Gender Classification of Weblog Authors

Xiang Yan
Department of Computer Science
Stanford University, Stanford, CA
xyan@stanford.edu

Ling Yan
Department of Computer Science
Dartmouth College, Hanover, NH
ling.yan@dartmouth.edu

Abstract

In this paper, we present a Naïve Bayes classification approach to identify genders of weblog authors. In addition to features employed in traditional text categorization, we use weblog-specific features such as web page background colors and emoticons. Our results in progress, although preliminary, outperform the chosen baseline. They also suggest room for significant improvement once more advanced functionalities of the classifier are implemented.

Introduction

Weblogs, also termed "blogs," refer to usually personal, informal writings listed in reverse-chronological order on a web site. Since its debut in 1996, blogging as an online activity has been growing rapidly (Herring *et al.* 2004). In recent years, the academic and business communities have displayed significant interest in blogs. In particular, researchers have attempted to mine writing moods, product opinions, and underlying social networks from blogs (Gruhl *et al.* 2004; Liu, Hu, & Cheng 2005; Mishne 2005). As an information retrieval problem, blog mining differs from traditional text mining in multiple ways: for instance, blog entries are typically short and unstructured, their word choices are highly subjective, and meta-data such as font color provide additional information that traditional text media fail to deliver.

To our knowledge, the attempt to identify attributes of weblog authors, or bloggers, in an automated manner is new. Although some bloggers provide their biographical information in a prominent section of their blogs, many do not. Extracting such information from blog entries may prove valuable. For example, a firm may find it useful to learn how different demographic segments of its customer base blog about its product.

In this paper, we present a proof-of-concept system that classifies a blogger's gender, given a set of his or her blog entries. We hope that with future extensions, this system will extract other biographical information of bloggers as well. Furthermore, it will be interesting to compare our methodology and results with past research on author gender classification for formal, lengthy writings (Koppel, Argamon, & Shimoni 2002).

Problem Formulation

A blogger authors individual blog entries in his or her blog. Listed in reverse-chronological order on a web site, entries generally exhibit uniformity in features: for instance, they share the same background image, use the same font sizes and colors. Based on this observation, it is then natural to study properties of a *set* of multiple entries by the same blogger, rather than properties of individual entries. We abuse terminology here and call such a set a blog entry as well[1].

We now formulate the study of blogs in a fairly standard document-vector model (Manning & Schütze 1999). Consider a blog entry $B = (C_1, C_2, ..., C_n)$, where C_i is some feature of interest: e.g., a word, background color. Given B, we would like to predict the gender of its author, $G \in \{\text{male, female}\}$. That is, compute probability $P(G|B)$ and classify accordingly. Due to Bayes' Rule,

$$\begin{aligned} P(G|B) &= \frac{P(G)P(B|G)}{P(B)} \\ &= \frac{P(G)P(C_1, C_2, ..., C_n|G)}{P(B)}. \end{aligned}$$

We now make the Naïve Bayes assumption[2].

$$P(C_1, C_2, ..., C_n|G) = P(C_1|G)P(C_2|G)...P(C_n|G),$$

which yields

$$P(G|B) = \frac{P(G)P(C_1|G)P(C_2|G)...P(C_n|G)}{P(B)}.$$

To classify gender, we compute the ratio of posterior probabilities $\frac{P(G=\text{male}\ |B)}{P(G=\text{female}\ |B)}$ and compare it to 1. Denoting male and female with g^0 and g^1, respectively, we have,

$$\frac{P(G=g^0|B)}{P(G=g^1|B)} = \frac{P(G=g^0)\prod_{i=1}^{n} P(C_i|G=g^0)}{P(G=g^1)\prod_{i=1}^{n} P(C_i|G=g^1)}.$$

[1] Such abuse does not alter the nature of the problem, but it simplifies our formulation.

[2] The assumption that features are independent, given document label, is clearly inaccurate in this case. But such assumption has proven to work surprisingly well in many problems. See (Manning & Schütze 1999) for details.

