^{U.S.}$ is slightly more complicated because the frequency of *k* can change over time. (For example, city and state populations have varying historical growth rates depending on their local economies.) The frequency of characteristic *k* for *year* in the US population is calculated as:

$$p_{k,year}^{U.S.} = \frac{U.S.\_population_{k,year}}{\sum_{k} U.S.\_population_{k,year}}$$

where $U.S.\_population_{k,year}$ is the number of people having characteristic *k* and born in *year*. Then $p_k^{U.S.}$ is defined as the weighted average of $p_{k,year}^{U.S.}$ over all years:

$$p_k^{U.S.} = \sum_{year} (p_{k,year}^{U.S.} * weight_{year})$$

where the weight is proportional to the number of Wikipedia entries for that year.

To obtain $p_{k,year}^{U.S.}$ for first names, the top 1,000 names and their respective U.S. population frequencies for each year between 1940 and 1989 were downloaded from the Social Security Administration (U.S. Social Security Administration, 1997). Names with similar frequencies were binned together to ensure enough counts for statistical analysis. For birth locations, state populations for each decade were obtained from the U.S. Census (U.S. Census, 2002) and states' populations for years in between the decades were interpolated.

Analysis and Validation

Case Study I. First Names

Name frequency in Wikipedia is compared to that of the American population to see if the methodology presented here provides results consistent with previous studies. If names do not play a factor in notability, then the following should be observed: if 5% of American men are named "John", then one would expect 5% of the American men in Wikipedia to be also named John. If the observed percentage in Wikipedia is actually higher, this would indicate a positive association between that name and presence in Wikipedia. This study calculates the relative risk (RR), which is the ratio of the name frequency in Wikipedia to name frequency of the American population. A RR greater than 1 indicates the name appears more frequently in Wikipedia than expected.

Results

When first names are processed with strict matching rules, we observe a decreased likelihood that people with common names (frequency greater than 1%) would appear in Wikipedia (RR:0.52 (95% Confidence Interval: 0.50-0.53) for males; RR:0.71 (95% CI: 0.66-0.75) for females) (Figure 1, striped bars). It is surprising that having a common

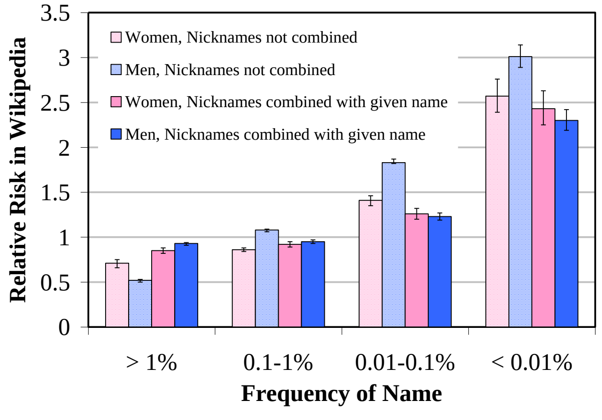

Figure 1: The likelihood of being in Wikipedia based on name frequency.

name would have markedly disadvantageous effects. However, analysis reveals that the proper name Michael was 42% less likely to appear in Wikipedia than expected, but its corresponding nickname “Mike” was 9 times more likely to appear. In general, males with nicknames are 2.39 times more likely to appear in Wikipedia while for females the effect is less but still present with a 1.32-fold increase. Therefore, those that use a nickname instead of their formal names have a greater chance of appearing in Wikipedia (e.g. Bill Gates instead of William Gates). We assume that Wikipedia individuals with nicknames had proper names legally assigned to them when born (e.g. Michael for Mike, William for Bill) and combined the proper name and their corresponding nicknames. Then the likelihood of appearing in Wikipedia for a common name approaches 1 (Figure 1, solid bars).

Our results show that rare names are more likely to appear in Wikipedia than expected. As nicknames have been combined with their respective proper names, a name like Mike is not counted as a rare name. For names between 0.01-1% in frequency, there is a ~25% increased likelihood of appearing in Wikipedia (RR=1.26 (1.20-1.32) for females; 1.23 (1.19-1.27) for males) (Figure 1, solid bars). For names less than 0.01% frequent in the population, the likelihood of being present in Wikipedia is more than double (RR: 2.43 (95% CI: 2.25-2.63) for females; 2.30 (2.19-2.42) for males) (Figure 1, solid bars). Similar results are obtained when individuals are subdivided into athletes and entertainers.

