Barely Legal Writers: An Exploration of Features for Predicting Blogger Age

John D. Burger and **John C. Henderson**
The MITRE Corporation
202 Burlington Road
Bedford, Massachusetts 01730
{john,jhndrsn}@mitre.org

Abstract

Accurate prediction of blogger age from evidence in the text and metadata of blog entries would be valuable for marketing, privacy, and law enforcement concerns. This paper offers an initial exploratory data analysis of candidate features for blogger age prediction.

Introduction

Personal blogs are an emerging publishing medium containing information of a different type from traditional broadcast news, newswire, and newsgroup data. Text found in blogs is often intimate and detail-oriented, and rarely crafted. The novel types of information prevalent in blogs are perhaps most useful in aggregate form. Phenomena with broad impacts such as disease spread, brand awareness, and reactions to rising fuel prices are interesting to a variety of communities. These trends can be measured in the aggregate in blogs.

Author ages can be found by inspecting profiles of bloggers. This is a new feature, prevalent only in association with this particular text medium. Accurate prediction of blogger age from evidence in the text and metadata of blog entries would be valuable for marketing, privacy, and law enforcement concerns. Models that can predict blogger age from text alone might also be used to predict the age of authors who are not publishing in blogs. While highly accurate prediction of blogger age is not yet attainable, we have investigated several informative features of blog posts in the course of evaluating candidate information sources for blogger age prediction. In this paper we look at a large sample of personal blogs and explore how blogger age relates to several other variables.

Data

A continuing collection of blog entries was initiated in July, 2004. One year of collection has yielded 84 million posts. Periodic network and hardware failures have interfered with collection to varying degrees. This has left us with some gaps in our collection, but those gaps do not significantly reduce utility. For this study, a subsample of 100,000 posts was randomly drawn from the continuous period of harvesting from 12/15/2004 through 1/15/2005. The number of posts per day in this subsample is shown in Figure 1. Some bloggers are more prolific than others, but 87,883 unique posters were found in this sample. Thus, for most of the bloggers represented in the sample, only one post was selected in the random draw.

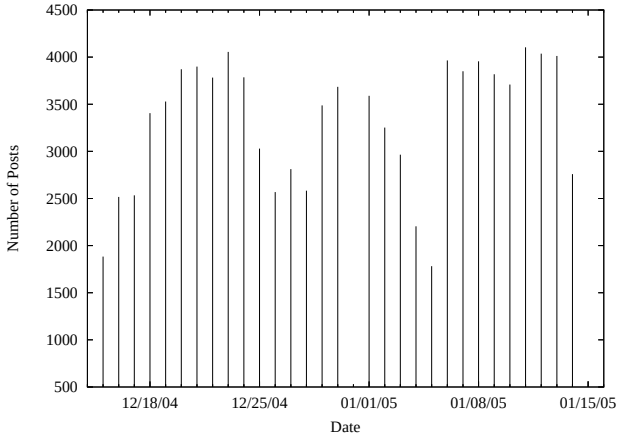

Figure 1: Posts per day in the sample.

To characterize these blogs in a qualitative sense, they are entries in personal journals. They are not community-oriented or agglomerations of expertise on technical subjects. They are subjectively-written rather than well-crafted texts. The topics such as announcements of the arrival of the coming weekend, discussions of best friends' attendance at parties, cries of depression at losing a job, intimate and generally personal subjects one would expect to find in a diary. Figure 2 shows an excerpted first paragraph from a blog entry in the sample.

When registering their blogs, bloggers are given the opportunity to declare some of their personal traits such as name, age and gender. The source from which we drew this sample allows registrants to omit the age if they prefer, or to fill in only partial dates. While bloggers can lie about these traits if they choose to, they can just as easily leave the fields blank to provide partial anonymity. In our sample, we found that 47,968 of the 87,883 bloggers (55%) filled out the birthdate field in its full MM-DD-YYYY form. These users were responsible for 52,449 posts in our sample. Some may have

Today has been different... good up until just now too. I woke up and went over to my neighbor's house and helped them hook up a new DVD player. That took me all of five minutes.. after that I ran up to the mall to run some errands for my mom. That proved to be semi-fruitless as well. I had to go to the Hallmark store to pick up something that my mom had them put on hold for her. The mall was busy too... I've never recalled it being this busy *sighs*... I thought I'd be coming home to a ghost town, but this place has grown. Probably due to the hurricanes and such. Anyway I went to the store and it was packed. There must have been thirty people in this little mall store... I walk up to the desk and ask the lady if she has what my mother put on hold, and I just watch all the intellegence drain from her face. Needless to say, she knew nothing of holding anything for my mom so... I left empty handed.

Figure 2: Text from a blog used in this study.

lied about their age to appear older or younger, but we felt that in this large sample such effects would be drowned out by those answering honestly.

