Word Usage and Posting Behaviors: Modeling Blogs with Unobtrusive Data Collection Methods

Adam D. I. Kramer
Department of Psychology
University of Oregon
adik@uoregon.edu

Kerry Rodden
User Experience Research
Google
kerryr@google.com

ABSTRACT

We present a large-scale analysis of the content of weblogs dating back to the release of the Blogger program in 1999. Over one million blogs were analyzed from their conception through June 2006. These data was submitted to the Text Analysis: Word Counts program [12], which conducted a word-count analysis using Linguistic Inquiry and Word Counts (LIWC) dictionaries [20] to provide and analyze a representative sample of blogger word usage. Covariation among LIWC dictionaries suggests that blogs vary along five psychologically relevant linguistic dimensions: Melancholy, Socialness, Ranting, Metaphysicality, and Work-Relatedness. These variables and others were subjected to a cluster analysis in an attempt to extract natural usage groups to inform design of blogging systems, the results of which were mixed.

AUTHOR KEYWORDS

Blogs, Personas, Cluster Analysis, PCA, Unobtrusive, Word usage, LIWC, User Modeling.

ACM CLASSIFICATION KEYWORDS

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

A weblog or "blog" is a web-accessible reverse-chronologically ordered set of essays (usually consisting of a few paragraphs or less), diary-like in nature, maintained and updated by a single individual (user) or a group of users. In 2003, 1.3 million blogs were estimated to exist, with over 870,000 being "actively maintained" [3]. As of the writing of this paper (Sept. 2007), current estimates put the total number of blogs at over 59 million, of which over 903,000 had been posted to during the prior day [3].

Researchers have studied the reasons why people create and maintain blogs. The most common type of blog is a personal diary, maintained by one person to describe the events in their life to the world. Blogs also take the form of news summaries, product announcements and reviews, communities, and other forms of static communication [15,18]. This is to say, the exceptionally interesting question of what blogs are *used for* has been addressed and examined for some time. The online, public nature of these blogs also provides an incredible resource for data mining [16]. For example, consider the question of how to classify and conceptualize the different uses and purposes of blogs: Li, Xu, and Zhang [16] discuss how titles, bodies, and comments can be used to correctly classify blogs with diverse topics. In this paper, we take a more user-oriented approach, and rather than asking how higher-level blog topics and interests can be clustered or induced from blog content, we are interested in how the blog content itself can be used to differentiate and classify users of a blogging system. This is similar to Mishne's [17] work on self-reported emotional state, though we focus on raw post text.

The purpose of this paper is to attempt to answer the questions of "what sorts of people are out there blogging" and "how do blogs vary" from the bottom up. We address these questions from the perspectives of a social psychologist and user experience researcher, and as such we seek to keep the wide range of individuals who use the product in mind. These individuals' similarities and differences may cut across the many varied topics and intentions behind their blogs. As such, the blog content we examine in this paper comes from two distinct sources: Metadata on the usage patterns of the bloggers (i.e., a blogger's posting habits and prolificacy, the blog's creation date, and whether the blog is to be considered "active"), and data on the blog's content (e.g., the relative frequency of certain word categories within posts). We use these data to cluster bloggers, or to discover which groups of bloggers are most differentiable. We then evaluate whether these clusters can be used as "personas." Personas are descriptive, understandable person descriptions, used as targets for product design [3,18], which are analogous to fictional "spokespeople" for different subsets of the user base. Reliance on personas, however, requires that the personas adequately represent the user population, lest some users be unrepresented during the design process [6]. As many systems for determining which personas represent a user base rely on opt-in data (e.g., survey responses), there is

CHI 2008, April 5–10, 2008, Florence, Italy.

some worry that those who choose to provide personal information of any sort may inadequately represent the user base. Blogs have a user base in which many people wish to remain anonymous and thus prefer *not* to be singled out (e.g., for a survey) [18], making this concern especially worrisome. As such, in this paper we concentrate only on unobtrusively collected variables.