Our method corrects for year of birth, as one must account for how name frequency changes throughout the years. For example, ‘Britney’ was ranked #758 by the Social Security Administration in 1981, which is the year performer Britney Spears was born. At that time, her name is rare and from this analysis, there is an increased relative risk for appearing in Wikipedia. However, in 2000, when Britney Spears achieved Best Selling Album by a Female Artist in the U.S., more girls were being named ‘Britney’ and its rank increased to #137. Based on this study, one may infer that Britney’s born in 2000 do not have the same rare-name advantage as Ms. Spears.

Comparison to Related Work

Past research has suggested that successful individuals tend to have rare names (Zweigenhaft 1977). This phenomenon does not imply causation: parents that give their children rare names may view their children as special and provide extra opportunities for development.

Zweigenhaft chose 218 rare male names and 218 common male names and then looked for their presence in *Who’s Who*, a sponsored compendium of “noteworthy and influential” people. Out of the 436 names, thirty were found, and a high proportion were rare (Zweigenhaft 1977). Our analysis of the Zweigenhaft dataset calculates an odds ratio of 3.6 (CI:1.5-8.5) for appearing in *Who’s Who*. The odds ratio for males appearing in Wikipedia with name frequency < 0.01% is 2.49 (CI:2.36-2.62). This odds ratio falls within the 95% confidence interval reported by Zweigenhaft, thereby validating our method.

Case Study II. Birthplaces

Results

We calculate birthplace trends in Wikipedia by comparing the frequency of birth states in Wikipedia versus the U.S. population distribution. For example, California residents accounted for 9% of the U.S. population for the period under study. In Wikipedia, they account for 15% of the entries which equates to a RR of 1.69 (CI: 1.65-1.74). Therefore, being born in California is associated with a 69% increased likelihood of appearing in Wikipedia. For the purposes of this study, the location of birth in

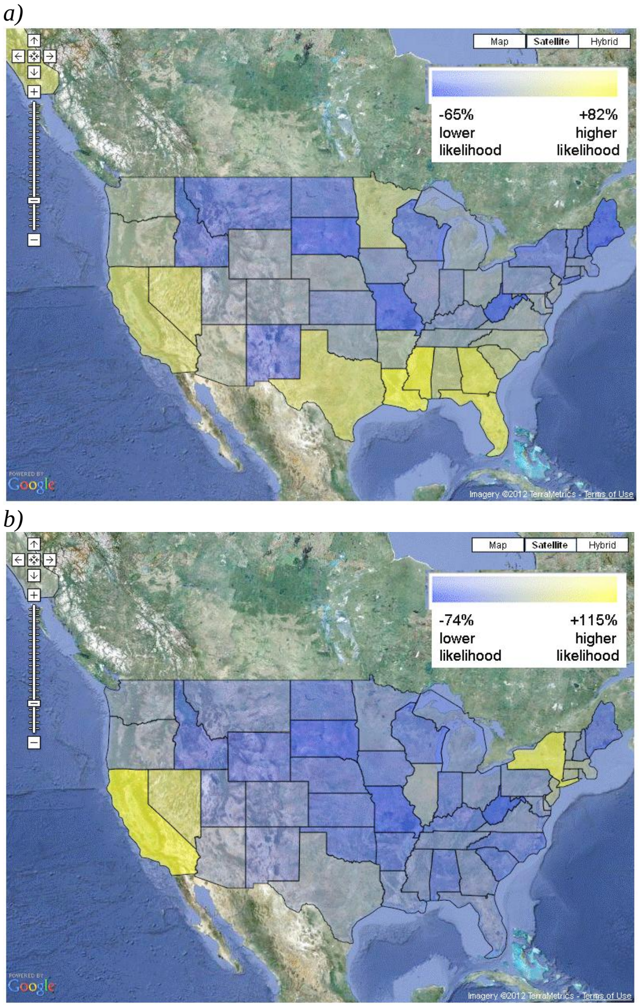

Figure 2: Relative risks for appearing in Wikipedia based on birth states for (a) athletes and (b) entertainers. Visualizations at http://paulinepi.com

Wikipedia biographies is assumed to be the location where the individual spent their formative years.