Exploring Blogger Age

Year of birth has been chosen to represent age in this study because it will more easily allow for longitudinal comparisons. The age of a person changes over time, but year of birth remains unchanged. Figure 3 shows the general distribution of the self-reported year of birth in our sample. The top plot shows the smoothness of the curve and the tightness of the predominance of posters in the age range of 14–24. The bottom plot shows the log-linear regularity of the curve as it moves to older bloggers (with earlier years of birth). According to this data, bloggers start posting around age 14 and taper off gradually after age 24. The cohorts from the years 1970–1990 each have at least 100 users, large enough samples to be examined in more detail in later graphs. The small set of posters around the age of 5 is inconsistent with the rest of the plot, and likely consists of blogs either written by or strongly encouraged by their parents.

Location

Figure 4 shows the mean age, standard deviation, and number of bloggers in countries that were represented by at least 100 bloggers. Unsurprisingly, the majority of bloggers in our sample state that they reside in the U.S. with a mean age of 21. The Philippines has the youngest bloggers and Israel's bloggers report the highest ages. The eight-year difference in mean age between bloggers in those countries is surprisingly wide, suggesting one or more cultural biases influencing the age of computer users or an underlying demographic difference.

Time

Figure 5 shows the mean and mode of the time of day that bloggers post as given by the submission time on their entry. The units shown are relative to the GMT timezone. The

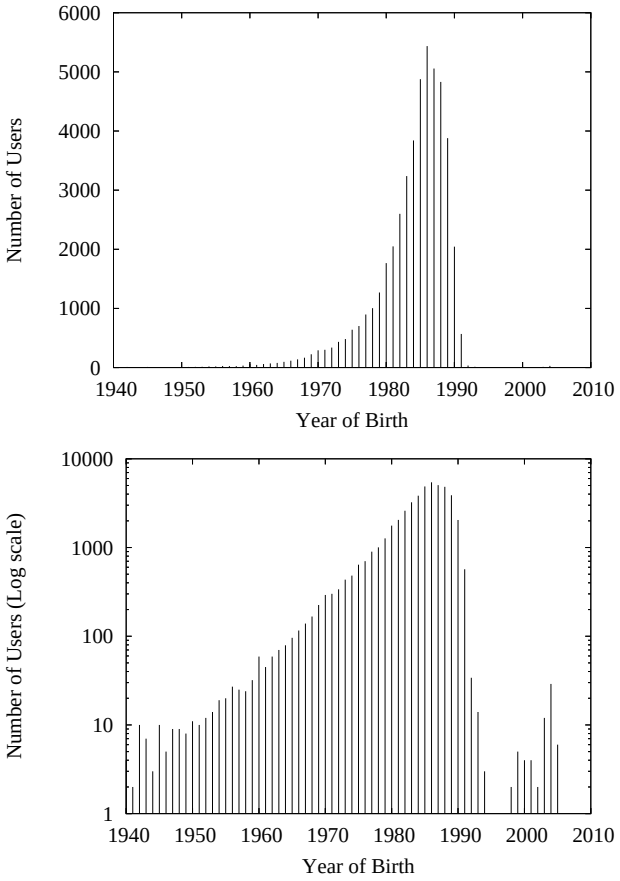

Figure 3: Number of bloggers indicating their year of birth.

Country	Year of Birth Mean	Std. dev.	# Users
Philippines	1985	3.5	300
Finland	1985	4.4	292
Spain	1984	4.7	138
Scotland	1984	5.6	113
Netherlands	1984	5.8	240
United States	1984	6.5	50450
Singapore	1984	7.1	351
Australia	1983	6.6	1278
United Kingdom	1983	6.7	2683
Canada	1983	7.1	3597
New Zealand	1983	7.3	183
None	1983	9.1	17361
Japan	1982	4.9	173
Germany	1982	5.7	348
France	1981	6.4	118
Estonia	1981	6.6	110
Russian Federation	1980	8.6	3979
Belarus	1980	9.1	114
Ukraine	1979	9.1	383
Israel	1977	7.6	242

Figure 4: Mean logger age by country ($n \geq 100$).

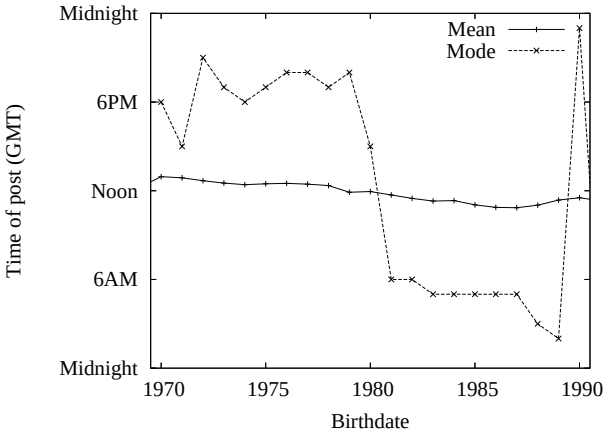

Figure 5: Mean (to the minute) and mode (binned by hour) of posting time (Greenwich Mean Time).

