Automated Blog Classification: Challenges and Pitfalls

Hong Qu, Andrea La Pietra, Sarah Poon

University of California at Berkeley
School of Information Management & Systems
314 South Hall, Berkeley, CA, USA 94720-4600
{hqu, lapietra ,sspoon}@sims.berkeley.edu

Abstract

Blogs are difficult to categorize by humans and machines alike, because they are written in a capricious style. In the early days of web, directories maintain by humans could not keep up millions the websites; likewise, blog directories cannot keep up with the explosive growth of the blogsphere. This paper investigates the efficacy of using machine learning to categorize blogs. We design a text classification experiment to categorize one hundred and twenty blogs into four topics: personal diary, news, political, and sports. The baseline feature is unigrams weighed by TF-IDF, which yielded 84% accuracy. We analyze the corpus, features, and result data. Our analysis leads us to believe that blog taxonomies need to support polyhierarchy—a given blog may be correctly classified under more than one category.

1. Introduction

The number of blogs is growing at an exponential rate. In 2004, the Pew Institute found that 2-7% of Internet users have a blog and that 11% read blogs [1]. Technorati's web crawlers indicate that there are about 12,000 new weblogs created each day; put another way, a new weblog is created every 7.4 seconds [2]. Given the popularity of blogs, it would be useful if we could devise a content classification system to automatically generate a directory of blogs. It is difficult to group blogs into categories because of the freestyle nature of the discourse. Bloggers write whatever is on their mind, sometimes inventing new vocabulary and grammar. Some blog intentionally deviate from rules of language and decorum to create a spectacle for the sake of attracting a larger audience.

Yahoo started out by creating a directory of websites by human labor. But the growth of the web soon overwhelmed indexers. Online blog directories are taking this manual approach. We believe that the growth of blogs will require automation. The challenges of automated blog classification are daunting.

We tried to classifying blogs using pure statistical measures such as TF-IDF. In addition, we experimented with giving more weight to linguistic features such as the title of individual posts in the blog and the anchor text from incoming links. These efforts proved to be ineffective, because blogs do not fit neatly in mutually exclusive categories: a particular blog can fall into multiple categories.

2. Previous Work

There are many definitions for what constitutes a blog. For our purpose a blogs is a website for personal expression composed of "webpages that are constantly updated with new commentary and links about a particular topic. Often very personal [3]." However, it may be too early to group blogs into a directory based on an arbitary taxonomy because "our collective conceptions of weblogs are changing too quickly to realistically capture them in such frameworks [4]."

Krishnamurthy proposes a classification system along two dimensions: personal vs. topical, and individual vs. community [5]. Hobbyist and experts write topical blogs: the most popular blogs are read by tens if not hundreds of thousand of people daily. Personal blogs, on the other hand, are written as a personal newsletter for the benefit of family, friends, and random strangers [6]. Topical blogs have clearly delineated topics, because the audience expects the blogger to be on topic. By contrast, personal bloggers are not confined to one topic. Personal blogs tend to meander across a range of topics, which has implications for content classification.

3. Processing the Corpus

First, we limited the scope of topics to four topic categories (personal diary, news, politics, and sports) and manually harvested and classified 30 blogs for each categories. Next, we parsed the RSS feeds from these blogs. We then used an open source RSS tool called Magpie to extract the title and body text from the blogs. Finally, we used NLTK [7] and Weka [8] to prepare the corpus and process the text. A more detailed explanation can be found on the web [9].

We measure the term frequency–inverse document frequency weigh of single word tokens (unigrams) in the corpus, removing stop words such as *the* and *of*. We used the Naïve Bayes Multinomial classification algorithm

because it was the fastest and most accurate Weka classifier [10]. Other Weka classifiers yielded significantly lower accuracy.

4. Evaluation

We found that the standard TF–IDF weigh worked better than the other two linguistic features. TF–IDF yielded an accuracy of 84%. The titles text of each blog post and the anchor text of inbound links yielded 76% and 80% respectively. Upon closer examination of results form the training data set and the testing data set, we noticed a very peculiar result: almost all the blogs were correctly identified except for *political* and *news* blogs

The classifier consistently mistook *news* blogs for *political* blogs. All the blogs in the test set were correctly classified except for 5 *news* blogs, which were misplaced into the *political* category. In the training set, 12 *news* blogs were correctly classified as *news*, but 7 *news* blogs were misclassified as *political*. This result indicates that the category *news* blogs is very difficult to pin down. *News* blogs often talk about politics. Even human judgment would have trouble determining whether a blog that talks mostly about politics is a *political* or a *news* blog.

We believe that the source of the erroneous classifications stem from a flawed taxonomy—the topic boundary between news and politics is blurry. Yet the blogsphere is full of blogs that address multiple topics. Consequently, the first step in building an automated blog classification system—taxonomy design—is a pitfall because some blogs belong in multiple categories.