It is likely that an understanding of the individuals who comprise the blogosphere would best be obtained by grounding ourselves in each individuals' blog, and forming an understanding of the particular individual in their individual blogging context (i.e., by joining the blog's readership). For a user base of millions of bloggers, however, it is infeasible for a single research team to ground themselves in the contexts of enough blogs to claim that their sample is truly representative via random selection. Further, several analyses have revealed that the systematic noninterpretive processing of natural language text can reveal things that would be missed by a grounded observer (e.g., [21]). As such, when attempting to determine qualities of a bloggers as a whole, we only examine variables which can be measured unobtrusively, and classify bloggers according to these qualities. This process is referred to as "User modeling" [4]. In these respects, our research parallels the work of Kraut and colleagues' [1], who unobtrusively analyzed response patterns for online communities (e.g., USENET).

VARIABLE SELECTION

To model users, it is necessary to select and justify a set of variables on which it is interesting to differentiate users. Our approach is to use the blog-per-person as our level of analysis (i.e., treating each person's contribution to each blog they are an author of as the unit of analysis), in order to understand how blog usage patterns naturally cluster across bloggers. Because we are interested in the differences among bloggers, it is necessary to choose variables that show variance at the individual level. Posting rate variables, such as the average time between posts, have been shown to vary meaningfully at the level of the blogger, as does the overall number of posts and the blog's lifespan [14].

We also used the Text Analysis and Word Counts (TAWC) system [12] to count the number of words in each post to each blog, as well as the number of words from each of the categories presented in Pennebaker and colleagues' Linguistic Inquiry and Word Count (LIWC) research (e.g., [1,13,20]). These categories map to psychological qualities, and have been shown consistently to represent and predict individual-level phenomena [21], and as such are perfect for an individual-oriented modeling task. Misspellings and other idiographic word usage patterns (such as groups using the word "awesome" to mean "intoxicated") were not explored.

This approach differs from keyword-based analyses because LIWC categories are defined from a psychometric standpoint. For instance, greater use of first-person singular pronouns has been shown to indicate greater self-focus [21] even when aggregated across the many different contexts in which the term is used. Keyword models, however, generally focus on identifying a blog's topic (whereas we focus on the user), and often assume that any use of the keyword indicates topic-relevant discussion (which is not always the case). The LIWC dictionaries have also been used in other more specific analyses of blogs and blogging behavior: Nowson & Oberlander [19] used LIWC to show that blog posts are similar to school essays, and that use of LIWC categories can be used to predict the personality of bloggers, using an opt-in survey methodology. Others have used LIWC to attempt to differentiate qualities of bloggers in a top-down fashion, (e.g. to differentiate genders and age ranges [22]), whereas our approach is bottom-up. In addition to the LIWC categories, we also included counts of the number of words with at least 6 letters as a measure of vocabulary complexity, and the number of emoticons that may be used as emotion expression in blogs [1], as well as the total number of words and posts per blog.

DATA SET

We used a collection of anonymized metadata and blog posts from 1,846,445 blogs, comprising 8,242,957,116 words used over the course of 50,070,469 posts, created using Google's "Blogger" web application [2] (the same corpus used in [14]). All data analyzed was at one point published on a public-accessible website, and as such these analyses are consistent with the Google and Blogger privacy policies [7]. This sample[1] contained post dates between August 23, 1999 (the day that Blogger was launched by Pyra Labs) and June 28, 2006. Blogs were not selected based on their activity levels or blog content. Further, ours is a longitudinal lifespan sample of blogs, including all posts to the represented blogs from the first post until the end of data collection, as in [10,14]. As such, we believe that this corpus is a representative sample of users of free blogging systems, and can serve as "normed" comparison data for future research.

[1] Blog selection was based on efficient data access, and was not random. Selection criteria were not related to any variables discussed in this paper. We also excluded blogs that had previously been identified as spam (i.e., posted to automatically for the sole purpose of attempting to manipulate search engine results), via proprietary spam detection methods. Only one other blog was removed, which contained two posts, one of which included only the word "Hi," which was repeated 2,623,991 times, improperly affecting the mean number of communication and other-reference words per post for the entire corpus.