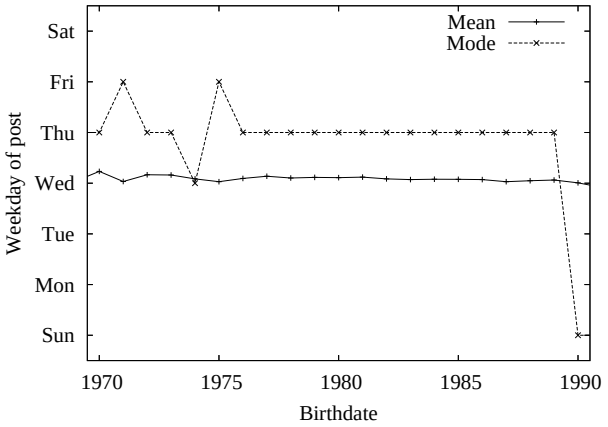

Figure 6: Mean and mode of weekday of posting.

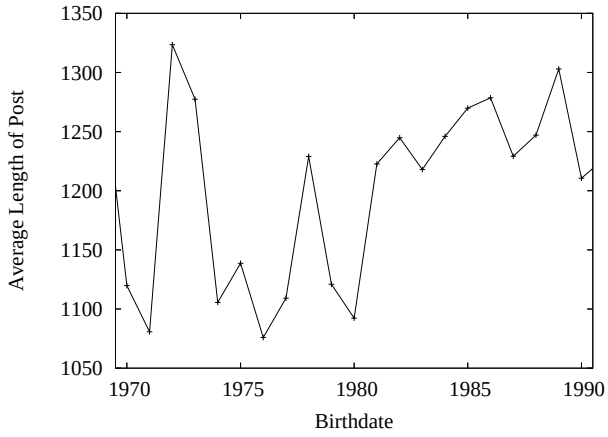

Figure 7: Mean post length.

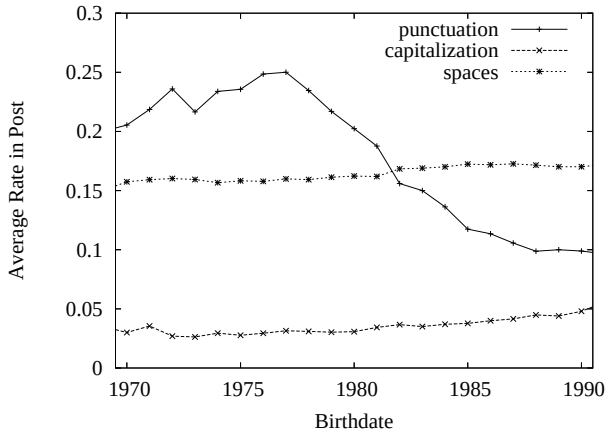

Figure 8: Rates of punctuation, uppercase and space character usage.

mode is the more indicative statistic for this data because of the periodic nature of the dependent variable. The trend of the mean with respect to the year of birth is indicative, but the offset is an artifact of the coordinate system with zero set to midnight.

Keeping in mind that the majority of the bloggers are from the U.S., we see that bloggers under the age of 24 are submitting their post late in the evening. 5 AM GMT is midnight EST, suggesting that younger bloggers in the three timezones of the continental U.S. are posting between 9 PM and midnight. Also notice that bloggers' posting times creep later in the evening until they reach age 23 or 24 (perhaps college graduation) at which point they start posting in the afternoons.

Figure 6 shows the mean and mode of the day of the week that blog posts are made. These curves suggest that there is no dominant change in the day of the week that bloggers post. The right-most point on the mode curve, at 1990, could be an artifact of flatness of the distributions, or suggestive of multimodality. These two phenomena are indistinguishable from this view, but the lack of a predominant differentiation by year of birth makes the outcome of further study of this variable pessimistic.

This dataset spans only 30 days of posting. While there were five opportunities to observe postings on most of the days of the week, there were only four opportunities to observe Monday postings and Tuesday postings. The holidays observed during that period could also have affected the counts from the day of the week. While these factors were independent of bloggers' ages, the effect from time-of-day of posting observed above could have interacted with observation of this variable. A follow-up study spanning more weeks could resolve these issues with more evidence.