6. Discussion

Blogs are inherently difficult to group into categories. For any given blog there is no definitive category that satisfies the judgment of the classifier, be it human or machine. We investigated the feasibility of applying machine learning to design a computational method for blog classification. The major stumbling block lies in defining the topic groups for categorizing blogs. At first glance, *news* would be a good topic, but it turns out that *news* blogs contain mostly political commentary. A more robust taxonomy for grouping blogs is needed.

In retrospect, we should have used a faceted classification that allows for polyhierarchy. Polyhierarchy would capture multiple facets of a blog. For example, the *Daily Kos* blog [10] supplements breaking news with political commentary. These are two facets of the blog; thus, it belongs to both the *news* and the *politics* categories.

One approach for designing a polyhierarchical blog classification system could be to divide a blog into individual posts. The algorithm would classify individual blog posts and then applying a percentage threshold to determine which facet(s) to assign the blog. The benefit of this approach is that a blog post is likely to be limited to one topic; the drawback is that a single post is a small document and is therefore more difficult to classify.

Ranganathan, the originator of faceted classification, wrote: "The presence of books with multi-faceted subjects was a fact. ...Library classification should reckon with them [11]." This observation is especially true for blogs, which jump from one topic to another. Hence, blog classification systems need to incorporate this insight regarding the nuances of taxonomy design.

Acknowledgements

The authors would like to thank Professor Marti Hearst for her teaching and mentoring.

References

[1] Amanda Lenhart, John Horrigan, and Deborah Fallow, "Content Creation Online", Pew Internet & American Life Project.

http://www.pewinternet.org/pdfs/PIP_Content_Creation_Report.pdf

[2] Sifry, David, “State of the Blogosphere, October 2004”.

http://www.sifry.com/alerts/archives/000387.html

[3] Access eGovernment
http://www.access-egov.info/glossary.cfm?xid=PA

[4] Savage, Sean, “Assignment 5: Squeezing Weblogs Into Little Square Boxes”

http://www.cheesebikini.com/208/archives/001022.html

[5] Krishnamurthy, S. (2002). “The Multidimensionality of Blog Conversations” *Internet Research 3.0*, Maastricht, The Netherlands.

[6] Schiano, D. J., Nardi, B. A., Gumbrecht, M., and Swartz, L. 2004. Blogging by the rest of us. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria, April 24 - 29, 2004). CHI '04. ACM Press, New York, NY, 1143-1146.

[7] Natural Language Toolkit (NLTK)

http://nltk.sourceforge.net/index.html

[8] Pfahringer, B. 2004. The Weka solution to the 2004 KDD Cup. *SIGKDD Explor*. Newsl. 6, 2 (Dec. 2004), 117-119.

[9] Applied Natural Language Processing Resources Page

http://www.sims.berkeley.edu:8000/courses/is290-2/f04/resources.html

[10] McCallum, A. and Nigam K. "A Comparison of Event Models for Naive Bayes Text Classification". In AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48. Technical Report WS-98-05. AAAI Press. 1998.

[11] Ranganathan, S. R., "Library Classification on the March," in "The Sayers Memorial Volume," Library Association, London, p. 84; 1961.

Appendix A

Linguistic Feature	Accuracy
Unigrams (baseline)	84%
Title + 1st sentence	76%
Anchor text	80%
Title + 1st sentence and Anchor text	73%

Appendix B

Training Dataset Results				
=== Confusion Matrix ===				
a	b	c	d	<-- classified as
25	0	0	0	a = personal
1	12	7	0	b = news
0	2	22	0	c = political
3	2	0	17	d = sports
Test Dataset Results				
=== Confusion Matrix ===				
a	b	c	d	<-- classified as
5	0	0	0	a = personal
0	0	5	0	b = news
0	0	5	0	c = political
0	0	0	5	d = sports

Appendix C

Personal Diary Blogs

www.all-she-wrote.net
www.hanagirl.blogspot.com
www.miyer.com
www.quirkychick.org
www.sayaprayer.blogspot.com

News Blogs

wiredblogs.tripod.com/sterling
dneiwert.blogspot.com
mediaminded.blogspot.com
www.amylangfield.com
www.tnr.com/blog

Political Blogs

rightwingruminations.blogspot.com
thecommonmanspointofview.blogspot.com
truthliesandcommonsense.blogspot.com
tsuredzuregusa.blogspot.com
wordsonapage1.blogspot.com

Sports Blogs

thesteelcurtain.blogspot.com
thiscouldbetheyear.joeuser.com
www.bulls.blogspot.com
rocketsnba.blogspot.com
www.basketball-overseas.com