RESULTS: LIWC

Means and standard deviations for the LIWC categories, as well as correlations among the categories, are available in the supplementary materials. Due to the large number of variables present in the LIWC data set, linguistic data were submitted to a principal components analysis (PCA). PCA is a procedure by which a large set of variables is reduced to a small "summary" set, based on the correlations among the larger set (see [8,9,11] for a thorough discussion of PCA factor extraction, rotation, and generalizability; the full correlation matrix is provided in the supplementary materials). This approach suggested a very strong unrotated first factor indicating that all categories correlate (due to baseline wordiness). This factor accounted for 62% of the variance among word categories, effectively controlling for baseline number of words per post. As discussed in [11], we removed this factor and analyzed the remaining covariance. Using VARIMAX rotation, we concluded that a five-factor structure was appropriate using methods suggested in [9], explaining 23% of the remaining variance in word usage. Factors are named based on which word categories loaded with an absolute value greater than 0.4:

- Factor 1 contained positive loadings for affective words, negative emotional words, and sad (but not angry) words, as well as words in the physical, body, eating, and grooming categories. Together, this appears to represent how "Melancholy" the blogger is feeling.
- Factor 2 contained positive loadings for school, job, leisure, home, sports, TV, music, and money, and a negative loading for total number of posts. These social activities appear to preclude frequent posting, indicating that blogs vary in terms of how "Social" the blogger is.
- Factor 3 contained positive loadings for anger, sex, swearing, and self reference, indicating variation in terms of the extent to which blogs are used for "Ranting."
- Factor 4 contained positive loadings for "religious," "metaphorical", and "death" words. This "Metaphysical" factor suggests that blogs vary in terms of whether their topics are metaphysical in nature.
- Factor 5 contained loadings for occupation words, school words, job words, and money. This "Work" factor indicates that blogs vary in terms of discussion of work.

Factor scores were extracted for all individuals on these five factors. Although blogs vary in terms of melancholy, socialness, ranting, metaphysicality, and work-relatedness, this does not mean that blogs are represented by the extremes: An individual blog can occupy any part of this 5-dimensional space [8], so though factors are named after their poles, individual blogs needn't be polar on any factor. Usage factors, such as how frequently they post, may also cut across these word usage factors [14].

Blogs exist with many purposes, and have certainly been shown to cover many topics, and further work is necessary to validate our chosen descriptions of the factors as in [17]. However, the words people are using when they post do vary on these dimensions consistently, providing some evidence that communication using the blogging format follows a five-factor structure.

RESULTS: CLUSTERING

As a first attempt in using this five-factor structure to attempt to classify blogs into natural subgroups, we conducted a cluster analysis. Cluster analysis is a data-driven means of grouping individuals together based on a set of variables (ours listed in Table 1) that can be used to differentiate groups. Using the open-source WEKA tool and the procedures for examining within-cells squared errors, discussed in [4], we determined that a five-cluster solution was appropriate. These five clusters describe how the bloggers in our sample were best differentiated based on our variables, and thus constitute natural groupings.

- Cluster 1 represented blogs that showed high Melancholy, Metaphysicality, and Work-relatedness, and less often showed Rantiness. These blogs rarely met the "Established Blog" criterion [14], indicating that the blog never really "took off," due to a small number of posts or a short lifespan. These appear to be diaries created for emotional expression during a "Sad period."
- Cluster 2 was similar to Cluster 1 but high in Ranting instead of Melancholy, indicating an "Angry period."
- Cluster 3 represented active, established blogs with a reasonable quantity of posts and a wide range of creation dates: These blogs were not all "old standards," nor were they very new. They tended to have a variable post rate, with occasional long gaps between posts. These blogs were not differentiated based on linguistic qualities.
- Cluster 4 represented blogs that were once established and varied across all linguistic categories, spanning all creation dates, which have now been "Abandoned."

Variable	Description
Melancholy	How much sadness is expressed
Social	How socially-oriented posts are
Rantiness	How much ranting the blog contains
Metaphysical	How metaphysical the blog is
Work	How work-related posts are
Creation	When was the blog created
Last date	Date of the last post
Total days	What is the blog's lifespan
Avg. time between posts	On average, how long between posts
Var. time between posts	How variable is the delay between posts
Established	At least 11 posts over 9 days [14]
Recency	The blog has a "recent" post [14]

Table 1. Variables used in the cluster analysis

- Cluster 5 represented blogs that were both created and posted to recently. These blogs were not reliably differentiated on other variables from other clusters. We call these "New blogs."