The top five states that are associated with an increased likelihood of generating notable individuals are California, Alaska, Nevada, Hawaii, and Louisiana (see http://paulinepi.com).

When restricted to athletes, six of the top ten sports-producing states are from the South with RRs ranging from 1.24-1.83 (Figure 2a: MS, LA, GA, FL, TX, AL). In this region, athletics plays a large cultural role. This is corroborated by the observation that 36% of the top 25 high school basketball, football, and baseball teams in 2010-2011 come from these six Southern states (26/75; p=0.002) (Source: www.maxpreps.com). Given the states' dedication to sports during high school, it is not surprising these states are successful in creating an environment that yields a higher proportion of notable athletes.

California and New York are the top two states with the highest relative risks for entertainers (RR=2.15 (CA); RR=1.78 (NV), Figure 2b). This is not surprising since Californians would have exposure and access to Hollywood and New Yorkers to Broadway. Correlation does not imply causality. For example, the entertainers already living in California and New York can readily expose their children to the industry. However, parents moving to one of these states may not necessarily enjoy an advantage without connections to the entertainment industry.

Conclusions and Future Work

To the best of our knowledge, this is the first work to use Wikipedia to identify sociologically important characteristics. The method is validated using first names and birth states. Results for names are consistent with a previous study, and the results with birth states are confirmed by independent means and states' cultural identities.

Many past sociological studies use a relatively small number of participants (hundreds to low thousands). Smaller studies tend to be designed as case-control rather than cohort studies so that only odds ratios can be calculated. This new methodology permits calculation of relative risk which is more interpretable and intuitive than odds ratios (Davies, et al. 1998). Furthermore, past studies have focused on high school or college students where future 'success' is uncertain, and/or self-reporting which can be unreliable. Using Wikipedia as a resource for successful individuals provides (a) a more unbiased criteria for success which helps address the problems of self-reporting or surveys taken during formative years (b) the use of a cohort design rather than case-control that enables calculation of relative risks and (c) sufficient amounts of data to generate narrow confidence intervals and allow analysis of subcategories (e.g. entertainers and athletes). These advantages lead to better interpretability and application to a general audience. Thus, one can make and verify statements such as: "People with rare names are ~2x more likely to appear in Wikipedia"; and this would be both interesting and accurate to a general audience.

There are precautions when using this methodology. Since Wikipedia is crowd-sourced, results will shift depending upon contributors' changing cultural biases. The Wikipedia entries, filtered for Americans, are currently dominated by entertainers and athletes (38% and 39% of Wikipedia biographies, respectively). An uneven geographical distribution of Wikipedia contributors could also influence geographical risk patterns and some people use Wikipedia to promote their own image. Filters such as requiring a minimum number of edits or unique contributors could be used to remove some of these possible biases (Lih 2004).

The method presented here can be used to validate characteristics that have been implicated in positive outcomes (e.g. education, birth order in family, marital status of parents). Furthermore, it can provide improved statistical confidence compared to prior studies because of the large number of Wikipedia entries. This type of analysis can be applied to any attribute as long as adequate control population data exists. Potentially, the raw text of the Wikipedia biographies may be mined to discover novel characteristics associated with positive life incomes.

References

Cook, T.D. and Campbell. 1979. *Quasi-experimentation: Design and Analysis Issues for Field Settings.* Boston, Mass.: Houghton Mifflin.

Davies, H.T.O., Crombie, I.K., Tavakoli, M. 1998. When can odds ratios mislead? *British Medical Journal* 316:989-991.

Davis-Kean, P.E. 2005. The Influence of Parent Education and Family Income on Child Achievement: The Indirect Role of Parental Expectations and the Home Environment. *Journal of Family Psychology* 19:294-304.

Leventhal, T. and Brooks-Gunn, J. 2000. The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes. *Psychological Bulletin* 126:309-337.

Lih, A. 2004. Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. In 5th *International Symposium on Online Journalism*. Austin, Texas.

U.S. Census Bureau. Measuring America: The Decennial Censuses From 1790 to 2000, September 2002. By Gauthier, J. G. Web. 18 Aug 2009.

U.S. Social Security Administration. Name Distributions in the Social Security Area, August 1997. By Shackleford, M.W. Web. 6 Oct. 2010

Zweigenhaft, R.L. 1977. The other side of unusual first names. *The Journal of Social Psychology* 103:291-302.