If we can estimate prior distribution $P(G)$ and conditional feature probabilities $P(C_i|G)$ from training data, our computation of $\frac{P(G=g^0|B)}{P(G=g^1|B)}$ shall complete. To do so, consider a training set of m blog entries whose author genders are known: $M = \{B_1, ..., B_m\}$. We then estimate $P(G)$ and $P(C_i|G)$ as follows,

$$P(G = g^0) = \frac{|\{B_j|g(B_j) = g^0\}|}{m},$$

$$P(C_i|G = g^0) = \frac{|\{B_j|g(B_j) = g^0 \text{ and } h(B_j, C_i) = 1\}| + 1}{|\{B_j|g(B_j) = g^0\}| + |Val(C_i)|},$$

$$P(G = g^1) = \frac{|\{B_j|g(B_j) = g^1\}|}{m}, \text{ and}$$

$$P(C_i|G = g^1) = \frac{|\{B_j|g(B_j) = g^1 \text{ and } h(B_j, C_i) = 1\}| + 1}{|\{B_j|g(B_j) = g^1\}| + |Val(C_i)|},$$

where $g : \{\text{blog entries}\} \rightarrow \{g^0, g^1\}$ maps a blog entry to its author gender, and $h : \{\text{blog entries}\} \times \{\text{features}\} \rightarrow \{0, 1\}$ satisfies $h(B, C_i) = 1$ if blog entry B exhibits feature C_i, and $h(B, C_i) = 0$ otherwise. We also use $Val(C_i)$ to denote the set of values that a feature C_i can take. Note that in computing $P(C_i|G)$, we use Laplace smoothing to avoid assigning zero probability to unseen features conditioned on a particular gender. See (Manning & Schütze 1999) for discussion on Laplace smoothing.

Finally, we classify B as male-authored if $\frac{P(G=g^0|B)}{P(G=g^1|B)} > 1$, and female-authored otherwise.

Experiments and Results

We collected 75000 individual blog entries authored by 3000 bloggers from Xanga, a free blog service with a large user base, averaging 25 entries per blogger. All these bloggers have posted their genders on their Xanga profiles. We first collected a list of blog URL's by starting at one Xanga user's blog, screen-scraping her page, parsing for other blog URL's that she links to, crawling onto those links, and recursively repeating the procedure. We then subscribed to RSS feeds designated by these URL's to retrieve blog entries.

We performed two sets of experiments with different features. In the first baseline experiment, features are unigrams and each blog entry is represented by a bag-of-words model. That is, each C_i records the occurrence frequency of a given word. See (Manning & Schütze 1999) for model details. In the second experiment, in addition to unigrams, we added the following "non-traditional" features:

- Background color. When background of the blog's resident page has more than one color, we do not use this feature in classification.
- Word fonts and cases.
- Punctuation marks.
- Emoticons: special sequence of punctuations that convey emotions. e.g., :-) and :-D.

In each set of experiments, we trained our classifier both with and without stop words as features, and on training sets of different sizes. We did not stem words, however, as the usage of plural versus singular forms, past versus present tenses, may be suggestive of author gender. A similar position on stemming has been taken in previous work (Kolari, Finin, & Joshi 2006).

Our results in Tables 1 and 2 show that the second experiment achieves significantly higher classification accuracy than the baseline[3], as measured by F-measure[4]. These results indicate that the addition of non-traditional attributes forms a better feature set.

	Entries classified as male-blogged	Entries classified as female-blogged
Male-blogged entries	5379	4621
Female-blogged entries	8210	6790

Table 1: Experiment 1 results (baseline): In test set, 10000 male-blogged entries, 15000 female-blogged entries, precision=.40, recall=.54, F-measure = .50 with $\alpha = .5$, stop words are not removed.

	Entries classified as male-blogged	Entries classified as female-blogged
Male-blogged entries	7101	2899
Female-blogged entries	3824	11176

Table 2: Experiment 2 results: In test set, 10000 male-blogged entries, 15000 female-blogged entries, precision=.65, recall=.71, F-measure = .68 with $\alpha = .5$, stop words are not removed.