Together, these five clusters do not provide a compelling case for focusing on specific subgroups of users, as would be necessary for a persona-based approach to interface design. Instead, these categories appear to be differentiating bloggers along a few specific variables, indicating that the variables we used may be more useful as individual difference metrics, rather than used together to attempt to categorize users. While the results of any clustering attempt depend almost entirely on the set of variables provided for analysis, and a different set of variables could provide a more compelling empirical grouping of bloggers, we know that the blogging community is quite heterogenous, and as such it is perhaps unsurprising that blogs are hard to categorize.

FUTURE DIRECTIONS

We have presented a bottom-up analysis of blog content and usage, suggesting a consistent and interesting five-factor structure of the psychologically relevant words used in blog posts, but mixed results when we attempt to examine subgroupings based on natural language factors and usage metrics.

One important direction for future research involves exploring the generalizability of these results to other subpopulations of bloggers: For instance, exploring what qualities of blogging products (e.g., community focus), populations (e.g., company blogs, autistic blogs) affect the usage patterns or contents of posts, or to examine differences between bloggers from different subgroups.

Another important route for future research would examine the cluster results of other sets of input variables or clustering algorithms. Our paper focused primarily on data collection and analysis of word categories, though it is possible that other variables could be used to generate more parsimonious groupings.

ACKNOWLEDGMENTS

This work was conducted during the first author's internships at Google in 2006-7. We thank Kimberly Angelo, Moira Burke, Eric Case, Nika Smith, Jim Lin, and the Blogger team for support, comments, and motivation.

REFERENCES

1. Arguello, J., Butler, B., Joyce, E., Kraut, R., Ling, S., Rosé, C., & Wang, X. (2006). Talk to me: Foundations for successful individual-group interactions in online communities. *Proc. CHI*, 959—968.
2. Blogger. http://www.blogger.com
3. blogpulse, a service of Nielsen BuzzMetrics. http://www.blogpulse.com
4. Chen, H. & Cooper, M. D. (2001). Using Clustering Techniques to Detect Usage Patterns in a Web-Based Information System. *J. Am. Soc. For the Information Science and Technology, 52, 888—904*.
5. Chi, Y., Tseng, B. L., & Tatemura, J. (2006). Eigen-Trend: Trend analysis in the blogosphere based on singular value decompositions. *Proc. CIKM*, 68—77.
6. Cooper, A. (1999) *The inmates are running the asylum*. Indianapolis, IN: Sams/Pearson Education.
7. Google privacy policy. (Sep. 6, 2007) http://www.google.com/privacypolicy.html
8. Gorsuch, R. (1983). *Factor analysis.* Mahwah: Erlbaum.
9. Gorsuch, R. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68,* 532-560.
10. Gurzick, D. & Lutters, W. G. From the Personal to the Profound: Understanding the Blog Life Cycle. In *Proc. CHI 2006*, 827-832.
11. Guttman, L. (1952). Multiple group methods for common-factor analysis: Their basis, computation, and interpretation. *Psychometrika, 17,* 209—222.
12. Kramer, A. D. I., Fussell, S. R., & Setlock, L. D. (2004). Text analysis as a tool for analyzing conversation in online support groups. *Proc. CHI,* 1485—1488.
13. Kramer, A. D. I., Oh, L. M., & Fussell, S. R. (2006). Using linguistic features to measure presence in computer-mediated communication. *Proc. CHI*, 913—916.
14. Kramer, A. D. I., & Rodden, K. (2007). Applying a user-centered metric to identify active blogs. *Proc. CHI,* 2525—2530.
15. Kumar, R., Novak, J., Raghavan, P., Tomkins, A. (2004). Structure and evolution of blogspace. *Comm. ACM 47,* 35-39.
16. Li, B., Xu, S., & Zhang, J. (2007). Enhancing clustering blog documents by utilizing author/reader comments. In *Proc. ACMSE*, 94—99.
17. Mishne, G. (2005). Experiments with mood classification in blog posts. In *Style2005, part of SIGIR.*
18. Nardi, B. A., Schiano, D. J., & Gumbrecht, M. (2004). Blogging as social activity, or, would you let 900 million people read your diary? *Proc. CSCW*, 222-231.
19. Nowson, S. & Oberlander, J. (2006). Differentiating document type and author personality from linguistic features. *Proc. Australasian Document Computing Symposium.*
20. Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count: LIWC (2nd Edition).* Mahwah, NJ: Lawrence Erlbaum Associates.
21. Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Ann. Rev. Psy., 54, 547—577.*
22. Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). *Proc. of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs, 2006.*