Text features

Shallow textual features include the fraction of characters in a post that are punctuation, or that are capitalized letters. The portion of characters that are spaces can stand in as inversely related to average word length. As Figure 8 indicate, these are opposites in that younger bloggers use more capi-

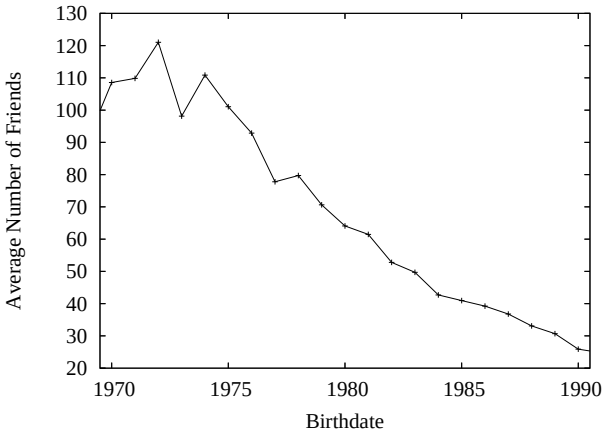

Figure 9: Mean number of declared "friends".

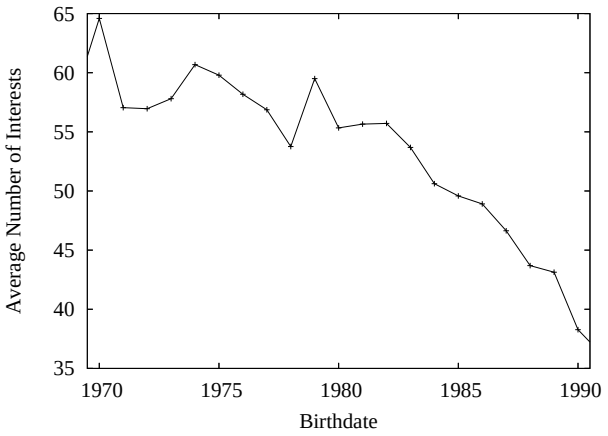

Figure 10: Mean number of interests.

talization, but less punctuation and slightly shorter words.

Figure 7 shows the length of the text part of a blog entry as measured in characters versus the age of the blogger. Posting length increases with year of birth. That is, older bloggers are posting shorter entries. This is a noisy trend, but it appears strongly when looking to either side of the 25-year-old dividing threshold (1980).

Friends

Figure 9 shows the number of "friends" indicated by the profiles of bloggers. These are counts of explicitly declared "friend links" to other bloggers' blogs, offered as an aide for quick navigation and declaration of a relationship. Older bloggers declare more friends. On average, they accumulate at a rate of roughly 4 per year between the ages of 15 and 30.

Interests

While creating their profile, some bloggers indicate their interests by selecting from a common taxonomy provided by the hosting website. Figure 10 is a plot of the number of interests declared by bloggers of various ages. It generally indicates that older bloggers have more interests. They acquire an additional 15 interests between the ages of 15 and 25.

	Year of Birth		
Interest	**Mean**	**Std. dev.**	**# Users**
story of the year	1987	2.7	1072
yellowcard	1987	3.2	1701
my chemical romance	1987	4.3	1681
brand new	1987	4.3	1822
the used	1987	4.3	1844
taking back sunday	1987	4.5	2430
aim	1987	4.8	1579
guys	1987	4.9	1838
the killers	1987	5.0	1518
blink 182	1987	5.3	1462
sushi	1981	7.4	1726
sex	1981	7.4	4103
fantasy	1981	7.4	2218
women	1981	7.7	1716
history	1981	8.5	2236
cooking	1980	6.8	2570
hiking	1980	7.1	1421
travel	1980	7.3	1518
sci-fi	1979	7.1	1014
science fiction	1977	8.9	1366

Figure 11: Mean birth year for blogger interests (youngest and oldest 10 with $n \geq 1000$).

Figure 11 shows the mean age for twenty of the most age-skewed interests declared by more than 1000 users. The youngest users report interests in musical bands and new technology such as "aim", while the older users are interested in more mature subjects and arcane technology such as "science fiction" books.

Conclusion

We have presented explorations of how several features of bloggers and blog entries correlate with blogger age, with an eye toward predicting age from textual content of a blog entry and other blogger metadata. Much of the blogger metadata is indicative of the birth year of the blogger, and some of the shallow textual evidence is weakly indicative.

In future work we plan to use these features together with more fine-grained text features to pursue the task of prediction of blogger age for previously unseen posts.. Following accurate prediction of blogger age, we would be interested in determining how well such models carry over to prediction of author age in other textual genres such as newswire, broadcast news transcriptions, chat logs, and forensic studies.

[0]Note for reviewers: If desired, this paper could expand into a longer paper by including a few more features and discussion of our initial age prediction results. We would separate discussion of age prediction using text content alone from prediction using text in conjunction with blog-specific metadata.