Training set size	F-measure	
	Stop words present	Stop words removed
2500	0.07	0.08
5000	0.11	0.10
10000	0.26	0.24
20000	0.37	0.33
30000	0.49	0.45
40000	0.63	0.61
50000	0.68	0.64

Table 3: F-measure vs. training set size in different experiments with or without stop words as features.

As illustrated in Table 3, F-measure improves monotonically as training set size increases. This leaves room for imagination should we train our classifier on a larger set of blog entries.

[3] Precision and recall are calculated with male-blogged blogs as target of interest.

[4] F-measure is defined as $\frac{1}{\alpha\frac{1}{\text{precision}} + (1-\alpha)\frac{1}{\text{recall}}}$, where α is a parameter that tunes relative importance of precision and recall.

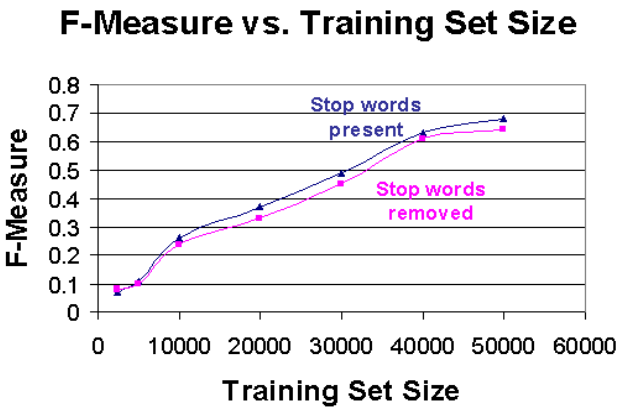

Figure 1: Learning curves for experiments with or without stop words as features.

Table 3 and Figure 1 demonstrate that removal of stop words as features surprisingly hurts classification accuracy. One might therefore conjecture that usage frequency of words like "he" and "then" may be suggestive of author gender.

Furthermore, we enumerate in Table 4 some frequent words that have occurred in blogs authored by one gender but not the other. They seem to align with our general perception of vocabulary usage by males and females in informal writing.

Words in male-authored blogs	Words in female-authored blogs
psst	*muah*
nba	*make-up*
poet	*jewelry*
income	*fabulous*
badass	*barbie*
furious	*layed*
wasup	*kissme*

Table 4: Frequent words that occurred in one gender-authored blog but not the other.

Among unigrams that have appeared in both male- and female-authored blogs, we list here those unigrams with the highest mutual information with the observed blogger gender distribution: *peace*, *shit*, *yo*, *man*, *fuck*, *damn*, *ass*, *sup*, *hit*, *played*, *gay*. These are the most "gender-discriminant" common words. See (Cover & Thomas 1973) for discussion on mutual information.

Future Challenges

To improve classification accuracy, we plan to increase our training set size, refine our parsing algorithm to capture certain features previously ignored (e.g., number and position of images, individual words in bloggers' user ID's), and perform feature reduction. Furthermore, a more ambitious project would be to factor into consideration blog link topologies during classification. That is, examine author gender bias of blogs that link to or are linked from a given blog entry of interest.

Acknowlegments

The authors wish to thank Peter Ciccolo for valuable discussions and implementations during the early phase of the project.

References

Cover, T., and Thomas, J. 1973. *Elements of Information Theory*. Wiley.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *WWW2004: the 13th international conference on World Wide Web*. ACM Press. 491–501.

Herring, S.; Schedit, L.; Bonus, S.; and Wright, E. 2004. Bridging the gap: A genre analysis of weblogs. In *Proceedings 37th Annual HICSS Conference*.

Kolari, P.; Finin, T.; and Joshi, A. 2006. Svms for the blogosphere: Blog identification and splog detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*.

Koppel, M.; Argamon, S.; and Shimoni, A. R. 2002. Automatically catagorizing written texts by author gender. In *Literary and Linguistic Computing*, volume 17(4). 401–412.

Liu, B.; Hu, M.; and Cheng, J. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings WWW 2005*. 342–351.

Manning, C., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Mishne, G. 2005. Experiments with mood classification in blog posts. In *Style2005 – 1st Workshop on Stylistic Analysis of Text for Information Access, at SIGIR 2005*.